International Journal of Current Trends in Engineering & Technology
Received 1 December 2018, Received in revised form, 20 January 2018, Accepted 15 Feb. 2018, Available online 25 Feb. 2018
www.ijctet.org, ISSN: 2395-3152
Volume: 05, Issue: 01 (January- February, 2019)

A Robust Privacy Preserving of Outsourced Data by Frequent Web Access Pattern and Substitution

Deepa Agrawal, Jayshree Boaddh, Akrati Shrivastava
Mittal Institute of Research Technology, Bhopal
deepaagrawalcs@rediffmail.com
+91, 6260801190

Abstract— *This period of huge database is currently a major issue. Although, the conventional information examination will most likely be unable to deal with such extensive amounts of information. So analysts attempt to build up a selected stage to productively investigate and keep up the calculations easy. Here proposed work has resolved this issue of digital information security by finding the highly frequent items in the dataset. Frequent Web Access Pattern and Substitution algorithm was developed in this work which find patterns in two scan. Here sensitive set of rules are perturbed by super class substitution. This type of substitution reduces the risk and increase the utility of the dataset as compared to other methods. Analysis is done on genuine dataset. Results demonstrates that proposed work is better as contrast with different past methodologies on the premise of assessment parameters.*

Keywords: Distributed Data, Data Mining, Encryption, Effective Pruning, Super class substitution.

I. Introduction

The requirement for information mining with security conservation has developed as an interest for trading sensitive data previously discharging information over the system. Additionally, the suspicious methodologies, and refusal of the information providers towards the assurance of data. Internet Phishing is an ill-conceived approach to acquire private data, for example, usernames, passwords, and charge card points of interest by disguising as a dependable substance in an electronic correspondence. In this manner, expanded online assurance against phishing attacks is a region of colossal intrigue. As these attacks are advanced in nature, they represent a few difficulties as far as shirking techniques. Internet phishing prompted a few security and financial strikes on the clients and undertakings around the world. Web payment gateways of internet banking have suffered and prompted generous money related misfortune [1, 2]. Consequently, enhanced information mining techniques with security are the need of great importance for secure data trade over the system. These days, putting away clients' data has an obligation with the end goal that their security isn't damaged. Among a few existing calculation, the Data Mining with protection produces outstanding outcomes identified with the inside perception of privacy preserving with information mining. The security should be consolidated onto all mining components including clustering, association control, and order [3, 13].

Distributed computing enabled the business collaborators to store the information for the advantages of all partners. This has prompted gather clients' individual information and nourished into information mining plans which ought to guarantee that there is no loss of protection. Furthermore, the elements like usage, order of protection regarding its benefits and negative marks are not been audited legitimately. A few protection safeguarding plans in information mining exists which incorporate K-secrecy, cryptography, buildup, L-diversity, randomization, techniques [8, 9]. The PPDM strategies secure the information by concealing some unique data with the goal that private data isn't uncovered. The design is to adjust an exchange off amongst secrecy and productivity. The utilization cryptographic strategies dependably have computational expenses to avoid data spillage [4, 6].

II. RELATED WORK

N. Muthu Lakshmi and K. Sandhya Rani [9] proposed a model to discover association rules for vertically divided databases considering the protection imperatives with 'n' number of sites alongside information data miner. This model compromises diverse cryptography strategies, for example, encryption, decoding and scalar item system to discover association runs productively and safely for vertically parceled databases.

F. Giannotti et al. [10] proposed an answer which depends on k-anonymity frequency. To counter frequency investigation 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. In the wake of embeddings the fake exchanges, any object in the perturbed database will have a similar frequency with in any event k − 1 different objects. At that point dada proprietors outsource their database to the server for the mining assignment. The server runs visit itemset mining calculation and returns the came 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. At that point, the information proprietor decodes the got itemsets with the amended backings higher than the frequency limit and produces association rules in view of the incessant itemsets. In these setting, information proprietor requires including itemset events fake exchanges to counteract fake exchanges. Utilizing this strategy for the vertically parceled database, information proprietors can't perform such 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 apportioned databases will bring about the leakage of the correct backings to information proprietors.

T. Tassa [12] proposed for secure mining of association runs in on a level plane disseminated databases. The proposed convention depends on the quick conveyed calculation, which is an unsecured dispersed variant of Apriori calculation. 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 as meager data about their (sensitive) crude information as conceivable to other information proprietors and outsiders. Symmetric homomorphic encryption procedure is utilized for calculation of help and certainty which guarantees the security of the information and mining result moreover.

