INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

ENHANCING THE PRIVACY OF CROWDSOURCING DATABASE USING STA

[1]M.Gokilavani, [2]Dr. A.V.Senthil Kumar

[1]Research Scholar, Karpagam University Coimbatore, Tamilnadu - India
[2]Department of Computer Applications (PG), Hindusthan College of Arts and Science
Coimbatore, Tamilnadu – India.

Abstract—Crowdsourcing is a platform where employers are allowed interact with the employees and get their work done.Crowd sourcing environment may contain sensitive attributes, which cause privacy leakage so that outsiders could link SA's with other public databases to reveal individual confidential information. Many techniques like randomization, k-anonymity, l-diversity have been proposed in recent years in order to protect the privacy of data. In this paper we introduce an enhanced approach in slicing, where the data is partitioned horizontally and vertically, that helps to overcome privacy violations to the maximum extent.

Index Terms— Crowdsourcing, K-Anonymity, database privacy, bucketization, generalization, Slicing.

1. Introduction

Crowdsourcing [1] database is platform where the employer gets his work done from a group of employees based on his requirement. The employer and the employee are the participants in crowdsourcing platform. The employers are allowed to upload their job details to the crowdsourcing platform and the registered employees are candidate employees who provide the labor when and where needed. Current crowdsourcing platforms, such as Amazon AMT1 and Crowdflower2, use the new LaaS (Labor as a Service) [1] model. Operators process each relevant answer from the workers and publish the records for further processing in the database. Some of the current crowdsourcing platforms [1] are Amazon AMT and Crowd flower which adopts a labor as a service model. For example, the HR agents [1], such as 51Job3 and ChinaHR4, receive thousands of requests from the employer and the employee. Millions of users register their curriculum vitae (CV) and thousands of companies submit their job positions to the agents. Each database consist of professional records and also personal details such as salary, educational qualification, working experience etc.,.These information cannot be revealed public which are considered to be sensitive information, whereas these informations has to be shared with the employers. Employers require these information to select the appropriate candidate suitable for the particular task. The primary task of operators is to collect the answers related to the queries which were published earlier. For example human resource agencies receive many applications from both the users and companies. Users submit the resumes and companies submit their job positions.

The interaction is not only restricted to both the parties even to the other employees. These agencies understand the

requirement of the companies and suggest the suitable candidates for a particular job. Crowd data sourcing, allow the employers to access the required information very fast and can understand the various skills of employees in a very short span. Sharing of data to the outside world has become important in recent years. The primary concern on privacy preserving data mining developing a strategy adapted to access exact data information about an individual. This privacy preserving has its wide range of applications in hospitals, educational institutions, government departments. There is a vast requirement for retrieving the relevant information of treatments provided for a patient, census database, and media related data and government department's database. There are many situations where the sharing of data leads to lots of benefit. The basic idea behind the research activities in privacy preserving data mining is to develop various techniques and algorithms that modify the original data in some way such that the private data and knowledge remains private even after passing the orginal data in mining algorithms. In Croudsourcing database systems, K-Anonymity techniques are used to protect the privacy of
Published data. After generalizing the data, K-Anonymity algorithm guarantees that any tuple in the database cannot be identified from the remaining K-1 tuples. However, K-Anonymity affects the performance of crowdsourcing, as generalization and grouping leads to information loss. The attacker is restricted from hacking the published data with other public databases to reveal the identity of a particular person. If the anonymized data is provided to the human workers, they may fail to return the correct answer.The system needs to address the tradeoff between the privacy and accuracy. Generalization transforms the Quasi-Identifiers values in each bucket into "less specific but semantically consistent" values so that tuples in the same bucket cannot be distinguished by their QI values. Where as in bucketization, SAs are separated from the QIs by randomly permuting the SA values in each bucket. The data processed in anonymisation technique consists of a set of buckets with permuted sensitive attribute values. In both approaches, attributes are partitioned into three categories [4]. Some attributes are identifiers that can uniquely identify an individual, such as Identity Number or Name .Some attributes are Quasi-Identifiers (QI), which are common among people but when taken together can identify the person such as Age, Sex, and Zipcode. Some attributes are Sensitive Attributes (SAs), which are unknown to the adversary and are considered sensitive. In both generalization and bucketization [4], identifiers are removed from the data and tuples are partitioned into buckets. In this paper we introduce a new approach called Slicing, which will overcome the limitations and ensure that the privacy is preserved.

