A Robust Privacy Preserving Approach of Outsourced Data by Modified Frequent Web Access Pattern

Tasneem Jahan
Truba Institute of Engineering and Information
Technology, India

Amit Saxena
Professor, Head of Department, CSE
Truba Institute of Engineering and Information
Technology, India

ABSTRACT

Current scenario of large databases is in point of fact a major issue. Although, the conventional information examination seems to deal the extensive amounts of information. But the data analysts also attempt to analyze the productivity of data. This proposed work is an attempt to resolve the issue of digital information security by finding the highly frequent items in the dataset. Modified Frequent Web Access Pattern algorithm was developed in this work which find patterns in two scans. Technique called as super class substitution will be used here for perturbation of sensitive set of rules. It offers an added advantage of reducing the risk and the utility of database is also increased. Our experiment is carried out on a genuine dataset. The outcomes here, have shown that proposed work has better results over the previous methodologies.

Keywords

Data Mining, PPDM, MFWAP, Super class substitution, Data Perturbation,

1. INTRODUCTION

The requirement for mining the information with security conservation has grown for maintaining consistent data flow through the system by not leading to any mishandling of sensitive data. Internet Phishing is often used to capture user's data, such as usernames, passwords and sometimes other confidential data. Internet phishing has shown to cause a potential damage from financial strikes to stocks breakdown around the globe. Provided the existence of advanced cryptographic methods, internet banking often encounters intrusions in payment gateways [1, 2]. Consequently, the enhanced and efficient information mining techniques with added security are the emerging need for secure data trade over the system. Nowadays, client's data are swiped out to third party database servers, so as to achieve distributed computing paradigm, but this must also ensure that security isn't compromised. So far, the progress carried out in data mining has produced excellent outcomes for the perception of privacy preserving information mining. The security is an aggregated term inclusive of the dimensions of mining components, i.e. -clustering, association control, and order [3, 13].

When we talk about Distributed computing, it is a model that enables business collaborators with an advantage of storing the information of all partners at distributed locations. This also arises the solutions that even the individual information is being gathered and processed with effective mining plans and there are no breaches in security. Mining must also incorporate the efficiency in usage of information and users should be benefitted in terms of order of protection. The security and protection schemes used in information mining implements the techniques such as, cryptography, buildup, K-secrecy, L-diversity, randomization techniques [8, 9]. The Privacy Preserving Data Mining strategies secure the information by hiding or concealing some sensitive data with the goal that private data isn't easily accessible. This offers a tradeoff between exchange of information with secrecy and productivity. PPDM also offers better implementation, because the utilization of cryptographic strategies to control data spillage would cost high computational expenses [4, 6].

2. RELATED WORK

N. Muthu Lakshmi and K. Sandhya Rani [9] explored the vertically partitioned databases and had proposed a model to identify association rules. It considers 'n' number of sites, and calculate the protection imperatives alongside information data miner. This model utilizes diverse cryptography strategies, for example - encryption, decoding and scalar item system to discover association rules for enhancing productivity and safety of vertically divided databases.

F. Giannotti et al. [10] proposed a solution which depends on k-anonymity frequency. To counter frequency investigation from intruder, the information proprietor embeds fake exchanges in the database to reduce the object frequency. Objects in the database are encoded with the 1-1 substitution words. When the fake exchanges takes place, any object in the perturbed database will have a similar frequency with in any event $k - 1$ different objects. At that point data proprietors outsource their database to the server for the mining assignment. The server runs itemset mining calculation and returns with about regular itemsets and their backings to the information proprietor. The information proprietor modifies these itemsets' backings by subtracting them with itemsets' relating event check in the fake exchanges separately. Then, the information proprietor decodes the received itemsets with the amended backings higher than the frequency limit and produces association rules in view of the incessant itemsets. In these process, the information proprietor needs to include itemset events to counteract fake exchanges. Utilizing this strategy for the vertically parceled database, information proprietors can't perform much computations.

J. Lai et al. [11] proposed a protection saving outsourced association pattern mining arrangement. This arrangement is powerless against frequency examination attacks. Applying this answer for vertically partitioned databases will bring about the leakage of the correct backings to information proprietors.

T. Tassa [12] proposed for secure mining of association runs on a level plane disseminated databases. It performs a quicker calculation by Apriori algorithm. The convention registers the union (or crossing point) of private subsets that each of the intriguing site hold. Likewise, the convention tests the incorporation of a component hold by one site in subset held by another. In any case, this arrangement is appropriate for level dividing, not for vertical apportioning.

