A REVIEW ON ANONYMIZATION TECHNIQUES FOR PRIVACY PRESERVING DATA PUBLISHING

Kavita Rodiya[1], Parmeet Gill[2]

[1]*Student, MIT College of Engineering, Aurangabad, Maharashtra, India*
[2]*Assistant professor, MIT College of Engineering, Aurangabad, Maharashtra, India*

Abstract

With the increased and vast use of online data, privacy in data publishing has now become very important issue. Now days, many organizations are collecting and storing huge volumes of data in large databases. Data publisher collected data from data holders then data publisher released this data to data recipient for research analysis and mining purpose. The released data reveals some information, which is considered to be private and personal. Privacy of information in such scenario becomes subject of research. In recent years, data anonymization techniques become subject of research. In this paper, we provide a review of the statistical Anonymization techniques that can be used for preserving privacy of publish data. Microdata publishing consist of variety of different anonymization methods like Generalization, Bucketization and Suppression. But for high dimensional data generalization loses significant amount of data and affected from curse of dimensionality. Whereas, bucketization fails to preserve membership disclosure and it needs a clear difference between quasi attributes and Sensitive attributes. Suppression reduces quality of data drastically. To deal with these problems, slicing method is introduced. Slicing is new anonymization technique for preserving publish data. In this data is partitioning horizontally as well as vertically. Dimensionality of the data is decreased by slicing. Utility is preserved & correlations between attributes which are highly-correlated together are maintained. Compare to generalization, slicing provides better utility of data. Compare with bucketization, slicing is more effective. High-dimensional data can be handled by slicing. Slicing also Provides attribute disclosure and membership disclosure.

Key Words: *privacy preserving Data publishing, microdata, Information Disclosure, Data Anonymization*

----------***----------

1. INTRODUCTION

In the information age, Organizations collect huge volumes of data from heterogeneous databases. Such type of data includes records for individual person. Example, many government agencies and private companies collected & used data commonly known as microdata [1] and also by web search engines [2] [3] which collects personal browsing histories. After collecting data next step is to publish data such publish data is useful for data mining & research. However, publish data usually contains personal information which may be sensitive; leakage of such sensitive information violates the individual privacy. Examples of some popular recent attacks are example of this like discovering disease of the Massachusetts governor [4], identifying the browsing history of an AOL user [5] etc. Due to such attacks, privacy preserving has become subject of research. Privacy preserving is nothing but to protect from disclosing the identity of individual.

1.1 Microdata Publishing

Microdata can be considered as medical data and census data. Every record has different attributes. This attributes can be classified as:

1. Key Attributes: Explicit Identifiers or key attributes are responsible for explicitly identifies record owners.
2. Quasi-Identifier: Quasi-identifiers are attributes that can potentially identify an individual when their values taken together. Examples are zip-code, nationality, gender and birthdate.
3. Sensitive Attribute: The attribute's values should not be disclosed by attacker. E.g. medical diagnosis, occupation.

Table 1 shows an example. In this name is key attribute, Age, gender and nationality is quasi- identifier and disease is sensitive attribute.

Table 1: Microdata

Name	Gender	Zipcode	Age	Disease
Reena	F	431201	40	Heart Disease
Madhuri	F	444806	23	Flu
kapil	M	444601	26	Heart Disease
Ajay	M	431001	35	Flu
kavita	F	431203	38	Viral Infection
Aarti	F	444604	20	Cancer

1.2 Disclosure Risks

It is necessary to give security to sensitive attribute of the individuals, while microdata is released. In the literature [7] [8] [9], information disclosures are of three types: attribute, identity, and membership disclosure.

- Identity Disclosure: In released data, individual is linked to specific record. The "anonymity" is broken.

- Membership Disclosure: To know that particular person record is in release data.
- Attribute Disclosure: New information is leaked.

1.3 Anonymizing Data

Even publish data have useful information to data miner, it also contains individual risk. Therefore, the aim is to maximize the benefit with minimize individual risk. Anonymization is approach for preserving privacy of publish Data which seeks to secure the sensitive data and identity of record owners. This can be accomplished by anonymization before publish. Initially in anonymization, the identity attributes i.e. explicit identifiers is remove. In Table 1, since Name can disclose the identity of a patient, the data owner removes Name from Table and releases it as shown in Table 2.

Table 2: A published data when the adversary has no background knowledge

Gender	Zipcode	Age	Disease
F	431201	40	Heart Disease
F	444806	23	Flu
M	444601	26	Heart Disease
M	431001	35	Flu
F	431203	38	Viral Infection
F	444604	20	Cancer

However, removing explicit identifier is not enough, as an attacker may have knowledge of the individuals in publish table. Attacker can have this information from personal knowledge, or from public databases such as voter registration list. From table 2 and table 3 one can conclude that Reena has Heart disease.

