Recent Trends in Intensive Computing
M. Rajesh et al. (Eds.)

doi:10.3233/APC210292

Enhanced Data Privacy Using Vertical Fragmentation and Data Anonymization Techniques

R Sudha [a,1], G Pooja [a], V Revathy [a] and S Dilip Kumar [a]
[a] *Department of Computer Science and Engineering, Arasu Engineering College, Kumbakonam, Tamilnadu, India*

Abstract. The use of online net banking official sites has been rapidly increased now a days. In online transaction attackers need only little information to steal the private information of bank users and can do any kind of fraudulent activities. One of the major drawbacks of commercial losses in online banking is fraud detected by credit card fraud detection system, which has a significant impact on clients. Fraudulent transactions will be discovered after the transaction is completed in the existing novel privacy models. As a result, in this paper, three level server systems are implemented to partition the intermediate gateway with better security. User details, transaction details and account details are considered as sensitive attributes and stored in separate database. And also data suppression scheme to replace the string and numerical characters into special symbols to overcome the traditional cryptography schemes is implemented. The Quassi-Identifiers are hidden by using Anonymization algorithm so that the transactions can be done efficiently.

Keywords. Anonymization, Quasi-identifiers, Credit Card fraud Detection System, Traditional Cryptography, Three-Level Server Systems, Intermediate Gateway, Data Suppression Scheme.

1. Introduction

The upcoming methodologies of communications techniques, online payment transactions are increasing day by day. The major drawback for internet banking in current business is that fraudulent transactions appear frequently like authentic ones and it is not efficient to use the simple pattern matching techniques. We can implement vertical clustering algorithm to cluster the datasets into more than one level. Subsets of properties (that is, segments) structure the pieces [1]. Rows of the parts that relate to one another must be connected by an identifier that identifies tuple. A vertical fragmentation relates to a procedure that undergoes projection on the table. The recombination of data derived from the fragmented tuples is done to get original dataset. The joint operator is utilized for identifying tuple to link the columns in vertical clustering in order to link column to the fragments whereas the union operation is used on the rows of the datasets in horizontal clustering.

[1] R Sudha, Department of Computer Science and Engineering, Arasu Engineering College Kumbakonam, Tamilnadu, India; E-mail: suthathilo@gmail.com.

Given person-specific personal data, produce an arrival of the information with logical ensures that the people which are the subjects of the information can't be re-recognized while the information remain basically helpful. If the output information contains the information of each person then it said to have k-anonymity property and it cannot be differentiated from at least k-1 individuals whose information also appear in the output dataset.

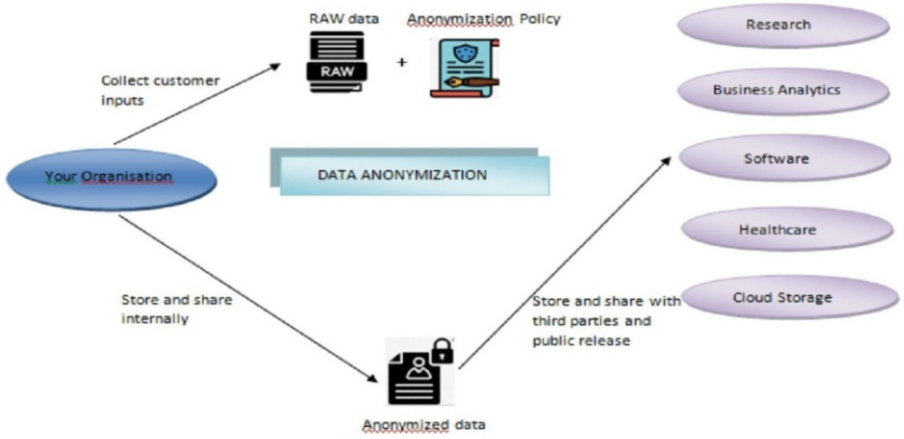


Figure 1. Data Anonymization

2. Existing models

In Existing paper, they have used traditional privacy preserving models, l-diversity and t-closeness. L-diversity make less the granularity in information portrayal utilizing strategies like generalization and suppression. If there were l no of well-versed values for the Quassi identifiers then that set of attributes is said to be l-diversity [2]. The distribution of the class which have equality is required by T-closeness which is moreover close to the attribute distributed in the whole table is shown in Figure 1. The necessitates of this method is the appropriation of a Quassi identifier in the identical class is near the conveyance of the sensitive attribute in the full table. These two strategies have certain drawbacks and prone to attacks called homogeneity attack and background knowledge attack.

