A Survey : Intrusion Detection System for database using data mining techniques

Archana Thusoo 1and Prof. G.B Jethava 2

1Parul Institute of Engineering and Technology, Vadodara, India thusooarchana@gmail.com

2Parul Institute of Engineering and Technology, Vadodara, India g.jethava@gmail.com

Abstract: There are various security mechanisms available the security of the database is compromised by various attacks such as sql injection attack, zero day attacks, insider threats and various unknown attacks. To overcome such issues Intrusion Detection System (IDS) are developed, to detect malicious activity occurred in database. The IDS system sometimes are not able to detect the attacks as a false positive or false negative so to overcome such problems the detection method should be modified and enhanced using advance techniques. The efficiency of detection is also less as the unknown attacks are not recognized thus a technique can be delivered by combining security methods to deliver efficient intrusion detection system. The objective of various method is to detect the anomalies using various data mining techniques and provide accurate detection for malicious and intrusive activity. The attack can be done by any external entity or the threat may be caused by the insider, thus to detect the malicious activity and to take action against it the various methods are employed to increase detection of the intrusion and the reduction of false positives. This work provides analysis of various data mining based approaches and shows their efficiency and drawbacks with a vision of an advanced IDS that provides accurate results and reduces false detection.

Keywords: data mining; association rue mining; log mining; transaction-based technique, false positives.

1.INTRODUCTION

The main goal of data security can be divided into three separate, areas as follows. Secrecy is concerned with disclosure of information. The terms confidentiality or non-disclosure are the synonyms for secrecy. information or processes. to information. The term denial of service is also used as a synonym for availability. The secrecy is concerned with the problem of confidential data where there is extreme necessity of hiding the information of the users stored in a database. For example, in a credit card system, the card no., user name, its code and other confidential details are store, if such kind of sensitive data is leaked, or it is accessed by some hacker there can be a big loss and the misuse of the data is possible. The ultimate target of any attacker is a database thus it is highly essential to protect it from intrusion so as to maintain secrecy and data integrity. These three objectives also differ with respect to understanding of the objectives themselves and of the technology to achieve them. It is easiest to understand the objective of secrecy. Integrity is a less tangible objective on which experts in the field have diverse opinions. Availability is technically the least understood aspect. In terms of technology, the dominance of the commercial sector in the marketplace has led vendors to emphasize mechanisms for integrity rather than for military-like secrecy needs. These are severe attacks possible on a database system, and many detection techniques have been found to detect as well as prevent the intrusions, still there is scope of improving the mechanism, so as to provide accurate results and detect the anomalies as well as misuse of the data. The main motivation of this has emerged with the deliberate amount of work still in progress and remaining in order to provide an intrusion detection system as such on database management level which detects the known and unknown attacks. Hence in order to increase the efficiency of the intrusion detection this method has been motivated by two observations made on existing systems. First, despite of many existing intrusion detection systems the attacks are present in the database system and the attacks can be performed by the insider of the organization thus it is difficult to find out the malicious user in the system. Second, it is important to note that the intrusion detection systems on database level are having an overhead of storing large datasets and to increase the efficiency of the intrusive activity detection Data mining has attracted a lot of attention due to increased, generation, transmission and storage of hugs volume data and an imminent need for extracting useful information and knowledge from them. In recent year's research have started looking into the possibility of using data mining techniques in the emerging field of computer security especially in the challenging problem of intrusion detection. Intrusion is commonly defined as a set of actions that attempt to violate the integrity, confidentiality or availability of a system. Intrusion detection is the process of finding important events occurring in a computer system and analyzing them for possible presence of intrusion. Intrusion detection is a second line of defense, when all the prevention technique is compromised and an intrusion has potentially entered into the system. In general, that are two types of attacks: (i) Inside attack are the ones in which an intruder has all the privilege to access the application or the system, but it perform malicious actions. (ii) Outside attack are the ones in which the intruder does not have proper rights to access the system. Detecting inside attack is usually more difficult compare to outside attack.