Table 3: A voter registration list

Name	Zip-code	Gender	Age	Disease
Reena	431201	F	40	Heart Disease
Madhuri	444806	F	23	Flu
kapil	444601	M	26	Heart Disease
Ajay	431001	M	35	Flu
kavita	431203	F	38	Viral Infection
Aarti	444604	F	20	Cancer
Glory	47605	F	30	Heart Disease
Harry	47673	M	36	Cancer
Ian	47607	F	32	Cancer

Further anonymization is required for preventing this type of attacks.

2. RELATED WORK

There are different ways of anonymizing the table before it is published. Two popular methods are grouping-and-breaking and perturbation.

1) Grouping-and-Breaking: This is an operation that divides the records horizontally into a number of partitions and then breaks the exact connection between the quasi value and the sensitive value in each partition [Sweeney, 2002a; Xiao and Tao, 2006b]. The goal of grouping is to make individuals in the same group indistinguishable so that the adversary cannot uniquely identify each individual in the group. The objective of breaking is to weaken the connection between the quasi values and sensitive values so that the adversary has less confidence in the linkage that can be inferred between an individual (which can be identified by the QI values) and a sensitive value in the sensitive attribute such as HIV. Grouping and breaking can be classified into namely suppression, generalization and bucketization.

a) Suppression: Suppression [1] means removing an entire tuple or attribute value. Replaces tuple or attribute values with special symbol '*' that means any value can be there. Table 4 is a table generated by suppression.

Table 4: A published data by suppression

Gender	Zipcode	Age	Disease
*	431201	*	Heart Disease
*	444806	*	Flu
*	444601	*	Heart Disease
*	431001	*	Flu
*	431203	*	Viral Infection
*	444604	*	Cancer

Drawbacks of suppression:

- Quality of the data drastically reduces.

b) Generalization: Samarati and Sweeney proposed to use generalization [1] [4]. Generalization replaces attribute values with semantically consistent but less specific value. Due to this replacement, many records have same QI values. Generalization replaces exact values with a more general description to hide the details of attributes, making the QIDs less identifying. If the value is numeric, it may be changed to a range of values. For example, age attribute value 28 can be changed to range 25-30. If the value is a categorical value, it may be changed to another categorical value denoting a broader concept of the original categorical value. For instance, city Aurangabad can be changed to state Maharashtra. It uses taxonomy to replace attribute value in more general form. Table 5 is table generated by generalization

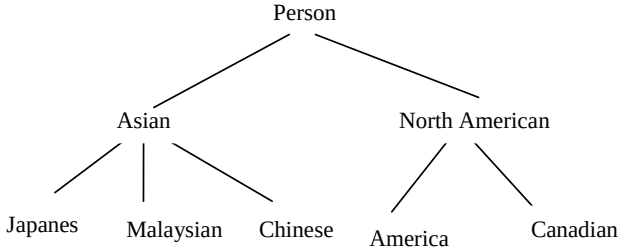

Fig -1: Taxonomy for attribute Nationality

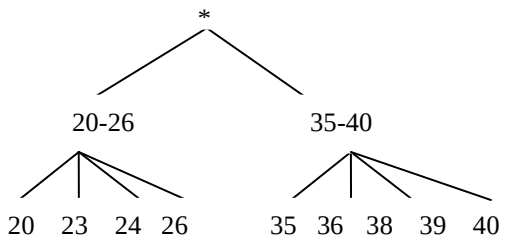

Fig -2: Taxonomy for attribute Age

Table 5: A published table by generalization

Zipcode	Gender	Age	Disease
431201	Person	35-40	Heart Disease
444806	Person	20-26	Flu
444601	M	20-26	Heart Disease
431001	Person	35-40	Flu
431203	F	35-40	Viral Infection
444604	Person	20-26	Cancer

Drawbacks of generalization:

- For high-dimensional data, generalization loses significant amount of information.
- Generalization affected from the curse of dimensionality.
- Generalized data reduces the data utility
- Correlations between different attributes are lost.

c) Bucketization: Bucketization [12] is similar to generalization, but it does not modify any QI attribute or sensitive attribute. Instead, after it divides the records into a number of partitions, it assigns a distinctive ID known as GID to each partition, and all tuples in this partition are said to have the same GID value. Then, two tables are formed, namely quasi attribute (QI) table and the sensitive table. Note that the grouping formed by bucketization is equivalent to the grouping formed by generalization, except that bucketization data contains all the original tuple values while generalization data contains some generalized tuples values. Bucketization has the advantage of allowing users to obtain the original specific values for data analysis.

Table 6: A published data by bucketization

Gender	Zipcode	Age	GID
F	431201	40	1
F	444806	23	1
M	444601	26	2
M	431001	35	2
F	431203	38	3
F	444604	20	3

a) QI table

GID	Disease
1	Heart Disease
1	Flu
2	Heart Disease
2	Flu
3	Viral Infection
3	Cancer

b) Sensitive table

Drawbacks of Bucketization:

- It does not prohibit membership disclosure
- It needs a clear difference between QIs and SAs.
- Correlations between the QIs and the SAs are breaks.

