International Journal of Engineering & Technology, 7 (1.1) (2018) 19-24

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Security and privacy concerned association rule mining technique for the accurate frequent pattern identification

T. Nusrat Jabeen [1*], M. Chidambaram [2], G. Suseendran [3]

[1]*Department of Computer Science, Bharathiyar University, Coimbatore, Tamil Nadu, India.*
[2]*Department of Computer Science, Bharathiyar University, Coimbatore, Tamil Nadu, India.*
[3]*Assistant Professor, Department of Information and Technology, School of Computing Sciences, Vels University, Chennai, India.*
**Corresponding author E-mail: nusratjabeent@gmail.com*

Abstract

Security and privacy has emerged to be a serious concern in which the business professional don't desire to share their classified transaction data. In the earlier work, secured sharing of transaction databases are carried out. The performance of those methods is enhanced further by bringing in Security and Privacy aware Large Database Association Rule Mining (SPLD-ARM) framework. Now the Improved Secured Association Rule Mining (ISARM) is introduced for the horizontal and vertical segmentation of huge database. Then k-Anonymization methods referred to as suppression and generalization based Anonymization method is employed for privacy guarantee. At last, Diffie-Hellman encryption algorithm is presented in order to safeguard the sensitive information and for the storage service provider to work on encrypted information. The Diffie-Hellman algorithm is utilized for increasing the quality of the system on the overall by the generation of the secured keys and thus the actual data is protected more efficiently. Realization of the newly introduced technique is conducted in the java simulation environment that reveals that the newly introduced technique accomplishes privacy in addition to security.

Keywords: *Privacy, anonymization, security, partitioning, encryption, quality of transaction database.*

1. Introduction

Data mining is typically an inter-disciplinary field with its basis formed in enterprise decision support. Data mining is not dedicated for the analysis of small datasets. It is regarded to be the job of finding fascinating and concealed patterns/data from huge chunks of information in cases where the data is found in databases, data repositories, OLAP or other information present in the repository [1]. Data mining is also involved with a combination of methodologies from several disciplines like database technology, statistics, machine learning, neural networks, fuzzy and rough set theory, knowledge representation, inductive logic programming, information retrieval and etc.

In data mining, the elementary issue is to discover the frequent item set found in the massive database [2]. The mining of frequent item set finds significance in a broad array of application fields like bioinformatics, web usage mining etc. Multiple numbers of diverse algorithms has been introduced for discovering the frequent item set. The Apriori [3] and FP- Growth algorithm [4] are the most widely accepted algorithms used in association rule mining. The Apriori algorithm is basically a bottom-up approach algorithm or level-wise search algorithm. A subset consisting of frequent item set should also be a frequent item set i.e in case {AB} is a frequent item set, then both A and B must be frequent item sets, known as Apriori property.

The FP-growth algorithm is one among the techniques, which avoids the production of a massive number of candidate item sets explored by Han et al [5]. It functions similar to a depth-first search algorithm. In the field of data mining, association rule mining is a widely used and critically examined technique for the discovery of exciting associations between variables in bulky databases. If the data is spread among various locations, then the discovery of the global association rules becomes a challenge since the privacy concerned with the information of the individual site has to be protected. Here, in this research work, a model is introduced in order to discover the global association rules by protecting the privacy of each sites' information while the information gets horizontally divided among n number of sites [6].

2. Related works

There are different distributed data mining algorithms available for discovering the frequent item sets and generating potential association rules. But few algorithms that can operate on reduction framework in distributed system are necessary [7]. The survey done over the earlier research in this domain is explained as follows.

Apriori algorithm was demonstrated by agrawal and srikant that discovers all the frequent item sets and then creates the rules from the frequent item sets and this process gets repeated till no more frequent item sets can be got that needs more time [8]. Fast Distributed Mining (FDM) algorithm was introduced by cheung that searches the support counts and refines all the unnecessary candidates sets [9]. Distributed Decision Miner (DDM) algorithm was presented by Schuster and wolff which related to apriori-based algorithms utilizing shared that determines the set of globally frequent item sets [10].

