International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/IJSRCSEIT

Document Fraud Detection

Ashifa. T[1], Sathya. R[2]

[1]B.Sc Computer Technology, Sri Krishna Adithya College of Arts & Science, Coimbatore, Tamil Nadu, India
[2]Assistant Professor, Department of Information and Computer Technology, Sri Krishna Adithya College of Arts and Science, Coimbatore, Tamil Nadu, India

ABSTRACT

In today scenario for data and fund transfer we are mainly depended on the internet. So prevention of fraud, abuse and data alteration through internet has become a major concern of many organizations. Our paper focuses the direction towards the document fraud detection. In this direction we proposed an efficient approach where we send the encrypted data via internet and maintain the log table of the data sends. The log table contains the information about individual word and numeric value along with the position and count. If any attacker attacks the data for updation or any violation again a log table is created based on the word as well as the position of the word and count. Our algorithm check the alteration based on the position and the frequency count. If any mismatch is detected server alerts the client about the document attack and resend the document to the client. The above scenario is explained by the result analysis which shows the effectiveness of the approach.

Keywords : Document Fraud Detection, Frequency Analysis, Encryption ad Decryption.

I. INTRODUCTION

There is bombarding of data on internet and we must rely on them by any reason in the day to day life. More and more information is stored in databases and turning these data into knowledge creates a demand for new, powerful tools.

Data analysis techniques used before were primarily oriented toward extracting quantitative and statistical data characteristics.

Data mining is known as gaining insights and identifying useful patterns from the huge amount of data stored in large databases in such a way that the patterns and insights are statistically reliable, previously unknown, and understandable [1][2][3][4]. Data mining is also define as a process that uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information and subsequently gaining knowledge from a large database [5].

If we analysing the property of data mining techniques then we understand the potential to solve the contradiction between effect and efficiency of fraud detection. It can be effective for detection by their association and frequent pattern analysis approach. So in our paper we use data mining technique to find the document fraud detection.

We provide here an overview of executing data mining services with fraud detection along with the security applications. The rest of this paper is arranged as follows: Section 2 introduces Literature Review; Section 3 describes algorithm and proposed

CSEIT195135 | Received : 01 March 2019 | Accepted : 10 March 2019 | March-April -2019 [5 (2) : 182-186]

work; Section 4 shows the result analysis; Section 5 describes the Conclusion.

II. RELATED WORK

In 2010, Shiguo wang et al. [6] categorizes, compares, and summarizes the data set, algorithm and performance measurement in almost all published technical and review articles in automated accounting fraud detection. Most researches regard fraud companies and non-fraud companies as data subjects, Eigen value covers auditor data, company governance data, financial statement data, industries, trading data and other categories. Most data in earlier research were auditor data; Later research establish model by using sharing data and public statement data. Company governance data have been widely used. It is generally believed that ratio data is more

In 2012, Clifton Phua et al. [10] observe that the credit application fraud is a specific case of identity crime. The existing nondata mining detection system of business rules and scorecards, and known fraud matching have limitations. To address these limitations and combat identity crime in real time, they propose a new multilayered detection system complemented with two additional layers: communal detection (CD) and spike detection (SD). CD finds real social relationships to reduce the suspicion score, and is tamper resistant to synthetic social relationships. It is the whitelist-oriented approach on a fixed set of attributes. SD finds spikes in duplicates to increase the suspicion score, and is probe-resistant for attributes. It is the attribute-oriented approach on a variable-size set of attributes. Together, CD and SD can detect more types of attacks, better account for changing legal behavior, and remove the redundant attributes. Experiments were carried out on CD and SD with several million real credt applications. Results on the data support the hypothesis that successful credit application fraud patterns are sudden and exhibit sharp spikes in duplicates. Although this research is specific to credit application fraud detection, the concept of resilience, together with adaptively and quality data discussed in their paper, are general to the design, implementation, and evaluation of all detection systems.

In 2012, Syed Imran Ahmed Qadri et al. [11] provide a security framework for server and client side. In this they provide some prevention methods which will apply for the server side and alert replication is also on client side. Content sniffing attacks occur if browsers render non-HTML files embedded with malicious HTML contents or JavaScript code as HTML files. This mitigation effects such as the stealing of sensitive information through the execution of malicious JavaScript code. In this framework client access the data which is encrypted from the server side. From the server data is encrypted using private key cryptography and file is send after splitting so that we reduce the execution time.

In 2012, V.Priyadharshini et al. [12] multilayered techniques for providing the security for the credit card frauds. The first layer is communal detection and second is Spike detection layers that highly provides security for detection of frauds like probe resistant and mark the illegal user through their input details and mark it in a list. Then it removes attacks like defense in depths on cards and by removing the data redundancy of the attributes and it is being processed with millions of the credit cards.

In [13] authors suggested about clustering approaches. In 2012, Namrata Shukla et al. [14] present an efficient approach for fraud detection. In their approach they first maintain a log file for data which contain the content separated by space, position and also the frequency. Then they encrypt the data by substitution method and send to the receiver end. They also send the log file to the receiver end before proceed to the encryption which is also in the form of secret message. So the receiver can match the data according to the content, position and frequency, if there is any mismatch occurs, they can detect the fraud and does not accept the file.