Lichun Li et al. [14] proposed a security protecting association run digging answer for outsourced vertically divided databases. In such a situation, information proprietors wish to take in the association administers or regular itemsets from an aggregate informational index and unveil crude (sensitive) information as conceivable to other information proprietors and outsiders. Symmetric homomorphic encryption procedure is utilized for calculation of help and it certainty guarantees the security of the information and mining results.

3. PROPOSED WORK

Work presented in this paper is a combination of two steps where first includes site creation and second includes distribution of columns on various sites. The encryption on rows is performed and saved on the sites. Explanation of the work is shown in Fig. 1.

Pre-Processing

Pre-Processing: As the initial dataset may contain many unnecessary information which needs to be removed for making proper operation. Hence the data is read and arranged in the form of matrix.

Modified Frequent Web Access Pattern

In this step, transactions which arrive in the dataset are passed in the network so that various combination of the items in the transaction are counted in each pass.

Main advantage of this proposed algorithm was that it count the patterns in just two passes of the dataset. In the first pass, various sets of patterns are identified which are present in dataset. So for the number of different items present in the dataset number of patterns are collect. This can be understand as:

Table 1 Session present in log

Sessions	Patterns
S1	a1, a2, a3, a4
S2	a1, a4, a3, a2
S3	a4, a3, a2
S4	a4, a2, a7, a5
S5	a1, a7, a5

From the above table, number of different patterns are {a1, a2, a3, a4, a5, a6, a7}. Now all possible combination of the items are created in our experiment. It was assumed that {a1, a2} is same as {a2, a1}, this act as symmetric property of the items. In this way we find all set of possible patterns present in the current dataset.

Now in second phase all set of possible patterns will be a tree node, where root node is null and other nodes act as the pattern count. In this MFWAP all elements are passed in the model in order of the pattern id so that symmetric property will be maintained. This can be understand by passing session S1 = {a1, a4, a3, a2} as {a1, a2, a3, a4} in the tree so this increase count of all node present in left of tree by one.

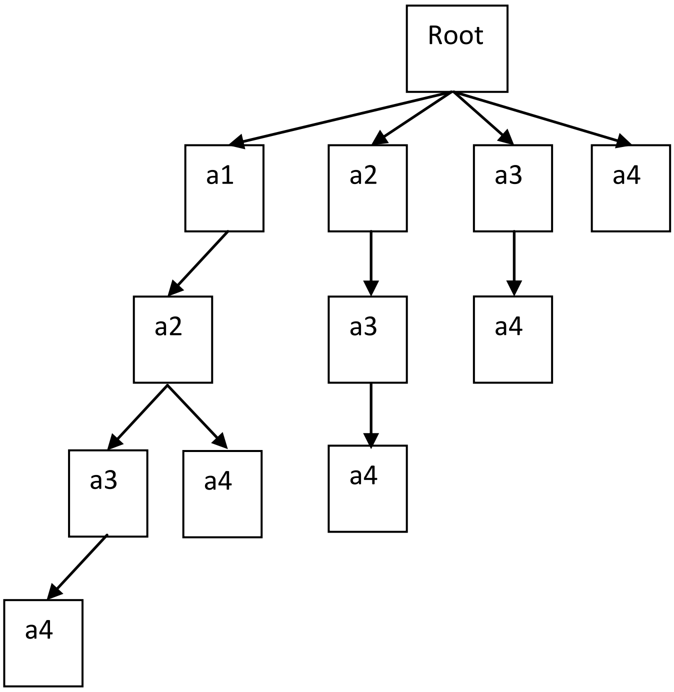

Fig. 1 Modified Frequent Web Access Pattern Diagram

Filter Sensitive Rule

Now from the generated rule, we can create bunch of rules and then we separate those rules from the collection into sensitive and non- sensitive rule set. Those rules which cross sensitive threshold are identified as the sensitive rules while those not containing are indirect rules. This can be understood as, let A, B →C where this pattern cross minimum threshold value so this rule is sensitive rule. If D, B→ C is a rule and is not crossing sensitive or minimum threshold then this rule is not sensitive rule.

Random Distribution

$$f(x \mid p) = \sum_{i=0}^{x} pq^{x}$$

Where q = 1 – p, x is a matrix of numbers while p is a probable value for the generation of series.