MLFPT (Multiple Local Frequent Pattern Tree) parallel algorithm was devised by zaiane that is dependent on frequent pattern–growth algorithm (FPGrowth). It does not create the candidate item for frequent item sets, rather it creates multiple frequent pattern trees [11]. ZigZag algorithm was introduced by otey that

follows a shared-nothing architecture and specifies where the data would be spread on various locations initially [12].
D(Distributed)-Sampling algorithm was presented by schuster and wolff that integrates a centralized sampling algorithm and DDM algorithm [13]. Association rule with logical AND operation algorithm was demonstrated by jabbar that, in turn, transforms the database transactions into binary formation and neglects any of the columns that is lesser compared to the threshold to search for a distinct column in (k-1)-item set [14]. Distributed Approximate Mining of Frequent Patterns algorithm was implemented by silvestri and orlando comprising of the exact distributed computation of locally frequent item sets [15].
Extracting Association Rules for Distributed Association Rules (EAR4DAR) Algorithm was suggested by salih that does the extraction of the association rule from local association rules into global association rules over distributed systems. Distributed Frequent Item Mining algorithm was formulated by lamine and ke-chadi that creates diverse clusters and grid environments [16]. In this technical work, the newly introduced system makes use of w-Tabular algorithm and gets over the drawbacks of Apriori and FPGrowth employing reduction framework [17].

3. Secured and privacy aware association rule mining

Secured and privacy based association rule mining is the most crucial research work that is undertaken by several business organization in order to share and then extract the frequent and amusing pattern. Due to this, the development of business can be in a good direction through the extraction of the resourceful information. The chief objective of this research technique is the implementation of the new framework, which can guarantee both the security and privacy aspects of business who want to share the transaction data details with other people in the business. This is guaranteed by the introduction of Security and Privacy aware Large Database Association Rule Mining (SPLD-ARM) framework. In this research framework, Improved Secured Association Rule Mining (ISARM) is brought in, for the purpose of horizontal and vertical partitioning of the data. It is employed to safeguard valuable data for a bigger transaction database considerably. Then privacy preservation is assured by making use of k-Anonymization approaches including suppression based Anonymization and generalization based Anonymization. It is exploited so as to protect the personal or private information. At last, Diffie-Hellman encryption algorithm is devised in order to preserve the sensitive information and for storage service provider to work on the encrypted information. The Diffie-Hellman algorithm is employed for increasing the quality of the entire system through the generation of the secured keys and thus the actual data is preserved with more efficiency. The overall organization of the research methodology is illustrated in figure 1 shown below.
The figure shown above illustrates the overall flow of the new research technique that tries to carry out secured and privacy critical data communication. The new research scheme is explained in detail in the sub sections that follow.

