ISSN: 0976-3104
SPECIAL ISSUE: Emerging Technologies in Networking and Security (ETNS)
Kumar and Gopal

ARTICLE OPEN ACCESS

IMPLEMENTATION OF IMPROVED APRIORI ALGORITHM IN INTERNAL

Vijay Kumar and Veni Devi Gopal*

Department of Information Technology, KCG College of Technology, Chennai, INDIA

ABSTRACT

The organizations nowadays mostly use user IDs and passwords as the entry gateway to authenticate the users. But somehow they share or give their credentials to their coworkers for their work and request them to aid co-tasks, thereby losing their security of their credentials and making it as the one of the entry points for an attack. Intruders are basically of two categories. First thing is the external intruder. They are those who are the unauthorized users of the machines they attack and the next are the internal intruders who are those, who have the permission to access and work on the system, but will not have access to some portions of the system and they are hard to find and detect because most of the intrusion detection systems and firewalls can identify and isolate malicious behaviors from the outside the system only. Therefore a security system called the Internal Intrusion Detection and the Protection System referred to as (IIDPS) is made to find and detect the insider who attacked the system and also to keep track of user's habit and finally to determine whether a valid user or not by comparing the current behaviors with the patterns collected and stored in the server before.

Published on: 08th– August-2016

KEY WORDS

Data mining, malicious behavior, user habits, intruder, security.

***Corresponding author: Email:** kvijay88859@gmail.com,,venidevig@gmail.com; **Tel.:** +91 9500088859

INTRODUCTION

Data Mining [1-2] is basically a concept of extracting or mining away the knowledge or the information needed from the large block-sets of data. Data mining is one of the way for identifying or to find the the intensive knowledge and information needed for use from those large amounts of data that is stored either in the databases, data warehouses, or the other repositories. Its trust and capacity is to discover valuable data efficiently, non-frequent information from those large databases. But it is certainly sometimes vulnerable to the wrong use of it. So, there might be some confusion among data mining and privacy. Basically, the mining process [3] will be done on this for effectively analyzing and obtaining the results. Privacy [4] basically deals with the extraction of sensitive information with the help of data mining technique. There is an alarming concern about the privacy of the individual users. Each individual has to control their information on their own. The general issues are the usage of other person's credentials or their information. Intrusion detection [5] is a research area where mining algorithms are used and analyzed for the effective detection of and protection for the individual's privacy concern and a brief can be found about it on the guide to IIDPS [6]. The Apriori Algorithm is an effective and suitable algorithm for mining the frequent item-sets for boolean association rules. There are two key concepts which are as follows. 1) The items must have minimum support. 2) The subset contained in the frequent item-set should be frequent.

Some of the possible threats to security are listed below.

Risk - Due to the malfunction of the hardware or due to the incomplete design or due to the incorrect software design because of which there may be an exposure of the information or the data or there may be some violation.

Vulnerability - Some known or suspected flaw because of which the hardware or software that is being used or the operation of a system that exposes the system to an accidental disclosure.

Attack – After the execution of a plan., a threat is carried out as a result

Penetration - A successful attack or a successful intruder one who can obtain the unauthorized access as a result to those files and programs to control the state of a computer system.

THE IIOAB JOURNAL
ISSN: 0976-3104

Intrusion detection systems (IDS) are basically focused on -

- Identifying possible activities of the user and the system.
- Gaining information about them.
- Analyzing the vulnerability.
- Assessing the file contents and the integrity of the system.
- Recognizing malicious activities and patterns that are deviating.
- Giving an alert to the administrator.

In addition, organizations use it generally for things like -

- Identifying issues and the problems associated with those identified issues.
- Analyzing the threats.
- Following the security policies.

RELATED WORK

Computer forensics [7] science views computer systems in order to identify the pattern, preserve, recover, analyze the results and present important facts and opinions on information collected. It analyzes about the attackers and their behaviors like spreading some computer viruses, some malwares and some malicious codes. The intrusion detection techniques helps on how to detect the malicious network behaviours [8], [9] It refers clearly that, from long searched and generated log files or patterns or the data-sets, these traces or patterns of misuse can be more accurately identified and reproduced when an unknown or a malicious user logs in to a system with the valid user credentials.