2. DATA MINING TECHNIQUE

2.1 Misuse detection or Signature based: In signature based approach a signature of known attack is generated. The generated attack signature has been kept in for intrusion detection. A signature is a feature of an intruder. This approach detects only known attacks. The problem with this approach is that it is not capable to detect new attack introduced by intruder that has a no signature in database. In signature based false negative alarm rate increase. Chung et al. [12] present DEMIDS, misuse detection system for relational database systems. This method assumes that the legitimate users show some level of consistency in using the database system. If this assumption does not hold, it results in a large number of false positives. Lee et al. [13] designed a signature based database intrusion detection system (DIDS) which detects intrusions by matching new SQL statements against a known set of transaction fingerprints. However, generating the complete set of fingerprints for all transactions and maintaining its consistency is a rigorous activity. Moreover, if any of the legitimate transaction fingerprints are missing, it can cause many false alarms. The main problem with this approach is that it is difficult to ensure that the fingerprints thus learned are indeed precise and complete.

2.2 Anomaly or Profile based: In profile based intrusion detection approach, a profile of normal user is used for intrusion detection. This approach is suitable for finding unknown attack in database. The profile of normal user is stored in database for intrusion detection. The problem with this approach is it requires more training data set. In this approach false negative alarm increased. Another problem is that significant time and effort is required for training. Zhong et al. [6] use query templates to mine user profiles. Bertino et al. [14] proposed a database IDS that has similarity with role-based access control (RBAC) model in profile granularity. The problems resume by Rao et al. [19] , this approach extracts the correlation among queries of the transaction. In this approach database log is read to extract the list of tables accessed by transaction and list of attributes read and written by transaction.

2.3 Association rule or dependency mining: Association refers to the correlation between items in a transaction. This approach work on data dependency, in which one item is modify another item refer with this also modify. Hu et al [16] determine dependency among data items where data dependency refers to the access correlations among data items. These data dependencies are generated in the form of classification rules, i.e., before one data item is updated in the database, which other data items probably need to be read and after this data item is updated, which other data items are most likely to be updated by the same transactions. Transactions that do not follow any of the mined data dependency rules are marked as malicious transactions. The problem with this concept is that they consider only those attribute that appeared more frequently either they are sensitive or not. They treat all the attributes at the same level and of equal importance, which is not always the case in real applications. In this approach there is no concept for attribute sensitivity. Some attribute may be accessed less frequently but their modification made a more inconsistency in database. Wang et al [17] have proposed a weighted association rule mining technique in which they assign numerical weights to each item to reflect interest/intensity of the item within the transaction .Tao et al [18] use weighted support for discovering the significant itemsets during the frequent itemset finding phase. Also have recently studied a proposed method the use of weighted association rule mining for speeding up web access by prefetching the URLs. These pages may be kept in a server’s cache to speed up web access. Existing techniques of selecting pages to be cached do not capture a user’s surfing patterns correctly. It use a Weighted Association Rule (WAR) mining technique that finds pages of the user’s current interest and cache them to give faster net access. This approach captures both user’s habit and interest as compared to other approaches where emphasis is only on habit. Data mining techniques can be used to mine these logs and extract association rules between the URLs requested by the users. The association rules will be of the form X → Y where X and Y are URLs. It means if a user accesses URL X then he would be accessing URL Y most likely. Database intrusion detection system is design using various approach here with we explain using data mining techniques. In this work, we have identified some of the limitations of the existing intrusion detection systems in general, and their incapability in treating database attributes at different levels of sensitivity in particular. In every database, some of the attributes are considered more sensitive to malicious modifications compared to others. Here with we explain an algorithm for finding dependencies among important data items in a relational database management system. Any transaction that does not follow these dependency rules are identified as malicious. The importance of this approach is it minimizes the number of false positive alarm. This approach generates more rules as compared to non-weighted approach. So there is a need for a mechanism to find out which of the new rules are useful for detecting malicious transactions. Such a mechanism helps in discarding redundant rules. However, the main problem with attribute dependency mining is the identification of proper support and confidence values.