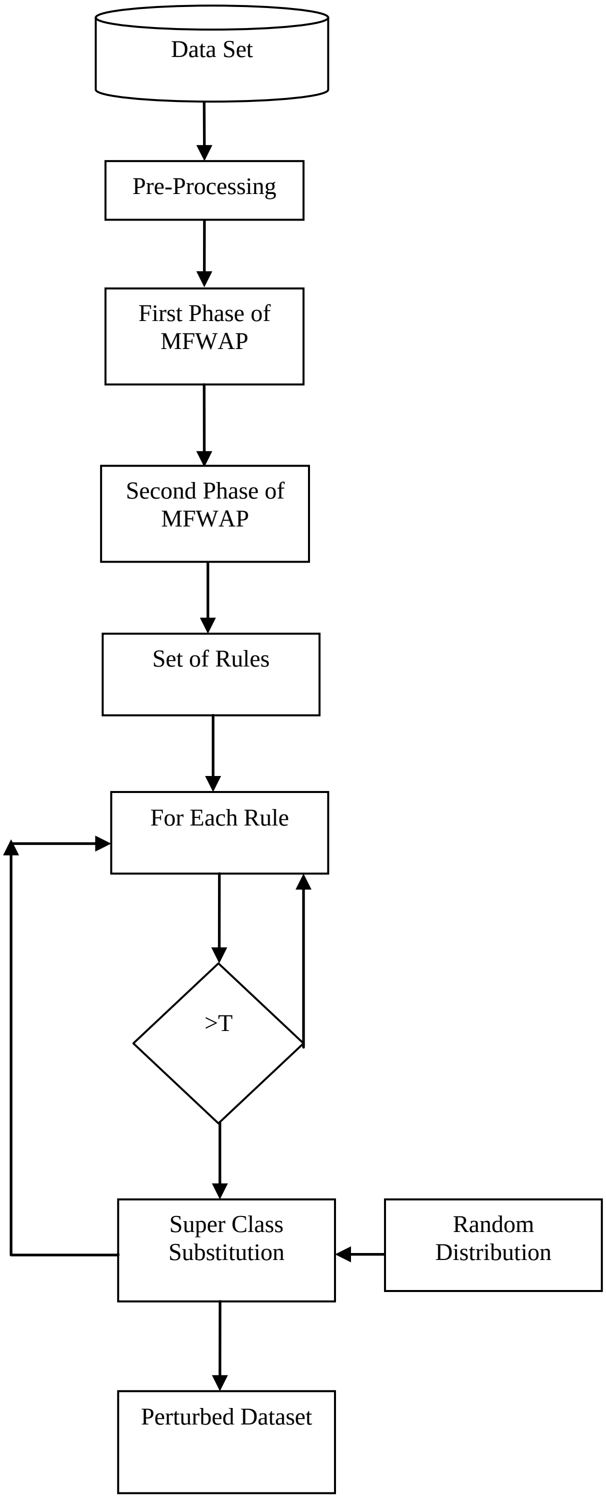

Fig. 2 Block Diagram of Proposed Work Rule Generation

In our approach, we will change the original dataset at random positions, where amount of change depends on the minimum threshold. The original values will not be in the same order as were in the original dataset. In [10] noise is generated by a Gaussian function that produce a sequence of number then add those sequences in the original position, and hence a kind of variation is developed for keeping the privacy of the original one, but it is limited to the numeric value only.

Sensitive Pattern Hiding:
Now in order to hide pattern, {X, Y), we will decrease its support to be lesser than user-provided minimum support transaction (Mini_Supp). In order to decrease the support value the approach is to lessen the support of the item set {X, Y}. So number of session required to decrease the support are calculated by formula:

$$Perturb\_session = \left(\frac{(Rule\_Supp - Mini\_Supp) \times |D|}{100} \right)$$

Where |D| is dataset size, Rule_Supp is Support of pattern.

Super Class Substitution

In this step, all multi attributes are replaced by their hierarchy value in the super modularity tree. While replacing, it is required to balance the dataset utility and risk by making required changes. This replacement is so designed that utility of the data get increase while risk remain below under some threshold value.

4. EXPERIMENT AND RESULT

Dataset

In our proposed work, we shall use a real-world dataset called chess [1], to analyze the proposed algorithm. This data set consists of 3196 records with 37 attributes (without class attribute). So in order to provide protection against the private data of the customer, we will utilize the concept of super modularity, which makes multiple copy of the same dataset with different values.