Mining of frequent item-sets is an important part in the association mining to search the frequent items from the list of item-set in the database. It is important to find the interesting patterns from the large datasets, such as the association, from the episodes, from the classifier, from the clustering and from the correlation etc. [10]. Apriori [11],[12] is like a subset of the candidate generation approach. It basically generates the candidate item-sets of length (k+1). It is based on the frequent item-sets of length (k). The item-set frequency can also be done by counting their occurrences during the transactions. Then at last in 2000,the Pattern growth was proposed by Han where he did some useful work regarding the FP tree.

RELATED DETAILS

This IIDPS basically uses two techniques to find the malicious behavior of the user who uses or accesses the account. The first one is basically like the physical entities. So many physical entities will be evaluated to find out about the malicious behavior. The physical entities such as finding where the user is accessing it and what system has been used frequently that would have been recorded before. So with that recorded information the behavior will be evaluated to find the data of who the intruder is.
The few characteristics of the physical entities with which the intruder can be traced out are –

- By finding the physical location of the intruder.
- By identifying the ip address of the system.

Next method is to find out the malicious behavior with the help of user's habits. The user who is accessing the account will have certain characteristics and these things are compared with those that are recorded in the server. From the result after comparing with the characteristics that are in the server, the malicious behavior can be obtained and the same can be reported to the admin as an alert or by mail.

The few characteristics of the user habitual entities with which the intruder can be traced out are

- With the users preferences
- A simple calculation.
- OTP.

THE IIOAB JOURNAL
ISSN: 0976-3104

So by using few combinations of the above, Security points are made, and the originality of the user can be found. If there is some unusual behavior found after the comparison, then the admin will be intimated from this. So, the admin has the right to stop the access after identifying the malicious behavior. In case if the same is done across in an organization then the admin can ask the user whether to grant or stop the access. In the other case if it's a college the admin can directly stop the access after knowing that malicious behavior of the user who tries to access the system.

THE IMPROVED APRIORI ALGORITHM

The following is an algorithm for finding the frequent item set using the improved apriori algorithm.[13]
Input: transaction database D; min-sup.

Output: the set of Frequent L in the database D.

(1)Find min-sup-count from D.

(2)Generate L1-candidates.

(3)Identify L1.

(4)for (k=2; Lk-1,k--)

(5){for each k-itemset (p, (xp) Lk-1 do

(6) for each k-itemset (q,(xq) Lk-1 do

(7) if (xp[1]=xq[1])!(xp[2]=xq[2])!…!(xp[k-2]=xq[k-2] then

(8){Lk-candidates.Xk= Xp* Xq;

(9) Lk-candidates.TID-set(Xk)

(10)for each k-itemset<Xk,TID(Xk)> Lk-candidates do

(11)Find sup-count

(12)Identify Lk

(13)set-count=Lk.item-count

(14)return L="kLk; [14]

Transactions in a database D=10

ISSN: 0976-3104

Table: 1. D of 10 Transactions

TID	Items
T1	It1,It2,tI4
T2	It2,It5
T3	It2,It3
T4	It1,It2,It5
T5	It1,It2
T6	It2,It3
T7	It1,It3
T8	It1,It2,It3,It4
T9	It1,It2,It3
T10	It1,It4

1) Scan D for count of each candidate.

Table: 2. Generation of C1

Item Set	Support Count
It1	7
It2	8
It3	4
It4	3
It5	2

2) Support count of the candidate-set is then compared with that of the minimum support. Suppose that the minimum transaction support count required is 2.

Table: 3. Generation of L1

Item Set	Support Count
It1	7
It2	8
It3	4
It4	3
It5	2

3)

4) Generate C2 candidates from L1 and scan D for count of each candidate.

Table :4. Generation of C2

Item Set	Support Count
It1,It2	5
It1,It3	3
It1,It4	3
It1,It5	1
It2,It3	4
It2,It4	2
It2,It5	1
It3,It4	1
1t3,It5	0
It4,It5	0

THE IIOAB JOURNAL
ISSN: 0976-3104

5) Compare candidate support count with minimum support. L2 is determined.Then D2 was determined from L2.

Table: 5. Generation of L2

Item Set	Support Count
It1,It2	5
It1,It3	3
It1,It4	3
It2,It3	4
It2,It4	2

Table: 6. D2(Updated Table)

TID	Items
T1	It1,It2,It4
T4	It1,It2,It5
T8	It1,It2,It3,It4
T9	It1,It2,It3

6) Generate C3 candidates from L2 and scan D2 for count of each candidate..