2.4 Database Intrusion Detection based on Improved Association Rule Algorithm [1] It presents an improved association rule algorithm, based on which it builds a database intrusion detection system on the basis of association rules. This system is a circularly and dynamically updated system, but it must first create a set of legitimate access rules from a static database as the basis for the system's judgment. After entering the dynamic intrusion detection management process, the normal data may be extracted from the historical network data stream through the intrusion detection, or the additional data judged to be legitimate, which strictly removes all the possible invasion data, so the rules extracted from the normal data are normal rules, of course, the better the more training data, thus the extracted rules are more complete. Apriori algorithm has the following two defects: 1) Algorithm must spend a lot of time to deal with huge candidate item sets. 2) It must repeatedly scan the transaction database to carry out pattern matching for the candidate item sets. Just because the above two flaws, it presents the technology based on the data partition to improve the adaptability and efficiency of the Apriori algorithm. It can use data partition technique for mining frequent item sets with only two times of the whole database scan. As shown in figure, it contains two main processing stages. The first phase, the algorithm will divide the transactional database D into n independent parts For each division (part), to mine all the frequent item sets in which, they are called local frequent item sets. In terms of the whole database D, a local frequent item set is not necessarily the global frequent item set, but any global frequent item set will certainly occur in the local frequent item sets obtained by the partition. This is very easy to get evidence to the contrary. Therefore, the local frequent item sets mined from n partitions can be as the candidate item set of the frequent item sets in the whole database D; and in the second stage again scanning the entire database for the support frequency of all candidate item sets, to finally confirm the global frequent item set. The partition size and number has the standard that each partition can be entirely placed into memory, so each stage only needs to read the database content once, and the entire mining needs to scan the entire database twice.

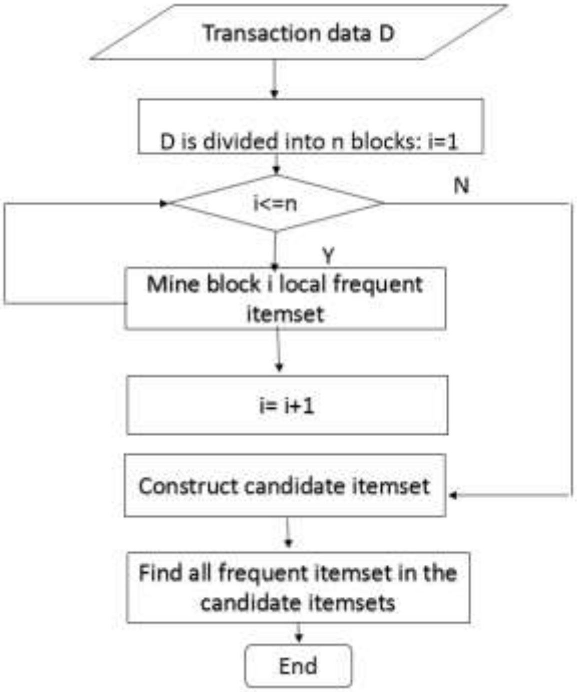
Transaction data D
D is divided into n blocks: i=1
i<=n
N
Y
Mine block i local frequent itemset
i= i+1
Construct candidate itemset
Find all frequent itemset in the candidate itemsets
End

Fig. 1 Data Partioning algorithm

2.5 Hybrid Approach for Database Intrusion Detection with Reactive Policies[2] It describes an approach for finding the intrusive activity using advanced apriori algorithm and also it introduces the concept of Reactive Policies. The Reactive Policies are the action rules defined to be taken against the intrusive activity. These policy are created based on the severity of an intrusion and an appropriate response is generated for the users who performed intrusive activity. Misuse and Anomaly detection are the two measure techniques used for IDS. Misuse detection is also known as signature based detection. Misuse detection technique is used for known attacks whereas Anomaly detection is based on finding the unknown attacks. Misuse detection is unable to detect the zero day attack and hence anomaly detection is employed in the system alongside the misuse detection. Both techniques have their own advantages and disadvantages: Misuse detection technique has already defined attack patterns so rate of false alarms is less, but it detects only known attacks. On the other hand anomaly detection technique detects novel attacks but the rate of false alarms is high. The important components of the HRDIDS are discussed below. A. Pre-processed Audit Log. The very first step in designing the DIDS is to collect the logs of user activity. The information regarding the user activity should be collected properly and need to be processed properly. Preprocessing of collected logs is done. This is done by collecting the user activity in a proper format. This format consists of attributes of a user and its corresponding activity. Each activity is identified by an operation ID and operation status ID. B. Data mining Data Mining is very useful for market-basket applications, to analyse the trends of market. Association rule mining is very popular and useful technique in extracting patterns in a large database; it is a very well researched technique. Large datasets are observed which contain items that frequently occur with each other and a threshold level is defined; if the percentage of threshold is crossed for certain association, a strong rule is generated. These rules can be very useful for deciding future trends and in our case; this will give us exact strategy of intruders. One of the very popular techniques in association rule mining is Apriori algorithm. Apriori algorithm gives us the important associations and gives us association patterns that can be very useful for detecting the intrusions. We are using data mining technique to find out the associations in a user activity. Each activity is monitored with several attributes and corresponding associations are observed. One of the most popular data mining approaches is to get frequent itemsets from a transaction dataset and derives association rules. Apriori algorithm generates the associations that are hidden in the operations. For every abnormal event a reactive policy is applied. Test phase is the most important part of IDS. The decision of an audit log being intrusive or normal is taken in test phase.

