Journal of Engineering and Applied Sciences 13 (4): 930-933, 2018
ISSN: 1816-949X

Privacy Preserving and Data Publishing using Tuple Grouping Algorithm

[1]G. Prabhakar Reddy, [1]K. Sai Prasad, [1]N. Chandra Shekar Reddy and [2]R. Karthik
[1]Department of Computer Science and Engineering,
[2]Department of Electronics and Communication Engineering,
MLR Institute of Technology, Hyderabad, India

Abstract: The key aspect of privacy protection is information sanitization intent for better data utility. The information is classified into three type's, i.e., first identifiers which labels the identity, second quasi-identifiers the type of disclosure that needs to be focused on identity disclosure. A quasi-identifier refers to a attributes subsets that can uniquely identify tuples like personal health information consists name, address, gender, mobile number and third is sensitive attributes like disease, salary must be disclosed form others. Privacy is a key issue where prospects will be harmed by inappropriate disclosure of few assets.

Key words: Bucketization, generalization, slicing, tuple grouping, data privacy, gender

INTRODUCTION

The term data mining (Goryczka *et al*., 2014) refers as the data analyzing process from data warehouse and summarizing that into useful information (known as knowledge discovery sometimes). The data is collected from the data warehouses and resource for the next stage pre-process data where the data is collected, cleaned and stored. Stored data can be searched by using refine queries and gives required results to the user. Figure 1 explains the process or procedure required for data mining.

Generally the data is divided into three categories like identifier, quasi-identifiers and sensitive attributes. Identifiers are those defined uniquely such as UIAD, security number and organization Id's. Quasi-identifiers (Andhalkar and Ingawale, 2014; Li *et al*., 2012) are the identifiers which can be considered together from the individual data example date of birth, zip code, gender. Sensitive attributes (Goryczka *et al*., 2014; Li *et al*., 2012) are considered as adversary information like disease, salary and location. In this study, we use different techniques like generalization (Goryczka *et al*., 2014), bucketization (Goryczka *et al*., 2014; Li *et al*., 2012) and slicing technique.

Table 1 explains us that the data in the organizations are collected and stored in database by using different types of techniques to provide easy accessibility and privacy too. For example, Table 1 diseases is the sensitive attribute and suppose Bob knows that Jack is 29 years old with 5671 zip code and his record is in table and it will be easy to him to track the data of the Jack if, we won't use any preventive measures.

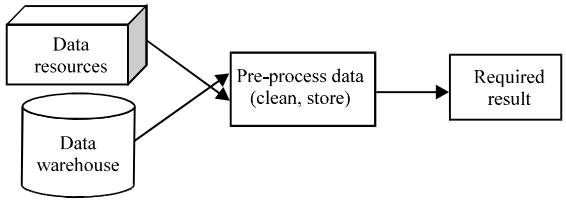

Fig. 1: Process of data mining

Table 1: Original data

Gender	Zip code	Age	Disease
F	5671	29	Heart disease
F	5674	22	Heart disease
M	5673	35	Heart disease
F	5682	34	Flu
M	5688	40	Cancer
M	5689	32	Cancer
F	5672	50	Flu
M	5687	22	Flu
F	5670	43	Cancer

Privacy: One more important aspect of data is privacy (Andhalkar and Ingawale, 2014), hiding the data from the unauthorized persons or third parties. Let's consider a organization of hospital were the personal information of the doctors, patients are stored in the database. That particular data should be accessed by the authenticated persons who belong to that particular organization and provide membership disclosure.

Data publishing: Data collection and publishing (Vasudha and JanakiRamaiah, 2013) are the scenario which are described as the collection of records from the user. The process of the data collection and data

Corresponding Author: G. Prabhakar Reddy, Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad, India

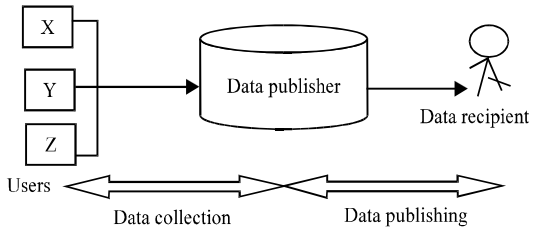

Fig 2: Data collection and publishing

Table 2: Protected data

Age	Zip code	Sex	Disease
2*	567*	*	Heart disease
2*	567*	*	Heart disease
2*	568*	*	Flu
3*	567*	*	Heart disease
3*	568*	*	Flu
3*	568*	*	Cancer
≥40	567*	*	Cancer
≥40	567*	*	Flu
≥40	568*	*	Cancer

*The reference

publishing is explained in Fig. 2 where the information from the different users x, y, z is collected and then it is published in the data publisher which is accessed by the data recipient.