Table: 7. Generation of C3

Items	Support Count
It1,It2,It4	2
It1,It2,It5	1
It1,It2,It3,It4	1
It1,It2,It3	2

Table: 8. C3 (Updated Table)

Items	Support Count
It1,It2,It4	2
It1,It2,It3	2

6) Compare candidate support count with that of the minimum-sup. The D2 is scanned in order to determine L3.

Table: 9.Generation of L3

Items	Support Count
It1,It2,It4	2
It1,It2,It3	2

7) The algorithm uses L3 to generate a candidate set of 4-itemsets, C4.

Table: 10. Generation of D3

TID	Items
T8	It1,It2,It3,It4

THE IIOAB JOURNAL
ISSN: 0976-3104

PROPOSED WORK

In this section, we have introduced the IIDPS system and the components of the IIDPS which are described in detail. An algorithm was also presented for generating a user habit file and detecting an internal intruder. The IIDPS system as in **Figure-1**, consists of a mining server, a detection server and three repositories such as user log file repository, user profile repository and an attacker profile repository. So when a user/attacker tries to access the system, the IIDPS will check if it is a valid user by checking it with the admin.

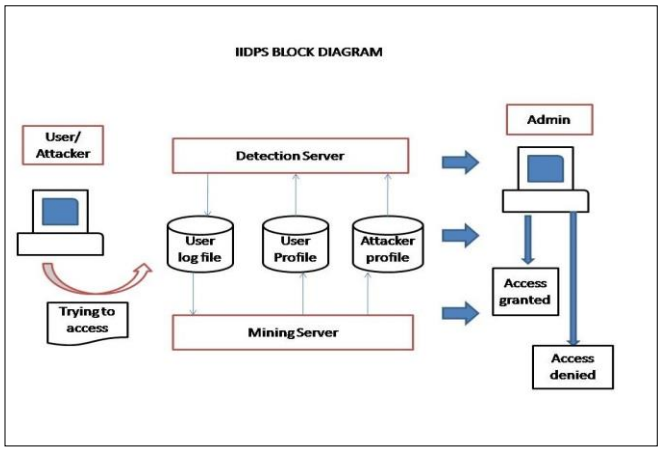


Fig: 1. IIDPS System

The admin will validate and has the power to either grant the access to the user or can revoke the access from the user. By this way, the malicious behaviour if present can be found out with the help of the system.

Support: It is defined as rate of occurrence of an item-set in a transaction database.

$$S(I1 \rightarrow I2) = \frac{Tn(I1,I2)}{Tt}$$

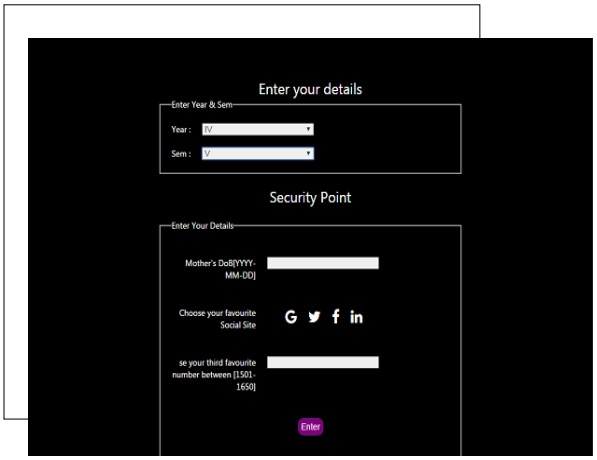
Where S is Support, Tn denotes number of transactions containing both item1 (I1) and item2(I2) . Tt denotes total number of transactions.

Confidence: It defines the ratio of data item-sets that contains Y in the items-sets of X in all the transactions.

$$C(I1 \rightarrow I2) = \frac{Tn(I1,I2)}{T(I1)}$$

Where C is Confidence , Tn denotes number of transactions containing both item1(I1) and item2(I2). T(I1) denotes number of transactions containing item1 (I1).