III. Proposed Work

Whole work is a combination of two steps where first include site creation while second include distribution of columns on various sites. While transferring whole row encryption was performed on the them to save on the sites. Explanation of whole work is shown in fig. 1.

Pre-Processing:- As the dataset is obtain from the above steps contain many unnecessary information which one need to be removed for making proper operation. Here data need to be read as per the algorithm such as the arrangement of the data in form of matrix is required.

Frequent Web Access Pattern :- In this step transaction comes in the dataset are pass in the network such that various combination of the items in the transaction are count in this pass. Main advantage of this proposed algorithm was that it counts the patterns in just two pass of the dataset. In first pass various set of patterns are identified which are present in dataset. So as per the number of different items present in the dataset number of patterns are collect. This can be understood 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

So from above table number of different patterns are {a1, a2, a3, a4, a5, a6, a7}. Now all possible combination of the items is created in the work, it was assumed that {a1, a2} is same as {a2, a1}, this act as symmetric property of the items. In this way one finds all set of possible patterns present in the current dataset.

Now in second phase all set of possible patterns are act as the tree node where root node is null and another node act as the pattern count. In this FWAPS all element is pass in the model in order of the pattern id so that symmetric property will be maintained. This can be understood 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.

Filter Sensitive Rule:-Now from the generated rule one can get bunch of rules then it is required to separate those rules from the collection into sensitive and non- sensitive

International Journal of Current Trends in Engineering & Technology
Received 1 December 2018, Received in revised form, 20 January 2018, Accepted 15 Feb. 2018, Available online 25 Feb. 2018
www.ijctet.org, ISSN: 2395-3152
Volume: 05, Issue: 01 (January- February, 2019)

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 the 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 not cross sensitive or minimum threshold then this rule is not sensitive rule.

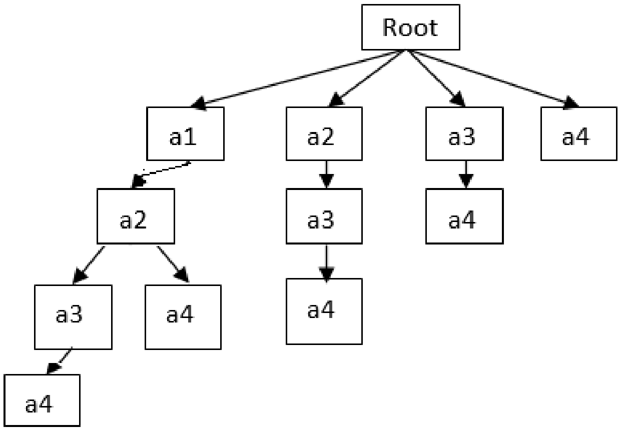

Fig. 1 Modified frequent web access pattern diagram

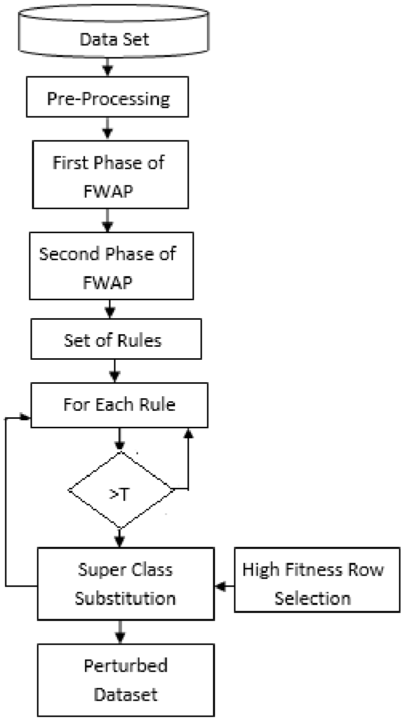

Fig. 2 Block diagram of proposed work rule generation.

Row Fitness Selection:- In this step all rows are selected first which have that rule items. Now find fitness of each row on the basis of information or item present in each row. As per the information present select rows for perturbation who have high information. So due to small change in loss of information is less.

In this approach original dataset is change in random portion where amount of change is depending on the minimum threshold. The original values but not in the same order as was in the original dataset. In [10] noise is generated by a Gaussian function that produce a sequence of number then add those sequence in the original position, so a kind of variation is develop over here for the privacy of the original one, but that was limited to the numeric only.