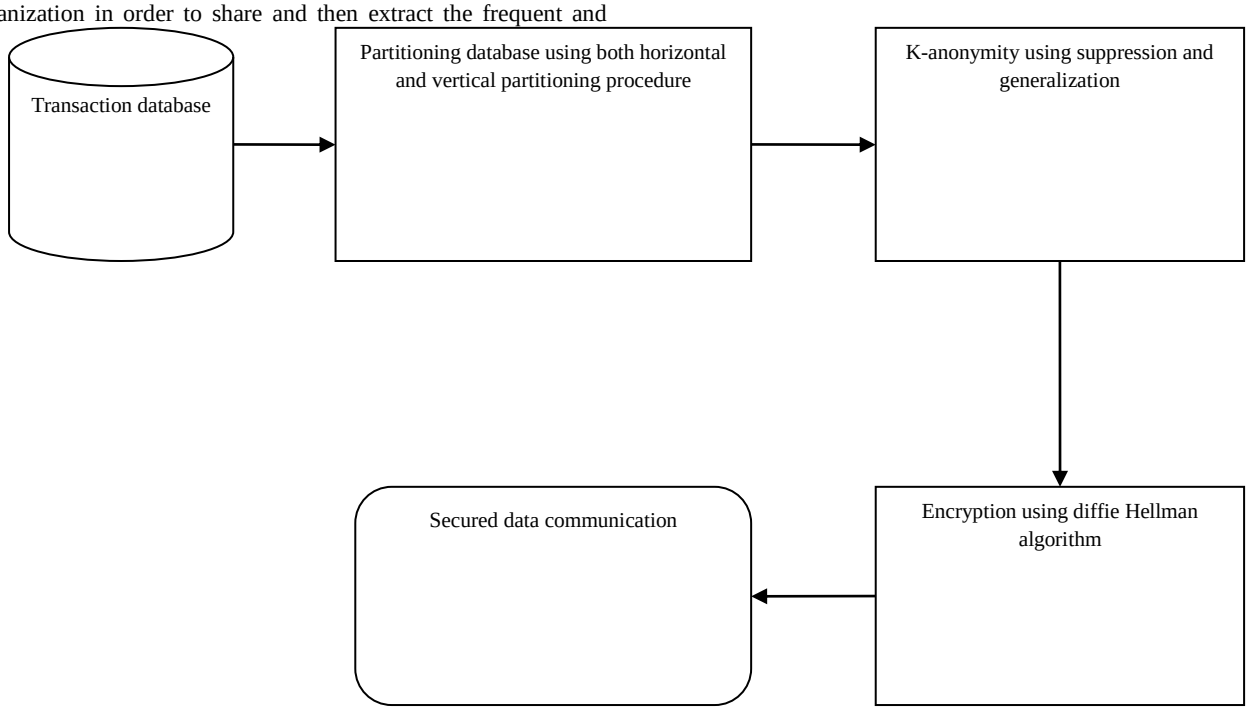

Figure 1: Overall flow of the proposed research method

Secured association rule mining with partitioning to handle large volume of database

The way in which data is partitioned hails to be one among the most vital factors in distributed data mining. Many algorithms on a majority are designed and evolved based on the data partitioning concept. Usually, two kinds of data partitioning exist, namely vertical partitioning and horizontal partitioning. In the case of vertical partitioning, the data that is available is stored at various geographical locations, for instance, assume that in a data mining process, a variety of data like corporate, medical, insurance, hospital, school and housing data have to be collected about various persons who live in the modern city.
Horizontally Partitioned: Horizontal partitioning splits the entire database into several number of smaller database based on the splitting of row. This is done in a way that the executing the query would become quick and also it will have the power to offer more amount of privacy to the partitioned database. Horizontally partitioned data could be exploited in which each fragment consists of a subset of records of R in the form of an association. Horizontal partitioning technique divides a table into different tables. Here, tables are divided in a way similar to the way query references are carried out by using lesser number of tables or large amount of UNION queries are employed to integrate the tables obviously during the moment of query that might, in turn, have an impact over the performance. For instance, assuming that in a project involving data mining it is required to explore the impacts of a drug over those patients who have special kind of illness. Particularly, for the purpose of getting different samples, there is a necessity to get the same data regarding this problem from various medical centers. In these kind of situations, it is stated that the data is horizontally partitioned.

Vertically Partitioned: Vertical portioning is a method that partitions the entire dataset into several number of small databases based on the column, such that the partitioned database does not have any duplicate information. There are primarily two kinds of vertical database namely normalized and row splitting. The data might be split into a set that consists of small files, which are physical, and every individual file comprises of the subset of the actual association, where the association stands for the database transaction that actually requires the subsets of the attributes given. Here, in the case of vertical partitioning, the data concerned with a set of similar entities are located in diverse places, for instance consider that in a data mining procedure it is required to gather various kind of data like financial, medical, insurance and housing data about a variety of individuals living in a city. In this procedure, a diverse amount of data regarding a set of similar entities has to be gathered, i.e. those individuals resident in that city, from the servers of various organizations like medical institutions, government servers, municipalities, banks and so on.
Let Item = {$item_1$, $item_2$, .$item_k$} refer to a set of items and Transaction = {$Trans_1$, $Trans_2$,… $Trans_l$} indicate a set of transactions where every $Trans_l \subseteq$ Item. A transaction $Trans_l$ has an item set X $\subseteq$Item only when X $\subseteq Trans_l$. An association rule implication takes the form X $\Rightarrow$Y(X $\cap$Y = 0) with support s and confidence C when S% of the transactions in T consists of X $\cup$Y and C% of transactions which have X also have Y. In the case of a database that is horizontally partitioned, the transactions are observed to be spread among n sites.