IMPLEMENTATION



THE IIOAB JOURNAL
ISSN: 0976-3104

Fig: 2. Security point check

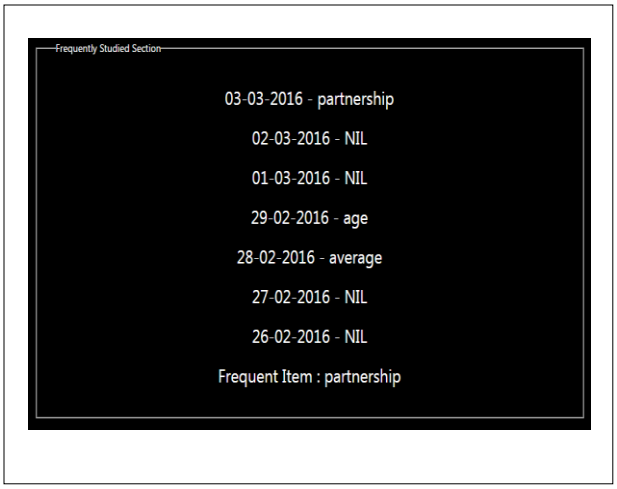


Fig: 3. Frequent item-set

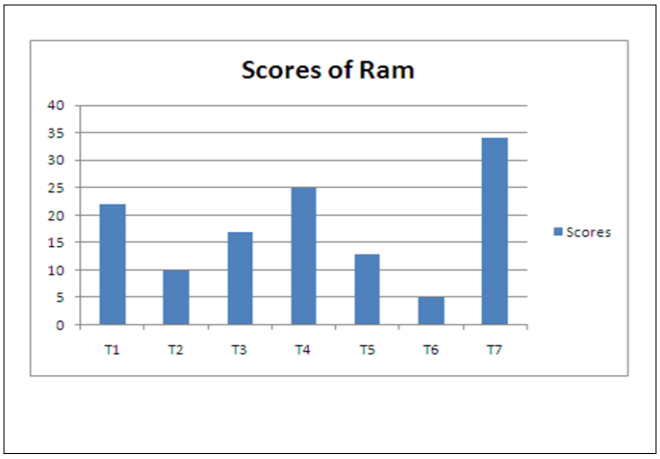


Fig:4. Scores of Ram

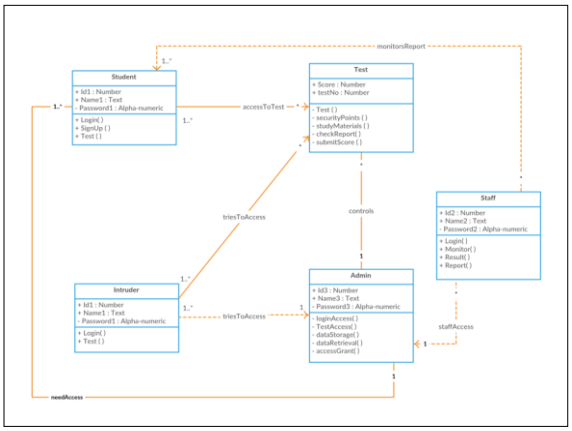


Fig:6. Representation of the overall functions of the system using IIDPS and IAA with the help of the class diagram

ISSN: 0976-3104

The IIDPS system is implemented for an aptitude test system for finding whether the real user of the system is taking up the test. After that, their performance is calculated. The **Figure- 2**, shows the security point which is used to check whether the user is real or not. The frequent item-set is generated after this which is shown in the **Figure-3**, i.e. frequent item-set generation, which is used to monitor the students overall materials that was used by the students.

After that, the staff can login and monitor individual students performance as in **Figure -4** or can monitor the overall performance. A Report can be generated based on the staffs need either semester wise or year wise, to monitor the students performance.

The results of the Classical Apriori Algorithm and the Improved Apriori Algorithm were noted based on its execution time. A comparison was made between those two algorithms and the results were analyzed and a graph was generated as in **Figure- 5**, which clearly indicates that the IAA is more efficient than the CAA.

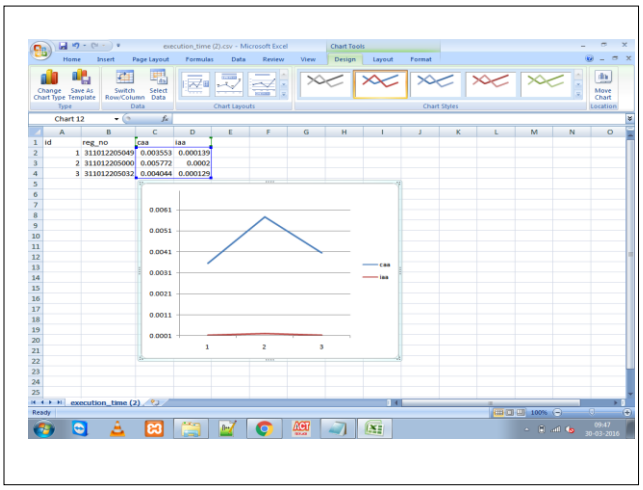


Fig:5. Comparative Results between Classical Apriori Algorithm and Improved Apriori Algorithm in terms of Execution time