Sensitive Pattern Hiding:- So in order to hide pattern, {X, Y}, this work can 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 calculate by below 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 whole multi attributes are replace by its 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.

IV.EXPERIMENT AND RESULT

In order to analyze proposed algorithm, it is in need of the dataset. There work uses real-world datasets called chess [1]. This data set consists of 3196 records. The data set has 37 attributes (without class attribute).

Table 2 Comparison of Execution time of Previous and Proposed work.

Dataset size	Ant Colony	FWAPS
3500	39.2224	2.567
5500	55.1239	2.337
7500	67.8972	1.694

From above table 2 it was obtained that proposed work has provide the privacy of the work in less time as compared to previous work [15]. Here use of FWAPS increase reduce the pattern finding time as previous work use ant colony for the same. From above table 3 it was obtained that proposed work has maintained the same dataset size as pass in the beginning while previous work [15] reduce the size of dataset. Here use of superposition concept increase utility of the dataset as previous work use deletion of the suspected session identified by ant colony for the same.

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

Dataset size	Ant Colony	FWAPS
3500	3362	3500
5500	5440	5500
7500	7482	7500

Table 4 Comparison of Information present in Dataset after applying Previous and Proposed work.

Dataset size	Ant Colony	FWAPS
3500	2162.5	2475.5
5500	2171.7	2418
7500	2201.6	2360.7

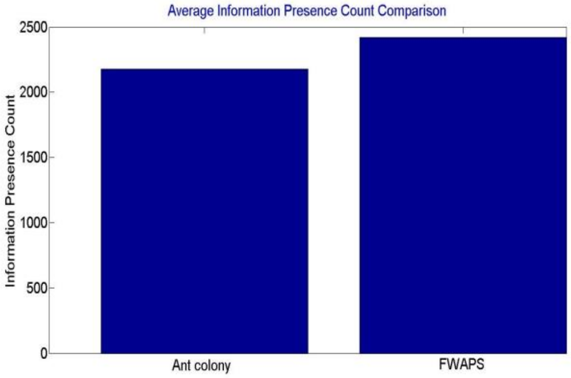

Fig. 3 Average information count comparison.

Table 5 Comparison of pattern generation execution time of Previous and Proposed work.

Dataset size	Ant Colony	FWAPS
3500	23.4823	4.0454
5500	37.75	4.4201
7500	15.6338	4.4583

From above table 4 it was obtained that proposed work has provide high information of the dataset in less time as compared to previous work [15]. Here use of substitution concept increase utility of the dataset as previous work use deletion of the suspected session identified by ant colony for the same.

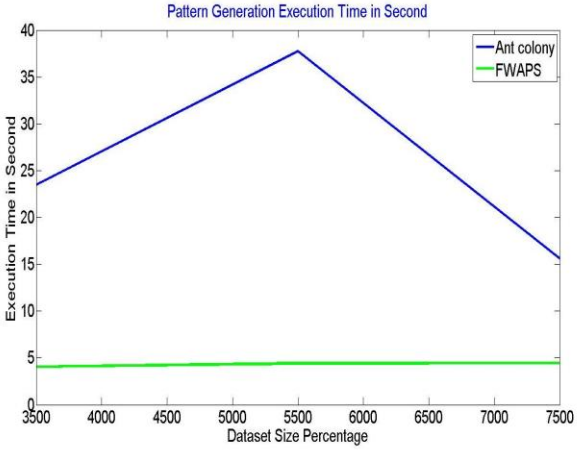

Fig. 4 Pattern generation time comparison of Ant colony and FWAPS methods.

From above table 5 it was obtained that proposed work has provide the privacy of the work in less time as compared to previous work [15]. Here use of FWAPS increase reduce the pattern finding time as previous work use ant colony for the same.

V. Conclusion

As scientists are chipping away at various field out of which finding a powerful vertical examples are measure issue with this becoming advanced world. This paper has proposed an information distribution algorithm with high privacy of data at various servers. By the utilization of FWAPS and super class substitution security of the information at server side get upgrade too. Results demonstrates that proposed work execution time get decrease. While batch passed time get decrease. By the utilization of programmed vertical example space cost is additionally diminished. As research is never end handle so in future one can embrace other example era method for enhancing the server execution.

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), Fast FDs: 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, "Privacy Preserving-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.