$$\text{Support (X Y)} = \text{Provability(X Y)} / |\text{Total Number of Transaction}|$$

The global support count of an item set indicates the sum of all the local support counts.

$$\text{Support g (X)} = \sum_{i=1}^{n} \text{Support}_i \text{ (x)}$$

$$\text{Confidence of rule (X Y)} = \text{Sup (X Y)} / \text{Sup(X)}$$

The global confidence of a rule shall be provided in terms of the global support.

$$\text{Confidence g } (X \Rightarrow Y) = \text{Support g } (X \cup Y) / \text{Support g(X)}$$

The goal of the privacy preserving association rule mining is discover each one of the rules with global support and global confidence greater than the end-user mentioned minimal support and confidence. The steps mentioned below, using secure sum and secure set union techniques that were explained before are brought into use. The algorithm is based on the Apriori algorithm which makes use of the (k-1) sized frequent item sets for the generation of the k sized frequent item sets. The issue of creating the size 1 item sets could be conveniently surpassed with secure computation over multiple number of sites.

1. Candidate Set Generation: Overlap the globally frequent item set of size (k-1) with locally frequent (k-1) item set to get candidates. From these, use the apriori algorithm to get the candidate k item sets.
2. Local Pruning: For each X in the local candidate set, scan the local database to compute the support of X. If X is locally frequent, it's included in the locally frequent item set.
3. Item set Exchange: Calculate a Secure union of the large item sets over all sites.
4. Support Count: Compute a Secure Sum of the local supports to get the global support.

Privacy preservation using K-anonymity

Anonymization indicates the identification of the information, which is eliminated from the actual data to preserve individual or private data. There exists several means of performing data anonymization. Essentially, this technique employs k-anonymization approach. In case, every row in the table cannot be differentiated from at least other k-1 rows by just searching through a set of attributes, and so this table is called to be K-anonymized on these attributes.
Suppression-based k-anonymization: Suppose the content present in table T = {t_1, t_2, ….,t_n}over the attribute set A. The concept is to generate the subsets of not differentiable tuples through the masking of the values of few selected attributes. Specifically, while making use of a suppression-based anonymization technique, it is masked with the unique value'*'. In this technique, the below notations are used.
Quasi-Identifier (QI): To find a particular individual that contains a set of attributes, which can be employed with particular external data could.
T[QI]: T[QI] refers to the projection of T to the set of attributes present in QI.
Generalization-based k-anonymization: In generalization-based anonymization technique, the actual values are substituted by few more generalized ones present in the database, based on a priori established value generalization hierarchies (VGHs).
Table 1 indicates the real dataset, which contains all the original data in the tuple form. Once the suppression based scheme is applied on the real dataset, it gets anonymized and displaying the anonymized records makes alterations in two QI and therefore the value of k=2 in table 2. So, table 3 reveals the result obtained of the generalized technique by substituting the value after the application of the data mining procedure. The "Data Mining" point could be generalized to a more certain value with "Database Systems". Thus continuing to replace the rest of the values in table with more generalized values, the actual dataset gets anonymized by employing the generalized technique and at last, once T is k anonymous, duplicate tuples can be deleted, and we the resultant set can be called as the witness set of T. Table 4 shows a witness set of Table 3.