The overall functions of the system with the help of IIDPS and IAA are represented with the help of the class diagram as in the **Figure -6.**

CONCLUSION AND FUTURE WORK

In this paper, we have proposed a system which is called the Internal Intrusion Detection and Protection System (IIDPS). The user's preferences are recorded and compared every time when he logs in to the account. So based on the usage profile and the patterns the IIDPS restricts the intruder. In today's world, internal intrusion detection is one of the major topics in which the research is still going on for its effective development. It can be extended to protect the system even at a higher level by using one of the following two ways. First one is that the IIDPS system can be still improvised by introducing new concepts and thereby protecting it from intrusion efficiently. Second one is by extending the system by using recent technologies such as finger print technology or a face reader technology to easily identify the user, thereby avoiding the malicious activity and to improve the IIDPS's reliability and performance. There are so many researches taking place in this field and a suitable one can be included to effectively develop the IDP system to protect the system.

ACKNOWLEDGEMENT
None

CONFLICT OF INTEREST
No conflict of interest

FINANCIAL DISCLOSURE
No financial support was received to carry out this project.

REFERENCES

[1] Fang-Yie Leu, Kun-Lin Tsai, Member, IEEE, Yi-Ting Hsiao, and Chao-Tung Yang "An Internal Intrusion Detection and Protection System by Using Data Mining and Forensic Techniques, *IEEE* 2015.

[2] J Han and Kamber.[2006] Data Mining: Concepts and Techniques, 2nd ed., The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor 2006.

[3] Pingshui WANG.[2010] Survey on Privacy Preserving Data Mining, *International Journal of Digital Content Technology and Its Applications* 4(9)

[4] MB Malik, MA Ghazi and R Ali.[2012] Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects, in Proceedings of Third International Conference on Computer and Communication Technology ,*IEEE*.

[5] Sheetal Thakare1 ,Pankaj Ingle2, BB Meshram.[2012] Intrusion Detection System the Survey of Information Security, *International Journal of Emerging Technology and Advanced Engineering* , 4 (8)

[6] Guide to Intrusion Detection and Prevention System‖, NIST, Technology Administration US Department of Commerce.

[7] ZB Hu, J Su, and VP Shirochin.[2007] An Intelligent Lightweight Intrusion Detection System With Forensics Technique, in Proc. IEEE Workshop Intell. *Data Acquisition Adv Comput Syst Technol Appl Dortmund, Germany*, 647–651.

[8] Q Wang, L Vu, K Nahrstedt, and H Khurana.[2010] MIS: Malicious Nodes Identification Scheme in Network-Coding-Based Peer-to-Peer Streaming, in Proc. IEEE INFOCOM, San Diego, CA, USA, , pp. 1–5.

[9] ZA Baig, "Pattern Recognition for Detecting Distributed Node Exhaustion Attacks in Wireless Sensor Networks, Compute. Commun., Vol. 34, No. 3,pp. 468–484, Mar. 2011.

[10] Rao S, Gupta R, Implementing Improved Algorithm Over Apriori Data Mining Association Rule Algorithm, *International Journal of Computer Science And Technology*, pp. 489-493, Mar. 2012.

[11] Xiang Fang.[2013] An Improved Apriori Algorithm on the Frequent Item set, International Conference on Education Technology and Information System (ICETIS 2013)

[12] SurajP. Patil1, U. M. Patil2 and Sonali Borse.[2012]The Novel Approach for Improving Apriori Algorithm for Mining Association Rule, Proceedings of "ational Conference on Emerging Trends in Computer Technology (NCETCT-2012)"Held at R.C.Patel Institute of Technology, Shirpur, Dist. Dhule, Maharashtra, India. April 21, 2012.

[13] Akshita Bhandari Ashutosh Gupta Debasis Das.[2014] Improvised Apriori Algorithm using Frequent Pattern Tree for Real Time Applications, *ICICT* 2014.

[14] Sakshi Aggarwal1, Ritu Sindhu.[2015] An Approach of Improvisation in Efficiency of Apriori Algorithm, *Peer Jpreprints*