2) Perturbation: Under perturbation [13], a value can be changed to any arbitrary value. For example, Male can be changed to Female and vice versa. Table 7 shows an example with perturbation.

Table 7: A published data by perturbation

Gender	Zipcode	Age	Disease
M	431201	40	Heart Disease
M	444806	23	Flu
F	444601	26	Heart Disease
F	431001	35	Flu
M	431203	38	Viral Infection
M	444604	20	Cancer

Drawbacks of perturbation:

- Reduces data utility.

3. SLICING

To deal with problems occur in generalization and bucketization, T. Li [2012] introduce slicing [15] a new technique to preserve privacy of publish. In this technique, slicing based on partitioning data. Partitioning is done vertically as well as horizontally. In vertical partitioning highly correlated attributes are cluster into column. Every column has subset of attributes that are highly correlated. In horizontal partitioning tuples are grouped into buckets. To break the linking between different columns values of column are randomly sorted. For example, table 7 is sliced data of table 2.

Table 7: A published data by slicing

Zipcode, Age	Gender, Disease
431201, 40	F, Heart Disease
444806, 23	F, Flu
444601, 26	M, Heart Disease
431001, 35	M, Flu
431203, 38	F, Viral Infection
444604, 20	F, Cancer

The dimensionality of the data is reduced by slicing. As compare to generalization and bucketization, preserves better utility. Slicing not only groups highly correlated attributes together but also maintains the correlations between attributes. It breaks the association between uncorrelated attribute, which in turn provide more privacy to publish data. Because these attributes are not rare and identification of this is simple task. Slicing provides better privacy protection because any tuple has more than one multiple matching.

Advantage of slicing:

- Compare to generalization, Slicing preserves better data utility.
- Slicing is more effective than bucketization.
- Slicing can also deal with high dimensional data.

3. CONCLUSION

Data anonymization is one key aspect of Micro data disclosures. Privacy is a key issue in publish data because improper disclosure of certain data assets will harm the prospects. Popular approaches of data anonymization like generalization, suppression, bucketization and perturbation have been used for preserving privacy of publish data which have some limitations like Generalization is unable to handle high dimensional data, it reduces data utility. Suppression reduces the quality of data drastically and Bucketization needs a clear difference between QIs and SAs and it also does not prevent membership disclosure. Perturbation reduces utility of data.

These limitations prompted the development of a novel technique called Slicing. Slicing technique involves partitioning of data horizontal as well as vertical. Compare with generalization, Slicing gives effective data utility as and compare to bucketization in workloads slicing is more effective. Most important benefit of using slicing as technique to preserve privacy of slicing is that it can deal with high dimensional data.

REFERENCES

[1] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information:-anonymity and its enforcement through generalization and suppression," SRI International, SRI-CSL-98-04, 1998.

[2] *http://www .google.com/psearch.*

[3] *http://myweb2.search.yahoo.com.*

[4] L. Sweeney, "K-Anonymity: A model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[5] M. Barbaro and T. Zeller, "A face is exposed for AOL searcher no. 4417749," New York Times, 2006.

[6] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in Proceedings of the IEEE Symposium on Security and Privacy (S&P), pp. 111–125, 2008.

[7] G. T. Duncan and D. Lambert, "Disclosure-limited data dissemination," Journal of The American Statistical Association, pp. 10–28, 1986.

[8] D. Lambert, "Measures of disclosure risk and harm," Journal of Official Statistics, vol. 9, pp. 313–331, 1993.

[9] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," in Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), pp. 665–676, 2007.

[10] C. Aggarwal, "On k-Anonymity and the Curse of Dimension- ality," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 901-909, 2005.

[11] D. Kifer and J. Gehrke, "Injecting Utility into Anonymized Data Sets," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 217-228, 2006.

[12] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.

[13] H. Kargupta, S. Datta, Q.Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in Proceedings of the International Conference on Data Mining (ICDM), p. 99, 2003.

[14] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, "Aggregate Query Answering on Anonymized Tables," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 116-125, 2007.

[15] Tiancheng Li, Ninghui Li, "Slicing: A New Approach for Privacy Preserving Data Publishing" IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 3, March 2012

BIOGRAPHIES

Ms. Kavita Rodiya obtained her Bachelor's Degree in Information Technology from BAMU University during 2009 and currently pursuing Masters Degree in Computer Science and Engineering from the same university. She is working in Computer Dept. of Shreeyash Polytechnic Aurangabad as Lecturer from last 5 years. Her research area includes data mining.

Prof. Parmeet Gill obtained her Bachelor's Degree in computer from Pune University during 2005 and Master Degree in Computer Science and Engineering from BAMU University during 2011. She is working in Computer Science Department of MIT Engineering Aurangabad. Her subjects of interest are Big Data, Data mining.