Table 1: Original Data set

Area	Position	Salary
Data mining	Associate professor	$90,000
Intrusion detection	Assistant professor	$78,000
Handheld system	Research assistant	$17,000
Handheld system	Research assistant	$15,000
Query processing	Associate professor	$100,000
Digital forensics	Assistant professor	&78,000

Table 2: Suppressed data with k=2

Area	Position	Salary
*	Associate professor	*
*	Assistant professor	*
Handheld system	Research assistant	*
Handheld system	Research assistant	*
*	Associate professor	*
*	Assistant professor	*

Table 3: Generalized Data with k=2

Area	Position	Salary
Database systems	Associate professor	[61k, 120k]
Information security	Assistant professor	[61k, 120k]
Operating systems	Research assistant	[11k, 30k]
Operating systems	Research assistant	[11k, 30k]
Database systems	Associate professor	[60k, 120k]
Information security	Assistant professor	[60k, 120k]

Table 4: Witness set

Area	Position	Salary
Database systems	Associate professor	[61k, 120k]
Information security	Assistant professor	[61k, 120k]
Operating systems	Research assistant	[11k, 30k]

Secured and privacy aware communication using diffie hellman key exchange algorithm

K-Anonymity is typically a strategy used for rendering preservation of privacy by guaranteeing that the information shall not be exhibited to people. The important objective is the protection of personal privacy. In the case of a k-anonymous dataset, in case any identification regarding the information is discovered in the real dataset having k tuples, and then quasi-identifiers are found first i.e. the tuple, which exactly differentiate the provided tuple present in database. After this, MW algorithm is applied for suppression based Approach. This algorithm is used for the identification of quasi-identifiers and a k-partition is computed, basically a group of disjoint subsets of rows, where every subset comprises of at least k rows and the union of these subsets gives the complete table. Then, every record with "*" is replaced. In the case of suppression based technique, diffie Hellman key exchange algorithm is applied for generating the private secure key. Then, AES (Advanced Encryption Standard)algorithm is exploited in order to encrypt and decrypt the information utilizing the key that is generated by means of the diffie Hellman key exchange algorithm. In this methodology, the encrypted data is brought into use and it is not directly related with the real data. If the end-user submits his information then it is encrypted by applying AES and all the data in table is also encrypted applying the same algorithm.
In case the information from user matches with the information in the table, then this tuple would be decrypted and then gets included into the table. In the case of Generalization based methodology, the value present in table is replaced with more generalized values. When the information submitted by the user matches the value that is being substituted by the generalized value, then this record will be replaced by the generalized value and these generalized values get inserted into the table. K Anonymization lets the database to have a suppressed and generalized kind of data so that data gets much more secured. The cryptography method is employed for securing the stored information in the database safe so that the data gets encrypted, saved and can be got back and then decrypted back to the actual one with particular authorization. The Diffie-Hellman Algorithm

Public key encryption scheme based on a commutative encryption function
1. Alice encrypts message M with her key: ka → {M}ka.
2. Alice sends {M}ka to Bob.
3. Bob in his turn encrypts the received message: → {{M} ka}kb
4. Bob sends {{M}ka}kb back to Alice.
5. Alice is able to decrypt the received message due to commutativity
{{M}ka} kb = {{M}b}ka :→ {M}kb
6. Alice sends {M} kb to Bob, who can decrypt it using his key kb→ M

Diffie and Hellman employ a commutative encryption function that is on the basis of discrete logarithm: Suitable prime p and generator g are selected, and it is generalized for each one of the users.

1. Alice chooses a secret random number xa (her private key) and publish ya = g^{xa} (Her public key).
2. Bob does the same with xb secret and yb = g^{xb} public.
3. Alice uses $yb^{xa} = g^{xa\ xb}$ encrypt a message to Bob.
4. Bob uses $ya^{xb} = g^{xa\ xb}$ to decrypt the received message.

In this new, system, AES algorithm and Diffie–Hellman key exchange algorithm is used. AES algorithm is employed for enhancing the quality of the system on an overall. The important cause behind employing AES is that AES operates under three permitted key lengths: 128 bits, 192 bits, and 256 bits. An algorithm begins with a random number, the key and the information that encrypted with it are then scrambled through four rounds of mathematical processes and the system is made more substantially strong. One more algorithm, Diffie–Hellman key exchange algorithm is applied for the exchange of cryptographic keys. This algorithm lets two parties, which are not aware of one another earlier can share a shared key for the purpose communications through the exchange of data over an open network. The curse of dimensionality problem, which was observed in the earlier system is solved. Also by enhancing the effectiveness of the system, it provides a quick response with substantial protection.