The basic assumptions of this paper are the various attributes present in the database. The attributes which explicitly identify individuals in a database are called explicit attributes. To identify individuals the attributes combine with other tuples which are known as Non-sensitive and Sensitive QID's. The different methodology and projects for creating anonymised information giving k-anonymity security have been licensed.

3. System architecture

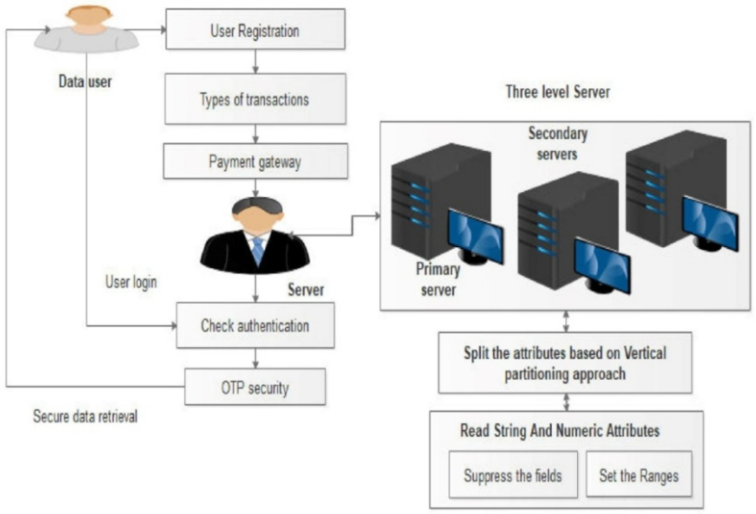


Figure 2. System architecture

4. Proposed Privacy Models

With the approach of correspondences strategies, web based business just as online installment exchanges are expanding step by step. In Figure 2, the serious issue for internet business today is that fake exchanges show up an ever increasing number of like real ones and straightforward example coordinating with strategies are not proficient to recognize misrepresentation [3]. Vertical clustering will do the projection operations on the table. Information from the pieces can be combined again to bring about the first informational index. For vertical clustering, the join administrator is utilized on the tuple identifier to interface the segments from the sections; in even fracture, the association administrator is utilized on the columns resulting from the pieces is shown in Figure 3. The different techniques and projects for producing anonymised information giving k-anonymity protection have been implemented.

4.1 Vertical Clustering Algorithm:

The Vertical Clustering Algorithm is a technique that fragments the records in column-wise manner and stores them in separate databases. The personal details, transaction details and account details in separate servers. There are two servers namely, primary server and a secondary server to store the fragmented record. The personal details are stored in primary server [4] . The Transaction details and account details are stored in a Secondary server. Vertical clustering is a strategy that segments the whole dataset into a few quantities of little data sets dependent on the segment, with the end goal that the fragmented data set doesn't have any copy data. There are basically two sorts of vertical information base specifically standardized and column portioning.

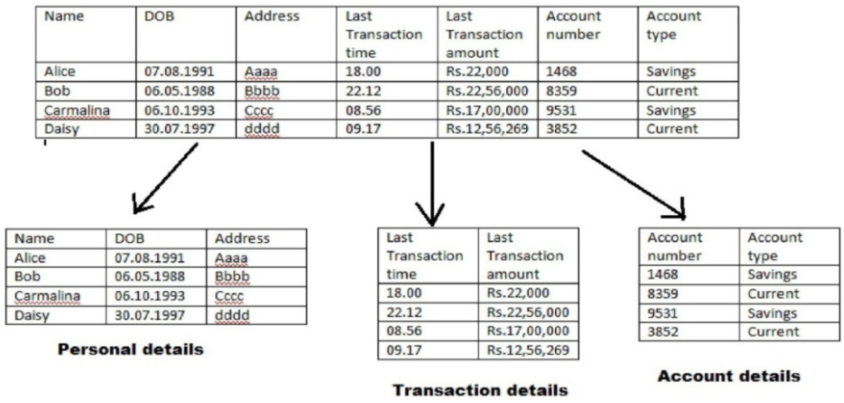


Figure 3. Vertical clustering

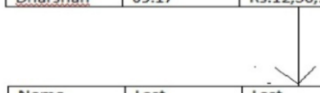
4.2 K-anonymity Algorithm:

The k-anonymity models have been implemented for protecting individual identification. Ongoing investigations show that a more refined model is important to secure the relationship of people to sensitive Quassi data. In Figure 4, generally, protection is estimated by the data gain of a spectator. Prior to seeing the delivered table, the spectator has some earlier conviction about the delicate trait estimation of a person [5]. Our methodology is that we separate the data acquire into two sections: that about the entire populace in the delivered information and that about explicit people. The following expressions are very difficult for understanding the remaining of the project: Quasi-identifier (QI): a bunch of characteristics which can be utilized with particular external data to distinguish a particular person. T, T[QI]: T is the given dataset represented in a relational form T[QI] is the protrusion of T to the arrangement of characteristics contained in QI. Tk [QI].