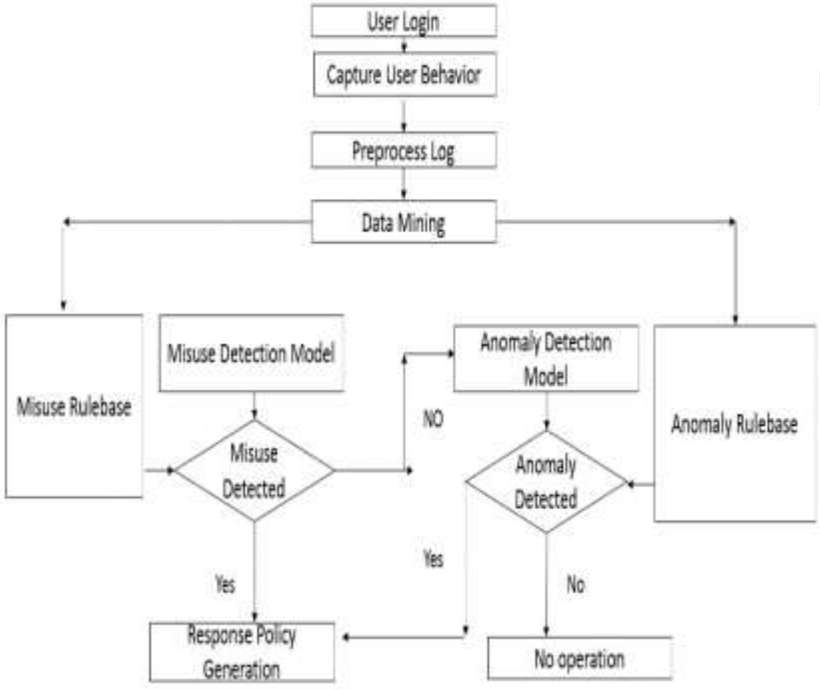

Fig. 2 System architecture of HRDS

We capture both normal as well as abnormal activities of the users and we test these activities with the normal patterns which we have identified during the training phase current data (audit log) is matched with trained rule base in case of anomaly detection process. If the process uses signature based intrusion detection technique then current data is matched with the updated rule base consisting of all previous intrusive signatures. System detects known as well as unknown intrusions and enhances the security by generating more selective and sensitive rules. This process is made faster by implementing the improved Apriori algorithm.

2.6 Detection of Malicious Transaction in Database using Log Mining Approach [4] This paper defines the log mining technique as automatic discovery for identifying anomalous database transactions. This approach can achieve desired true and false positive rates when the confidence and support are set up appropriately. The implemented system incrementally maintain the data dependency rule sets and optimize the performance of the intrusion detection process. There are two phases in which the approach is divided 1. Training phase 2. Detection phase. Training phase is to capture the behaviour of database objects, this monitor and audit the system operation. This auditing system helps to collect necessary data for building database profiles. To be more accurate, whatever technique the profiler utilizes to build the profiles, data gathered by auditing system provides necessary input for it. Depending on the suspicious level or sensitivity of intrusion, detection mechanism can contribute to access control system to deny access and prevent the intruder from causing malicious transaction. The log file consists the information about the committed transactions those are executed in the secure environment by the authorized users. Transactions profile are considered as authorized profiles and stored at the system, after that these authorized transactions profile are used at the detection phase.