2. EXISTING SYSTEM

Generalization [2][3] is a well-known method for privacy preserving data mining. Generalization is the process of replacing the original data with some dummy values. Basically quasi-identifier values are replaced with some values that are not similar but semantically consistent. There are many drawbacks in generalization algorithms such as heavy loss in information, especially for the high-dimensional data.

Bucketization [3], is the process of partitioning the tuples in the table into small units called buckets.Then buketisation algorithms separate the sensitive attribute and non-sensitive attributes by randomly partitioning the sensitive attribute in each bucket. The processed data set consists of the buckets with sensitive attributes. Bucketization preserves better data utility than generalization. Partitioning the tuples into buckets and within each bucket, here we apply an independent random permutation to each column.

A Feedback based algorithm [1] is a straightforward solution is to iterate all possible K-Anonymity strategies and evaluate each strategy based on the samples. A heuristic approach is used here by combining the sample-based feedbacks and the multidimensional K-Anonymity approach [1]. Each cut will result in a set of new cells.The samples in those cells are then anonymized based on the cell ranges. The anonymized samples of crowdsourcing jobs are published to collect the feedbacks for the cells.

3 PROPOSED SYSTEM

Slicing [4] is the process of partitioning the dataset both vertically and horizontally. Data utility is highly preserved and the data is protected against membership disclosure. One more aspect of slicing is that it handles huge dimensional data. Vertical partitioning is attained by grouping attributes into various columns based on the correlations

among the attributes. Each column contains a subset of the attributes that are highly correlated. Horizontal partitioning is attained by grouping the tuples into buckets. The values in each column are randomly sorted and processed to break up the link between the columns. The basic idea of slicing method is to break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of the data and preserves better utility than generalization and bucketization methods. One-attribute slicing partitions attributes so that highly-correlated attributes are in the same column. Here we have sliced the tuples based AGE (AGE<54 and AGE>54). In Two-attribute partitioning two fields namely (AGE, SEX and ZIPCODE, EDUCATION) are combained.

AGE	SEX	ZIPCODE	EDUCATION
34	M	642134	4
34	M	642323	4
34	M	642323	4
45	F	642134	3
45	F	642134	3
56	F	644545	4
55	F	645534	3

Table 1 : One-atrribute Slicing

AGE	SEX	ZIPCODE
34	(34,M)	(642134,4)
34	(34,M)	(642323,4)
34	(34,M)	(642323,4)
45	(45,F)	(642134,3)
45	(45,F)	(642134,3)
56	(56,F)	(644545,4)
55	(55,F)	(645534,3)

Table 2 : Two-atrribute Slicing

AGE	SEX	ZIPCODE	EDUCATION
34	(34,M)	(642134,4)	(34,M,4)
34	(34,M)	(642323,4)	(34,M,4)
34	(34,M)	(642323,4)	(34,M,4)
45	(45,F)	(642134,3)	(45,F,3)
45	(45,F)	(642134,3)	(45,F,3)
56	(56,F)	(644545,3)	(56,F,3)
55	(55,F)	(645534,3)	(56,F,3)

Table 3: Slicing Technique Algorithm

This is good for both utility and privacy. Data utility is achieved by grouping highly-correlated attributes preserve the correlations among those attributes. Privacy is achieved by the association of uncorrelated attributes and guarantees higher identification risks. Therefore, it is better to break the associations between uncorrelated attributes, in order to protect privacy. One-attribute-per-column slicing in table-1 preserves attribute distributional information. It

does not guarantee to preserve attribute correlation, because each attribute is in its own column. In two-attribute sliced table shown in Table 2 the correlations between Age and Sex and correlations between Zipcode and Education are preserved. The sliced table in table-3 encodes the same amount of information as the original data with regard to correlations between attributes in the same column.