4. Experimental results

This section provides the performance analysis for the newly introduced algorithm, Security and Privacy aware Large Database Association Rule Mining (SPLD-ARM) is used and the results of its analysis are given. Target algorithms include WFPMDS [18] for mining the weighted frequent patterns over

the sliding window-based data streams and the earlier innovative algorithms Fp-Growth Algorithm and Elliptic Curve Cryptography (FPG-ECC) [20] and Federation Rule Mining (FRM) [20]. All the algorithms have been written in JAVA language, and their execution was done on 3.33 GHz CPU, 3 GB RAM, and WINDOWS 7 OS environment.For experiments involving runtime and memory utilization, actual datasets that are available at http://fimi.cs.helsinki.fi/data/, Accidents, Pumsb, Retail, and Mushroom, were brought into use. The performance evaluation metrics considered include total runtimes and maximum memory usage.

Graphs shown in Figs. 2 (a)-(b) indicate the results obtained from the runtimes of every algorithm and dataset, in which the windows possess fixed sizes and these values are changed in accordance with every dataset as illustrated in the figures. Fig. 2 (a) provides the results of the Accidents dataset, where SPLD-ARM ensures the most remarkable runtime performance in every case. In addition, the newly introduced algorithm reveals nearly continuous runtime results with no regard to the minimum support threshold, whereas those of the other ones tend to become higher when d is gradually reduced. Fig. 2 (b) illustrates the execution times taken for mining patterns with the Pumsb dataset. These results prove that SPLD-ARM can carry out mining operations with more rapidity in comparison with the earlier cases.

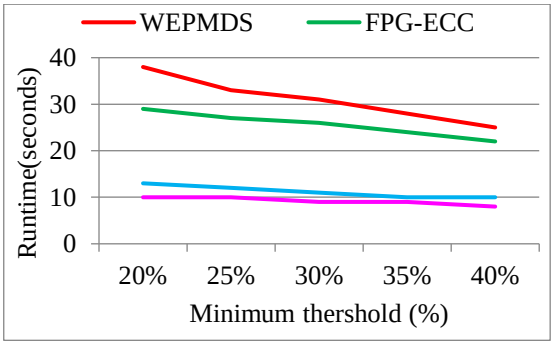

Fig. 2(a): Accidents dataset (W_2)

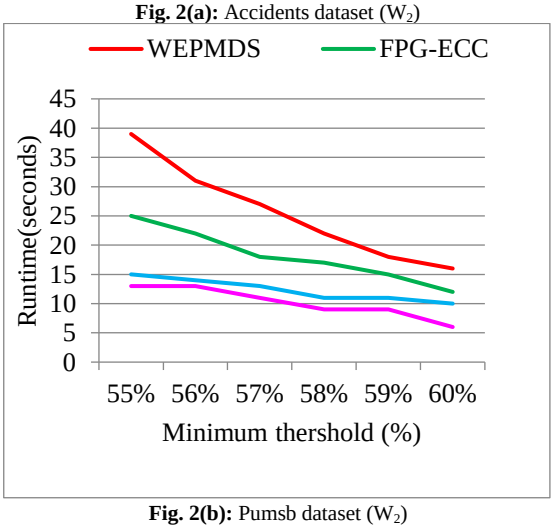

Fig. 2(b): Pumsb dataset (W_2)

Results of memory usage: The experiments conducted in this section show the memory usage results for every original dataset. The parameters used are same as those of the runtime experiments. Fig. 3 exhibits the results of memory usage for the Accidents dataset, in which it can be shown that all the algorithms have stable memory consumption irrespective of d. But, SPLD-ARM needs less memory in every case, and this tendency is shown in a similar manner in remaining experiments. In Fig. 4 for the Pumsb dataset, even though the gap between SPLD-ARM, FRM, FPG-ECC and WMFP-SW and others is lesser compared to that in the Accidents, SPLD-ARM still performs better than the others, as the new SPLD-ARM work closed frequent itemset gets mined in the weighted based frequent mining task.