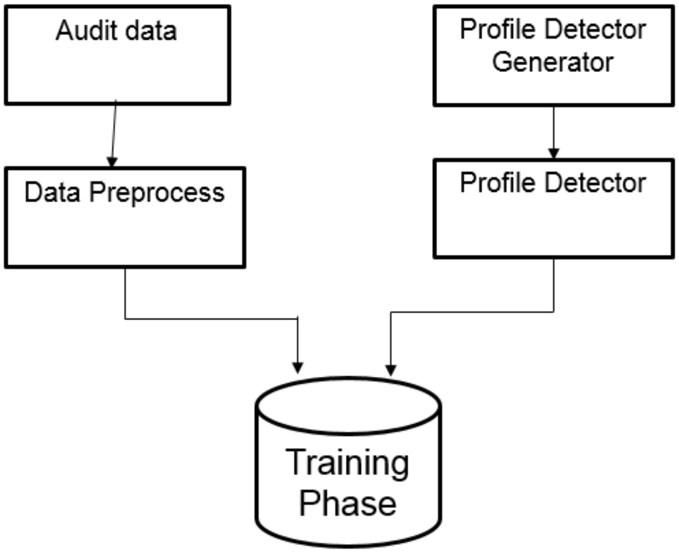

Fig. 3 Training phase the training phase for proposed system.

To capture the behavior of database objects, this monitor and audit the system operation. This auditing system helps to collect necessary data for building database profiles. To be more accurate, whatever technique the profiler utilizes to build the profiles, data gathered by auditing system provides necessary input for it. Detection system for Database. Depending on the suspicious level or sensitivity of intrusion, detection mechanism can contribute to access control system to deny access and prevent the intruder from causing malicious transaction.

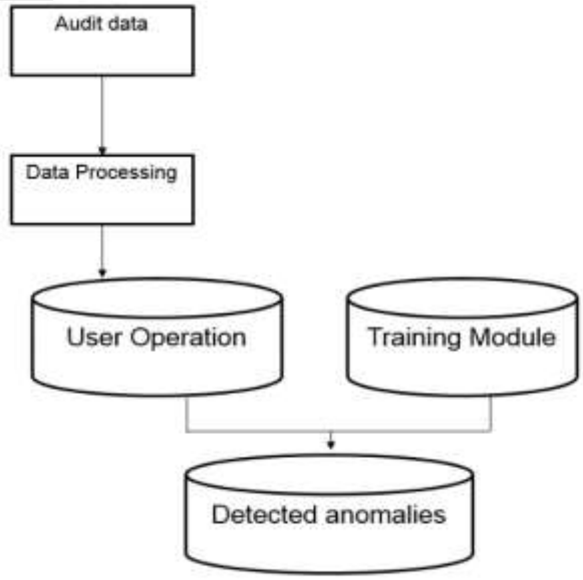

Fig. 4 Detection phase

The log file consists the information about the committed transactions those are executed in the secure environment by the authorized users. Transactions profile are considered as authorized profiles and stored at the system, after that these authorized transactions profile are used at the detection phase.

2.8 Database Intrusion Detection by Transaction Signature [3] The method evaluated here is located on the level of database management system .It focuses on security policies permitted on database system, it is designed to mine audit log of legitimate transaction performed with database and generate signature for legal transactions. The transaction which does not match the signature are declared a malicious transaction according to the policies defined. False positives are valid transactions identified as malicious transactions. In this mechanism the existence of false positives depends on how complete the definition of authorized transaction is. The proposed approach is based on using transaction signature and has learning, detection and response phase. Very briefly, the behaviour of database transaction is collected as a first step to feed the learning phase. Once the database utilization signatures is established, the behaviour learned from audit data is used to concurrently detect database intrusions in detection phase. For intrusive behaviour, this mechanism will alert database administrator. The central theme of my approach will be to learn and create signature from the collected audit data. A basic foundation for intrusion detection is collecting various normal behaviours of Database describe our architecture model in three phases.