Evaluation Parameters

Originality:
This specifies the percentage of the privacy provided by the technique that we have adopted. Here total number of cells are counted which are originally passed without any changes.

$$Originality = \frac{\sum Same\_cell}{Total\_cell}$$

Utility:
Under this parameter, the summation of information is done where highest subclass gets higher value of utility. Every set of attribute are mapped to different set of subclass, so utility of sharing information vary as per value pass in the perturbed dataset.

$$U = \log \frac{U(i,j)}{j}$$

Results

Table 2 Comparison of Execution time of Previous and Proposed work

Dataset size	Ant Colony	MFWAP
1000	49.4308	0.2657
2000	135.2244	0.2383
4000	126.6372	0.7333

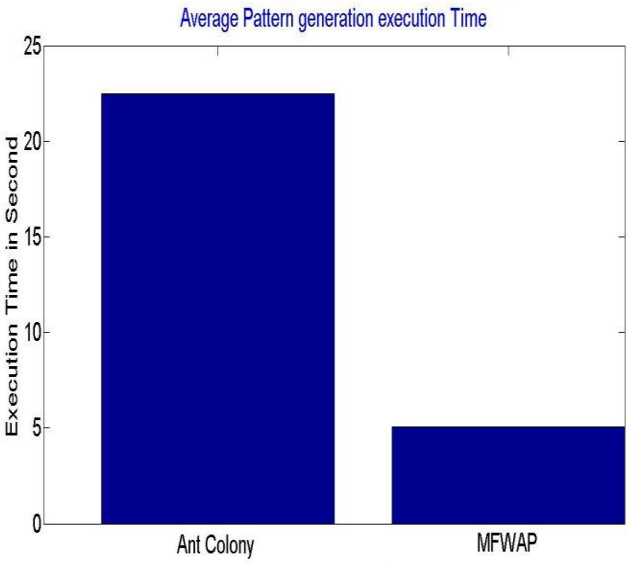

Fig. 3 Average execution time of ant colony and MFWAP

From Table 2 and Fig. 3, it was obtained that proposed work has provided the privacy in lesser time as compared to previous work [15]. Here use of MFWAP has reduced the pattern finding time as previous work use ant colony for the same.

Table 3 Comparison of Perturbed Dataset Size between Previous and Proposed work

Dataset size	Ant Colony	MFWAP
1000	966	1000
2000	1967	2000
4000	3980	4000

From Table 3 it was obtained that proposed work has maintained the same dataset size as passed in the beginning while previous work [15] reduced the size of dataset. Here the use of superposition concept increases utility of the dataset as previous work used deletion of the suspected session identified by ant colony for the same.

Table 4 Comparison of Utility of Previous and Proposed work

Dataset size	Ant Colony	MFWAP
1000	637.2158	867.9368
2000	1219.6	1568.7
4000	2168.8	2461.4

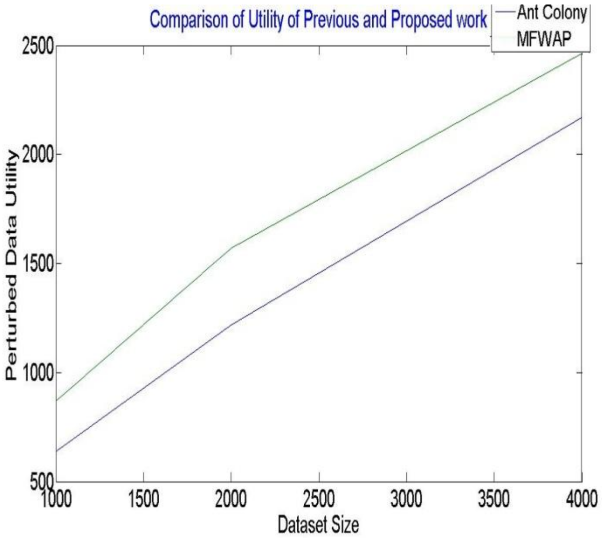

Fig. 4 Perturbed dataset utility comparison of MFWAP and Ant colony

From Table 4 and Fig. 4, it was obtained that proposed work has provided high utility of the dataset in less time as compared to previous work [15].