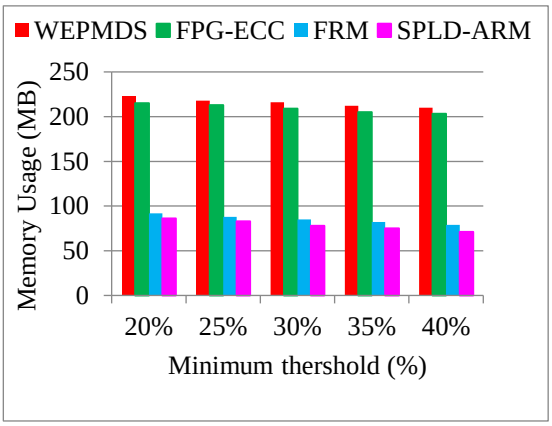

Fig. 3: Accidents dataset (W_2)

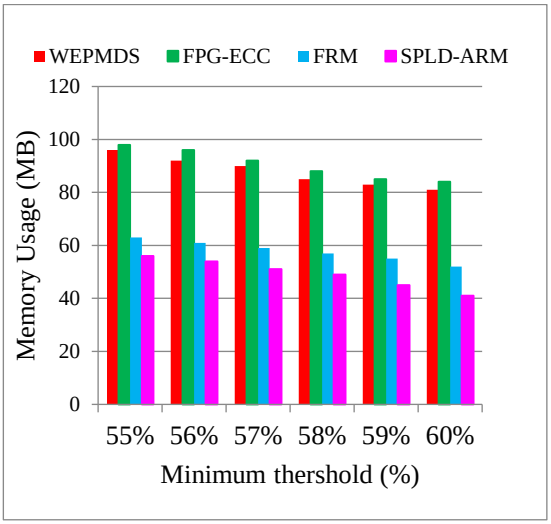

Fig. 4: Pumb dataset (W_2)

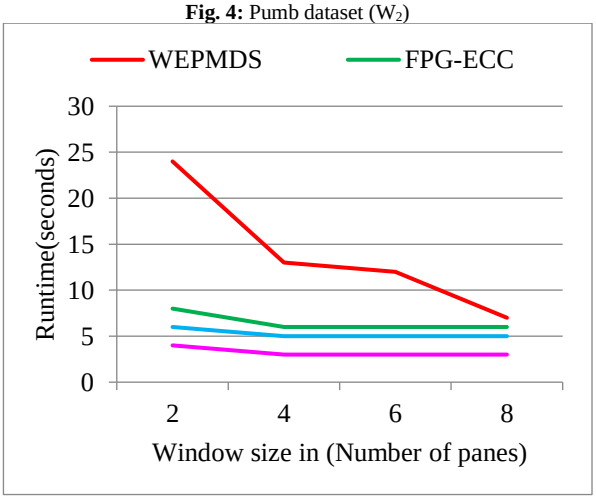

Fig. 5: Runtime of mushroom dataset (ρ = 5%)

Contrasting to the results obtained of the runtime experiments in Figs.4, the runtime usages are slowly reduced as the window sizes tend to become bigger and due to the fact that the sizes of the trees with respect to the current windows increases owing to the rise in the windows.

5. Conclusion

Secured and privacy related association rule mining has become the most cumbersome task practically, where different business organizations try to share their transaction database with one another. Here security has the emerged to be the important concern in which the shared databases might collude with one another and it can get degraded. In this research framework, Improved Secured Association Rule Mining (ISARM) is introduced for partitioning the data, both

horizontally and vertically. It is employed to preserve valuable information for bigger transaction database considerably. Then privacy preservation is guaranteed by making use of k-Anonymization techniques including suppression based Anonymization and generalization based Anonymization. It is applied for the protection of personal or private information. At last, Diffie-Hellman encryption algorithm is presented in order to safeguard the sensitive data and for the storage service provider to work on the data that is encrypted. The Diffie-Hellman algorithm is employed for increasing the quality of the system on an overall by creating the secured keys and thus the actual data is protected more efficiently. The overall analysis of the newly introduced research technique is performed in the java simulation environment, which helps in proving that the newly introduced technique accomplishes privacy and also security in an effective manner that is revealing by well-specified performance metrics. Privacy can be further improved in future by the introduction of the attribute based encryption processes. Moreover, Sensitivity of the transaction data can be incorporated by yielding generalized focus for the respective parameters and then decoding employing modern methodologies.