Algorithm for Slicing

INPUT

Microdata table T

OUTPUT

The Sliced Table S

Algorithm

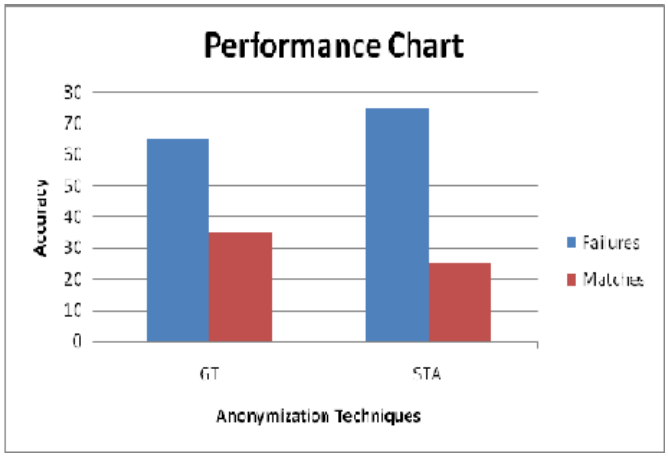
Step 1: Partition the Table T into several subsets of A, such that each attribute fits to exactly one subset. Each subset of attributes are called as columns.

Step 2: Two correlated attributes are combained in a single column.

Step 3: These Correlated attributes partitioned into subset based on Age of the Employee.

4 EXPERIMENTAL RESULT

The anonymization techniques play a major role in preserving privacy for publishing the data. Each anonymization techniques differ in nature but it maintain the data utility and the time consuming for the process must be less in nature to achieve its efficiency. Previous anonymization techniques such as generalization. buketization, feedback-based approach have a loss in utility than the proposed method slicing technique algorithm. The micro data table is taken and retrieved from the database. The preprocessing steps must be applied on the table before the workload experiments done on the data. After computing the preprocessed data the sensitive attribute and quasi identifier are examined. In order to measure the performance level of slicing techniques against the several privacy threats such as identity, membership and attribute disclosures the accuracy can be measured. The accuracy can be determined by the matching of fake tuples and buckets to the original data. This experiment demonstrates that slicing preserves better data utility than generalization is more effective than bucketization in workloads involving the sensitive attribute and the sliced table can be computed efficiently.



5. CONCLUSION AND FUTURE WORK

In recent years development of data analysis and processing techniques led to privacy disclosure problem about individuals when releasing or sharing data with the public. Published data quite often contains personally identifiable information (QI's) and therefore releasing such data may result in various privacy breaches. In this paper, we pre-

sented the issues in privacy preserving methods and the merits of slicing (horizontally and vertically partitioned data) technique. Our future work will be to develop an efficient algorithm which partitions the data vertically and horizontally that would minimize the privacy breaches. Our future work will be focused on M-Privacy and Overlap slicing

6. REFERENCES

[1] Sai Wu, Xiaoli Wang, Shen Wang, Zhenjie Zhang and Anthony K.H. Tung, "K-Anonymity for crowdsourcing database" 2013.

[2] R. J. B. Jr. and R. Agrawal, "Data privacy through optimal k-anonymization",in ICDE, 2005.

[3] Sweeney.L, "Achieving k-anonymity for privacy protection using generalization and suppression," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, 2002.

[4] Tiancheng Li, Ninghui Li, Jian Zhang, Lan Molloy, "Slicing: A new approach to privacy preserving data publishing".120-150, 2012.

[5] Samarati and L. Sweeny, "Generalizing data to provide anonymity when disclosing information" SIGART Symposium on Principles of Database Systems. (PODS 98).

[6] A.V. Evfimievski, J. Gehrke, and R. Srikant, "Limiting Privacy Breaches in Privacy Preserving Data Mining," Proc. ACM Symp.Principles of Database Systems (PODS), pp. 211-222, 2003.

[7] B.C.M. Fung, K. Wang, and P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation," Proc. Int'l Conf. Data Eng.(ICDE), pp. 205-216, 2005.

[8] K. LeFevre, R. Ramakrishnan, and D. J. DeWitt "Mondrian multidimensional k-anonymity", in ICDE, 2006.

[9] Sweeney.L, "k- Anonymity: A Model for Protecting Privacy,"International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, volume 10, issue 5, pp.557-570, 2002.

[10] A. Machanavajjhala, J. Gehrke ,D. Kifer, M.Venkitasubramaniam (2007). ℓ-Diversity: Privacy Beyond k-Anonymity. ACM Transactions on Knowledge Discoveryfrom Data, Vol 1(1), Article: 3.

[11] N. Li, T. Li, and S. Venkatasubramanian 2007 t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, Proc. IEEE 23rd Int'l Conf.Data Eng. (ICDE), pp. 106-115.