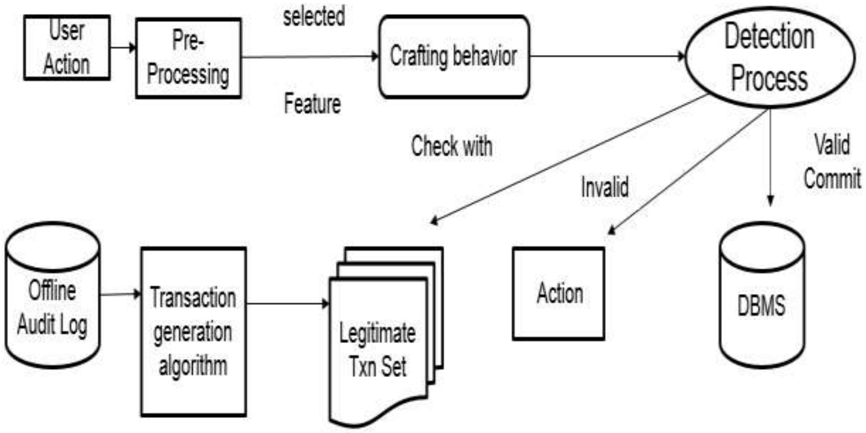

Fig. 5 Architecture of Transaction Signature Model

1) Phase 1: Learning Phase For propose system first phase can be identify as learning phase. Normal or legitimate behaviour is understood in this phase. We have records in audit log of DBMS which contains users' action as per security policy of system. In propose system as in Fig. 5 ,we have offline audit log data store and Legitimate Transaction signature generation modules, which both actually used for identifying normal and acceptable behaviour of user's database transactions. All historical transaction with database is accessed to understand behaviour of legitimate transactions. Various techniques like trigger generation or enabling audit log with database is used for same purpose. 2) Phase 11: Generating signature for user's action Preprocessing module as in Fig. 5 is used for extracting necessary information from transactions user performs with database. Normally user transaction is between BEGIN and END statement in transaction and contains various clause like select, insert, update etc. and attributes of database upon which operations are performed. In pre-processing we will extract key words, operations and target entities and kept it in dataset. So output of pre-processing is dataset or transaction set which can be used for next module. 3) Phase lll: Response This module will be responsible for deciding action depending on whether user's action is legitimate or not. Whatever the output of the signature generation algorithm, will be compared with signature derived from historical data. Based on this comparison this module will decide what action should be taken. B. Design and Implementation Here we are going to use sample dataset of transactions with database. We are using sample database with table order, product, order-line and stock. Various operations are performed on this database.

2.9 An Immune Based Relational Database Intrusion Detection Algorithm [9] In this paper, intrusion detection approaches for relational database systems were studied. An immune based intrusion detection algorithm for relational databases was pro- posed. According to the algorithm, the data to be detected were encoded into binary strings after preprocessing. Intrusion detection was fulfilled by comparing the strings of audit data with immune detectors. The results show that the immune based intrusion detection algorithm for relational databases is more effective reducing the false alarm ratio and promoting correctness ratio. Immune detectors functioned like the immune cells in biological immune systems. The detectors attempted to recognize suspect user behaviors, which patterns were highly similar to the patterns of the detectors. The matched behaviors were thought as anomalies. The philosophy of negative selection was adopted to generate the immune detectors. 1) Candidate detector generation: Candidate detectors are the initial binary strings generated for training. In some computer immunology systems, the candidate detectors were generated randomly. In this paper, since the length of each binary code is 8, there are at most 28 = 256 binary strings. As the number of possible detectors is limited, the candidate detectors were not generated randomly, but by enumerating all the 256 binary strings. 2) Mature detector generation: The philosophy of negative selection in biological immune systems was adopted to generate the mature detectors. That is, only the lymphocytes that have immune reactions to the extern antigens can live. The lymphocytes that have immune reactions to the self-cells will be killed. The mature detectors in the immune based intrusion detection system were generated similarly. Each string in candidate detectors set was compared with all the binary strings in self set. Only such detectors that cannot match any selfstring should be reserved. The candidate detectors that matched the self-strings were deleted. All the reserved strings made up of the mature detectors set. Once mature detectors generated, they can be applied to detect anomalies by comparing with the strings of audit data collected in real time. The r contiguous bit matching rule is also adopted. Since after negative selection, none of the mature detectors may match self-strings. When a string that matches one of the mature detectors is discovered, an anomaly is detected.