References

[1] Frawley W, Piatetsky-Shapiro G & Matheus C, "Knowledge Discovery in Databases: An Overview", *AI Magazine, Fall*, pp.213-228, (1992).

[2] Santhi MA, "Application of Data Mining Using Snort rule for intrusion detection", *SSRG International Journal of Computer Science and Engineering*, Vol.1, No.8, (2014).

[3] Agrawal R & Srikant R, "Fast algorithms for mining association rules", *Proc. 20th Int. Conf. Very Large Data Bases*, pp.487-499, (1994).

[4] Agrawal R & Srikant R, "Fast algorithms for mining association rules in large databases", *Proc. 20th VLDB*, (1994).

[5] Han J, Pei J & Yin Y, "Mining frequent patterns without candidate generation", *Proc. ACM SIGMOD Int. Conf. Manage. Data*, pp.1–12, (2000).

[6] Muthu Lakshmi NV & Sandhya Rani K, "Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques", *International Journal of Computer Science and Information Technologies*, Vol.3, No.1, pp.3176–3182, (2012).

[7] Cheung DW, Han J, Ng VT, Fu A & Fu Y, "A fast distributed algorithm for mining association rules", *In Int. Conf. on Parallel and Distributed Information Systems*, pp.31–44, (1996).

[8] Yi X, Rao FY, Bertino E & Bouguettaya A, "Privacy-preserving association rule mining in cloud computing", *Proceedings of the 10th ACM symposium on information, computer and communications security*, pp.439-450, (2015).

[9] Solanki SK & Patel JT, "A Survey on Association Rule Mining", *Fifth International Conference on Advanced Computing & Communication Technologies*, (2015).

[10] Galárraga LA, Teflioudi C, Hose K & Suchanek F, "AMIE: association rule mining under incomplete evidence in ontological knowledge bases", *Proceedings of the 22nd international conference on World Wide Web*, pp.413-422, (2013).

[11] Seol WS, Jeong HW, Lee B & Youn HY, "Reduction of association rules for big data sets in socially-aware computing", *IEEE 16th International Conference on Computational Science and Engineering (CSE)*, pp.949-956, (2013).

[12] Han J, Kamber M & Pei J, "Data mining: concepts & techniques", *Elsevier*, (2011).

[13] Anupriya E &Iyengar N.Ch.S.N., "A framework for optimizing the performance of peer-to-peer distributed data mining algorithms", *International Journal of Computing Science and Communication Technologies*, Vol.3, No.1, (2010).

[14] Le-Khac NA, Aouad L & Kechadi T, "Distributed knowledge map for mining data on grid platforms", *International Journal of Computer Science and Network 98 Security*, Vol.7 No.10, (2007).

[15] Emad Kadum Jabbar, "New Algorithms for Discovering Association Rules", *PHD. Thesis, Department of Computer Sciences of the University of Technology*, (2005).

[16] Silvestri C & Orlando S, "Distributed Approximate Mining of Frequent Patterns", *ACM Symposium on Applied Computing, Italy*, (2005).

[17] Zaiane OR, El-Hajj M & Lu P, "Fast Parallel Association Rule Mining without Candidacy Generation", *ICDM*, pp.665–668, (2001).

[18] Ahmed CF, Tanbeer SK, Jeong BS & Lee YK, "An efficient algorithm for sliding window-based weighted frequent pattern mining over data streams", *IEICE Transactions*, Vol.92-D, No.7, pp.1369–1381, (2009).

[19] Jabeen TN & Chidambaram M, "Privacy Preserving Association Rule Mining in Distributed Environments using Fp-Growth Algorithm and Elliptic Curve Cryptography", *Indian Journal of Science and Technology*, Vol.9, No.48, (2017).

[20] Jabeen TN & Chidambaram M, "Frequent Pattern Technique using Federation Rule Mining", *Indian Journal of Science and Technology*, Vol.9, No.38, (2016).