Literature review: The concepts of generalization (Andhalkar and Ingawale, 2014; Valeti *et al*., 2011), bucketization, slicing techniques on the data of the database by which the data can be stored effectively and membership disclosure (Goryczka *et al*., 2014; Li *et al*., 2012) is obtained for the personal information of the members of the organization. Data set is the collection of the data with all the noisy data or data duplication (Aggarwal, 2005). All the information in the dataset is undergo the process of generalization and bucketization where the identifiers are removed from the data. Later on sliced (Goryczka *et al*., 2014), both horizontally and vertically and the data is published into the database. In Table 2, we can observed that, the original data is protected by using the membership disclosure (Andhalkar and Ingawale, 2014). The attributes age, zip code last values and the sex of the original data are disclosed the unauthorized person as the personal information of the user is stored the database and that should be secure. Both bucketization and slicing perform much better than generalization. We compare slicing with optimized slicing in terms of computational efficiency. We fix and vary the cardinality of the data (i.e., the number of records) and the dimensionality of the data (i.e., the number of attributes). Table 2 provides the protected data required for the process. Various comparisons have been with respect to age and disease.

MATERIALS AND METHODS

Slicing: The data in the database is partitioned both horizontally and vertically for data utility purpose. Attributes are grouped in vertical partitioning (Vasudha and JanakiRamaiah, 2013) based on the correlations. Partitions are formed into tuples and later they are grouped as buckets (Valeti *et al*., 2011) in the horizontal partition. The advantage of slicing is that it reduces the data dimensionality over generalization and bucketization (Aggarwal, 2005). Slicing with tuple grouping algorithm provides efficient random tuple grouping for micro data publishing. Each column contains Sliced Bucket (SB) that permutated random values for each partitioned data. The frequency of the value in each one of the scan's-diversity algorithm checks the diversity in each sliced table.

Steps:

- Load data set
- Attribute partition
- Process tuple partition and buckets
- Slicing
- End

RESULTS AND DISCUSSION

In addition to the generalization and slicing techniques, we propose a one more approach called tuple grouping algorithm like tuple space search by which we can increase the time efficiency of the dataset in the database. By which the database can be used effectively and efficiently (Fig. 3).

Methods used

Generalization: In generalization process we provide membership disclosure (Goryczka *et al*., 2014; Andhalkar and Ingawale, 2014) by considering quasi-identifiers that are already known and can be taken together like zip code, dob and gender. These values are grouped, so that, tuples in the same group could not be identified by QI values through which better data utility is obtained. Generalization may leads to the loss of information, hence, we use other technique called bucketization. Highly correlated attributes are present in columns. Tuples grouped into buckets in horizontal partitioning values in each column are randomly permutated to break the linking between different columns in each column. Table 3 provides the generalization data. Various comparisons have been made with respect to age and disease.

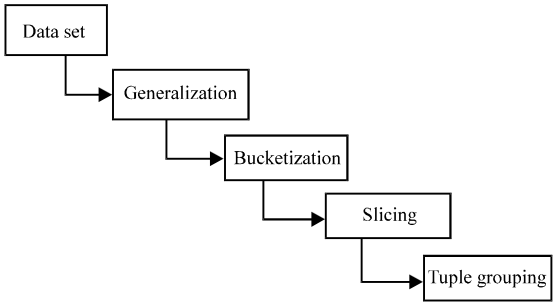

Fig. 3: Architecture

In Table 3, we can observe the concept of generalization where the quasi-identifiers like sex, zip code disclosed in order to provide security and age is generalized as 20-50, 50-80. After separating them according to age and zip code the information is disclosed.

Bucketization: Bucketization (Goryczka *et al*., 2014; Valeti *et al*., 2011) process separates sensitive attributes from quasi-identifiers randomly permitting the sensitive attribute values in each bucket. In this process, we can avoid the information loss but cannot provide membership disclosure as quasi-identifiers are separated.

In Table 4, we observe the process of bucketization (Aggarwal, 2005; Senthil and Vidya, 2013) where the data is formed into buckets according to the zip code and age but there is no membership disclosure to the information but the sensitive attributes are separated from the quasi-identifiers (Vasudha and JanakiRamaiah, 2013).

Slicing: In this process bucketized data is sliced both horizontally and vertically. It preserves better data utility than generalization and provides more attribute correlations with sensitive attributes (Andhalkar and Ingawale, 2014) than bucketization. Dimensionality of data is reduced and attribute disclosure is obtained effectively.