3. CONCLUSION

There are various methods for detecting the intrusion but still the intrusive activity occurs and malicious transaction takes place. Due to such intrusion the security of confidential and sensitive data is compromised. There is a scope for an improvement in the detecting methods for intrusive activity in database management system and optimizing the detection rate The future work will be enhance the approaches and to overcome the limitation of the processing power and the data storage issues to handle huge amount of information.

REFERENCES:

[1] Zhang Yanyan, Yao Yuan “Study of database intrusion detection based on improved association rule algorithm”, 3rd IEEE International Conference on Computer Science and Information Technology, Vol.4, pp 673-676, 2010.

[2] Rajashree Shedge, Lata Regha “Hybrid approach for database intrusion detection with reactive policy”, 4th IEEE International Conference on Computational Intelligence and Communication Networks, pp 724-729, 2012.

[3] Y.Rathod, M.Chaudhari and G. Jethava, “Database intrusion detection by transaction signature”, Proceedings of 3rd International Conference on Computing & Networking Technologies, India, pp. 1-5, July 26-28, 2012.

[4] Apashabi Pathan and Madhuri A. Potey “Detection of malicious transaction in database using log mining approach”, International Conference on Electronic systems, signal processing and computing technologies, pp 262-265, 2014.

[5] Gongxing Wu and Zhejiang Gongshang “Design of A New Intrusion Detection System based on Database “International Conference on Signal Processing Systems” , China, 2009

[6] Yong Zhong, Xiao-lin Qin. Database Intrusion Detection Based on User Query Frequent Itemsets Mining with Item Constraints [1]. Proceedings of the 3rd international conference on information security .2004:224-225. 676

[7] Yawei Zhang, Xiaojun Ye, “A Practical Database Intrusion Detection System Framework”, IEEE 9th International Conference on Computer & information Technology, Vol. 2, pp 342-347, 2009

[8]]N.Ye, Q.Chen. “An Anomaly Detection Technique Based on a Chi-Square Statistic for Detecting Intrusions into Information Systems, Quality and Reliability Engineering Internationa”l,2001, 17(2):I 05-112.

[9] Xiaomei Dong, Xiaohua Li , “An Immune Based Relational Database Intrusion Detection Algorithm”, IEEE 2009 Ninth International Conference on Hybrid Intelligent Systems, pp. 295-300

[10] J. Fonseca, M. Vieira, and H. Madeira, “Integrated Intrusion Detection in Databases”, In proceedings of Dependable Computing, Vol. 4746, pp. 198-211, 2007.

[11] U. P. Rao, and D. R. Patel, “Design and Implementation of Database Intrusion Detection System for Security in Database”, International Journal of Computer Applications, Vol.35, No.9, 2011.

[12] C. Y. Chung, M. Gertz and K. Levitt, “DEMIDS: A Misuse Detection System for Database Systems”, In Proceedings of the Integrity and Internal Control in Information System, Pages 159-178, 1999.

[13] S.Y. Lee, W. L. Low and P. Y. Wong, “Learning Fingerprints for a Database Intrusion Detection System”, In Proceedings of the 7th European Symposium on Research in Computer Security, Pages 264-280, 2002.

[14]. Bertino, E., Terzi, E., Kamra, A., Vakali, A: “Intrusion Detection in RBAC-Administered Databases”. In: Proceedings of the 21st annual computer security applications conference (ACSAC), pp. 170–182 (2005)

[15] A. Kundu, S. Sural, A. K. Majumdar, “Database Intrusion Detection Using Sequence Alignment”. International Journal of information security volume 9, number 3, 179-191, DOI: 10.1007/s 10207-010-01025.

[16] Y. Hu, B. Panda, “A Data Mining Approach for Database Intrusion Detection”, Proceedings of the ACM Symposium on Applied Computing, pp. 711-716 (2004).

[17] W. Wang, J. Yang, P. S. Yu, “Efficient Mining of Weighted Association Rules”, Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 270-274 (2000).

[18] F. Tao, F. Murtagh, M. Farid, “Weighted Association Rule Mining using Weighted Support and Significance Framework”, Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 661-666 (2003).

[19] Rao, U.P., sahani, G.J., Patel, D.R., “Detection of Malicious Activity in Role Based Access Control (RBAC) Enabled Databases”. In Proceeding of Journal of Information Assurance and Security 5 (2010) 611-617