In this module, we can implement K-Anonymity for secure the data privacy. Some secret data maintains a property called K-anonymity to hide the private data. The two main types of methods for maintaining k-anonymity for all the values of k. The Suppression technique, particular estimations of the credits are replaced by special characters '*'. Suppose all or may be a few estimations of a segment might be replaced by '*'. In Generalization technique the replacement with the most common category is done in this method, so that each and every value of attributes are replaced by with a broader category. We tracked down that a speculation testing approach gave the best authority over re-ID hazard and decreases the degree of data misfortune contrasted with pattern k-anonymity [6]. Certainly, The guarantee of achieving k-anonymity with the procedure of replacing every selected cell with a special character *, but this will result in leaving the database with no meaning. The expense of K-Anonymous answer for a data set is the quantity of *'s presented. A base expense k-anonymity arrangement chooses the least number of cells important to ensure k-anonymity.

Original data

Name	Last Transaction time	Last Transaction amount	Account number	Account type
Anbu	18.00	Rs.22,000	1468875759	Savings
Baskar	22.12	Rs.22,56,000	8359997888	Current
Chithra	08.56	Rs.17,00,000	9531987597	Savings
Dharshan	09.17	Rs.12,56,269	3852897597	Current



Name	Last Transaction time	Last Transaction amount	Account number	Account type
Anbu	18.00	Rs.22,000	14********	Savings
Baskar	22.12	Rs.22,56,000	83********	Current
Chithra	08.56	Rs.17,00,000	95********	Savings
Dharshan	09.17	Rs.12,56,269	38********	Current

Anonymized data

Figure 4. Data anonymization example

5. Conclusion

The main goal is to protect the personally identifiable information i.e. Quassi identifier in data privacy which is most wanted in online banking web application. Generally, if the information is linked directly or indirectly to a person or individual which they possess is called personally identifiable information. Thus, at the point when individual information are exposed to mining, the attribute esteems related with people are private and should be shielded from information theft. Instead of learning from the characteristics of a single individual, miners may learn from global models. In this paper, we can conclude that the proposed system provide improved security in cloud data. Vertical partitioning and K-Anonymity are two approaches that we can use. In data mining, K-Anonymity is a privacy-preserving technique for preventing the disclosure of private information. When anonymizing a database table, the procedure usually entails generalizing table entries, which results in the loss of relevant data.

References

[1] Chahar, H., Keshavamurthy, B. N., & Modi, C. (2017). Privacy-preserving distributed mining of association rules using Elliptic-curve cryptosystem and Shamir's secret sharing scheme. Sādhanā, 42(12), 1997–2007. https://doi.org/10.1007/s12046-017-0743-4

[2] Gunawan, D., & Mambo, M. (2018). Set-valued Data Anonymization Maintaining Data Utility and Data Property. Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication, 1–8. https://doi.org/10.1145/3164541.3164583

[3] Xiong, X., Chen, F., Huang, P., Tian, M., Hu, X., Chen, B., & Qin, J. (2018). Frequent Itemsets Mining With Differential Privacy Over Large-Scale Data. IEEE Access, 6, 28877–28889. https://doi.org/10.1109/ACCESS.2018.2839752

[4] Wang, J., Deng, C., & Li, X. (2018). Two Privacy-Preserving Approaches for Publishing Transactional Data Streams. IEEE Access, 6, 23648–23658. https://doi.org/10.1109/ACCESS.2018.2814622

[5] Wang, H., He, D., & Tang, S. (2016). Identity-Based Proxy-Oriented Data Uploading and Remote Data Integrity Checking in Public Cloud. IEEE Transactions on Information Forensics and Security, 11(6), 1165–1176. https://doi.org/10.1109/TIFS.2016.2520886

[6] Liang, K., Huang, X., Guo, F., & Liu, J. K. (2016). Privacy-Preserving and Regular Language Search Over Encrypted Cloud Data. IEEE Transactions on Information Forensics and Security, 11(10), 2365–2376. https://doi.org/10.1109/TIFS.2016.2581316