In Table 5, we observe the slicing process where the information is sliced in the form of two dimensional (age, sex) and (zip code, disease). By using this the information can be effectively stored and easy to retrieve.

Tuple grouping: In tuple grouping algorithm we are using the tuple space search algorithm which is same as slicing but in slicing data utility and data anonymity is not in effective manner. But when we use algorithm partition (T, B).

Table 3: Generalization

Age	Zip code	Sex	Diseases
20-50	567*	*	Heart disease
20-50	567*	*	Heart disease
20-50	567*	*	Heart disease
20-50	567*	*	Flu
20-50	567*	*	cancer
20-50	568*	*	Flu
20-50	568*	*	Cancer
20-50	568*	*	Cancer
20-50	568*	*	Flu

*Represents the reference

Table 4: Bucketization

Age	Zip code	Sex	Diseases
22	5674	F	Heart disease
29	5671	F	Heart disease
35	5673	M	Heart disease
43	5670	F	Cancer
50	5672	F	Flu
22	5687	M	Flu
29	5682	F	Flu
40	5688	M	Cancer
32	5689	M	Cancer

Table 5: Slicing

Age, sex	Zip code, diseases
22, F	5674, Heart disease
29, F	5671, Heart disease
35, M	5673, Heart disease
43, F	5670, Cancer
50, F	5672, Flu
22, F	5687, Flu
29, F	5682, Flu
40, M	5688, Cancer
32, M	5689, Cancer

Algorithm 1; Partition (T, B):

1. A = {B}; SB = {}
2. While A is not empty
3. Remove first bucket B from A
   A = A-{B}
4. Split B into two buckets B1 and B2
5. If check (T, A U {B1, B2} USB)
6. QA = A U {B1, B2}
7. Else SB = SBU{B}
8. Return SB

By using this process, we can retrieve the data quickly. Imagine the data set contains 200 records in slicing it may takes 5.50 msec and if we use tuple space search algorithm, it may takes 1.23 msec.

CONCLUSION

In this study data publishing is processed by using the generalization (Andhalkar and Ingawale, 2014), bucketization, slicing along with tuple grouping algorithm where the data in the dataset is represented in the form of buckets by eliminating the quasi-identifiers and provides membership disclosure to the information of user. Senthil and Vidya (2013), Reddy *et al*. (2014) is performed both horizontally and vertically by which we obtain the

efficiency to the data. By using tuple grouping algorithm along with slicing represents the quick retrieval of information which can be observed by considering time complexity of the retrieval of the data. As, we research with large amount of data we need to use effective algorithms and techniques for the better usage of this research. In this study, we worked on two dimensional data and specified the efficient accessing of data. This research can be extended by using the multidimensional dataset.

RECOMMENDATIONS

In addition to this, we can use technique like slice and dice in the future research. The data which is slice based on categories will be later on undergo the process of dice which means that sub cubes are formed by the sliced data and the data storage and the data retrieval will be more effective on the searching process.

REFERENCES

Aggarwal, C.C., 2005. On K-anonymity and the curse of dimensionality. Proceedings of the 31st International Conference on Very Large Data Bases, August 30-September 2, 2005, ACM, Trondheim, Norway, ISBN:1-59593-154-6, pp: 901-909.

Andhalkar, A. and P. Ingawale, 2014. Slicing: Privacy preserving data publishing technique. Intl. J. Comput. Organ. Trends, 5: 85-89.

Goryczka, S., L. Xiong and B.C. Fung, 2014. $ m $-Privacy for collaborative data publishing. IEEE. Trans. Knowl. Data Eng., 26: 2520-2533.

Li, T., N. Li, J. Zhang and I. Molloy, 2012. Slicing: A new approach for privacy preserving data publishing. Knowl. Data Eng. IEEE. Trans., 24: 561-574.

Reddy, E.U., M.V.U. Rani and M.S. Rao, 2014. Privacy in anonymizing horizontally partitioned data. Intl. J. Innov. Sci. Eng. Technol., 1: 493-496.

Senthil, R.M. and B.D. Vidya, 2013. Enhancement of privacy preservation in slicing approach using identity disclosure protection in ITSI. Trans. Electr. Electron. Eng., 1: 37-42.

Valeti, N., A. Chittineni and M.S.S. Sai, 2011. Privacy preserving based on tuple space search methods. IJDCST., 1: 34-37.

Vasudha, T. and B. JanakiRamaiah, 2013. Sensitive micro data disclosures based on tuple grouping methods. Intl. J. Adv. Res. Comput. Sci. Software Eng., 3: 103-110.