Table 5 Comparison of Pattern generation execution time of Previous and Proposed work

Dataset size	Ant Colony	MFWAP
1000	15.235	3.7607
2000	22.6321	3.2551
4000	29.5076	8.1260

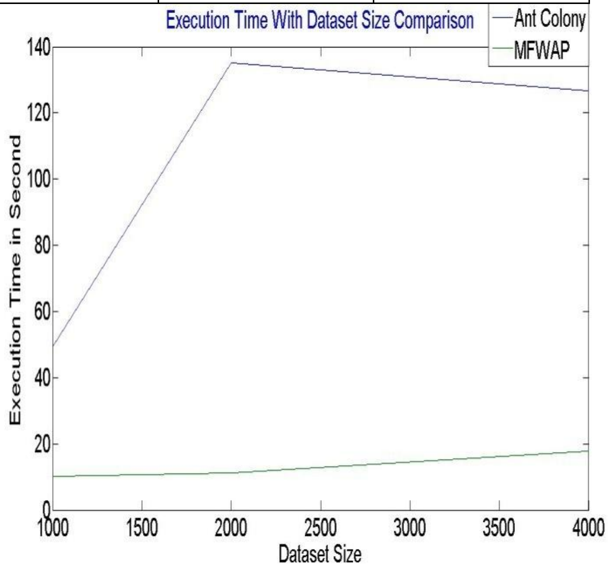

Fig. 5 Comparison of Pattern generation execution time of Previous and Proposed work

From above Table 5 and Fig. 5, it was obtained that proposed work has lesser execution time as previous work.

5. CONCLUSION

This paper has proposed an information distribution algorithm with high privacy of data at various servers. By the utilization of MFWAP and super class substitution security of the information at server side gets upgrades. Results have shown that in proposed work execution time gets decreased. By the utilization of programmed vertical example space, cost get additionally diminished. As research gets never end, so in future one can embrace other example method for enhancing the server execution.

6. REFERENCES

[1] R..Agrawal and R..Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases, pp. 487-499, 1994.

[2] T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277-292, 2010.

[3] F. Kamiran and T. Calders, "Classification with no Discrimination by Preferential Sampling," Proc. 19th Machine Learning Conf.Belgium and The Netherlands, pp 1-6, 2010.

[4] Huhtala, Y., Karkkainen, J., Porkka, P., and Toivonen, H., (1999), TANE: An Efficient Algorithm for discovering Functional and Approximate Dependencies, The Computer Journal, V.42, No.20, pp.100-107.

[5] Lichun Li, Rongxing Lu, Kim-Kwang Raymond Choo, Anwitaman Datta, and Jun Shao. "Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases". IEEE Transactions On Information Forensics And Security, Vol. 11, No. 8, August 2016 1847

[6] Shyue-liang Wang, Jenn-Shing Tsai and Been-Chian Chien, "Mining Approximate Dependencies Using Partitions on Similarity-relation-based Fuzzy Databases", IEEE International Conference on Systems, Man and Cybernetics(SMC) 1999.

[7] Yao, H., Hamilton, H., and Butz, C., FD_Mine: Discovering Functional dependencies in a Database Using Equivalences, Canada, IEEE ICDM 2002.

[8] Wyss. C., Giannella, C., and Robertson, E. (2001), FastFDs: A Heuristic-Driven, Depth-First Algorithm for Mining Functional Dependencies from Relation Instances, Springer Berlin Heidelberg 2001.

[9] N. V. Muthu Lakshmi1 & K. Sandhya Rani, "Privacy Preserving Association Rule Mining in Vertically Partitioned Databases," In IJCSA, vol. 39, no. 13, pp. 29-35, Feb. 2012.

[10] F. Giannotti, L. V. S.Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases," IEEE Syst. J., vol. 7, no. 3, pp. 385- 395, Sep. 2013.

[11] J. Lai, Y. Li, R. H. Deng, J. Weng, C. Guan, and Q. Yan, "Towards Semantically Secure Outsourcing of Association Rule Mining on Categorical Data," Inf. Sci.., vol. 267, pp. 267-286, May 2014.

[12] T. Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases Scalable Algorithms for Association Mining," IEEE Trans.Knowl. Data Eng., vol. 26, no. 4, Apr. 2014.

[13] Thasneem M, S.Ramesh, Dr. T. Senthil Prakash. "An Effective Attack Analysis and Defense in Web Traffic Using Only Timing Information". International Journal of Scientific Research & Engineering Trends Volume 3, Issue 3, May-2017, ISSN (Online): 2395-566X, www.ijsret.com

[14] L. Li, R. Lu, S. Member, K. R. Choo, and S. Member, "PrivacyPreserving-Outsourced Association Rule Mining on Vertically Partitioned Databases," IEEE Trans. Info. Foren. Secur., vol. 11, no. 8, pp. 1847–1861, Aug. 2016.

[15] Jimmy Ming-Tai Wu, Justin Zhan, And Jerry Chun-Wei Lin. "Ant Colony System Sanitization Approach to Hiding Sensitive Itemsets". Digital Object Identifier 10.1109/ACCESS.2017.2702281 June 28, 2017