In 2013, Animesh Dubey et al. [15] propose an efficient partition technique for web based files (jsp, html, php), text (word, text files) and PDF files. They are working in the direction of attack time detection. For this motivation they are considering mainly two factors first in the direction of minimizing the time, second in the direction of file support. For minimizing the time they use partitioning method. They also apply partitioning method on PDF files. In the result section they also provide the comparison with the traditional technique which shows the effectiveness of their approach.

Our method applies effectively to sufficiently large texts. If our file is attacked by intruder and it is successfully altered by the intruder, for this case we maintain a data mining log file for identify the fraud.

In our approach client first request the document which is needed to the client. After the client authentication server prepare the data for sending to the client. Server applies the above algorithm for encryption and sends it to the client. Before sending the data to the client server maintaining the log detail of the file and makes two copies, one copy is for the self-use and another copy for the client. If there is no attack then according to the server database there is no change in the database. If any alteration is performed in between the data sending and receiving to the client. Then server again maintains the log file after attack in the client side. This also contains the individual word their position and the total occurrences or the frequency in the database. Position and the frequency are matched to the initial log which is created on the time of data send. If it is matched then there is no alteration and the data is useful, but if the logs are contradicted in their position of word and frequency count means the data is altered and not useful. Then the server alerts the client about the attack and resends the data for their use. In this manner the detection is achieved in the above approach. For better understanding our above approach we provide the examples in the result section.

III. Algorithm and Proposed work

- ✓ In the Caesar cipher, the following algorithm is used
- ✓ If n is the number of a letter in the alphabet, this letter is replaced by another letter,
- ✓ whose number is (n+k) modulo 26 (shortly (n+k) mod 26)
- ✓ This is a remainder of division of (n+k) by 26
- ✓ For example, take k=5 and take letter X
- ✓ Its number n = 24
- ✓ (n+k) mod 26 = 24 + 5 mod 26 = 3
- ✓ So X is replaced with C
- ✓ Count the number of appearance of each letter and divide it by the total number of words in the ciphertext
- ✓ Compare the results with the frequency table

Our Detection approach also detects the position of the alteration which also help to reaccelerate the data. In this approach we also maintain the attack detection time which is useful for correlation when the attack is performed and the time will be send to the client.

IV. RESULT ANALYSIS

In this section we provide the results of our proposed approach. First client request the data from the server, if server authenticates the client and ready to send the file to the client, then a connection will be established between server and client. Server sends the data and maintains the status and the log file. Server maintains the name of the file decryption key that is used by the client, sending time, receiving time, client name and an identification bit that is 1. If any attack will be detected on the client side the identification bit is changed to 0. If any cahange either in the position or in the number of count it will be detected and registered in the client side and notify to the server so that server resend the file for the client.

Figure 1 shows the numeric data and the concern log file is shown in table 1. Figure 2 and Table 2 shows the data after the attack. The red highlighted text shows the attack file. Figure 3 shows the numeric data and the concern log file is shown in table 3. Figure 4 and Table 4 shows the data after the attack. The red highlighted text shows the attack file.

10 10 20 10 20 30 10 20 30 40 50

Figure 1: Numeric Data

Table 1: Numeic Database Log

Data	Position	Count
10	1 ,2 ,4 ,7	4
20	3 ,5 ,8	3
30	6 ,9	2
40	10	1
50	11	1

10 10 20 10 20 30 10 20 30 40 50 10 50

Figure 2: Numeric Data (After Attack)

Table 2: Numeic Database Log (After Attack)

Data	Position	Count
10	1 ,2 ,4 ,7,12	5
20	3 ,5 ,8	3
30	6 ,9	2
40	10	1
50	11,13	2

a b b c d e r y u oi er oi er er an an d b b c q an

Figure 3:Text Data

Table 3: Text Database Log

ab2log		
Data	Position	Count
a	1	1
b	2 ,3 ,18 ,19 ,23	5
c	4 ,20	2
d	5 ,17	2
e	6	1
r	7	1
y	8	1
u	9	1
oi	10 ,12	2
er	11 ,13 ,14	3
an	15 ,16 ,22	3
q	21 ,24	2

Figure 4:Text Data (After Attack)

Table 4: Text Database Log (After Attack)

ab2log		
Data	Position	Count
a	1	1
b	2 ,3 ,18 ,19 ,23	5
c	4 ,20	2
d	5 ,17	2
e	6	1
r	7	1
y	8	1
u	9	1
oi	10 ,12	2
er	11 ,13 ,14	3
an	15 ,16 ,22	3
q	21 ,24,25	3

a b b c d e r y u oi er oi er er an an d b b c q an q

V. CONCLUSION

The main focus of this paper is to identify the attack from the attackers on any text file document which is encrypted and send to the receiver. For this identification we apply frequency based data mining technique to identify the fraud. Our approach is better in terms of detection in client side on the basis of frequency and the position of bit

VI. REFERENCES

[1]. Elkan, C. (2001). Magical Thinking in Data Mining: Lessons from COIL Challenge 2000. Proc. of SIGKDD01, 426-431.

[2]. Kai Li, Peng Li, " A Selective Fuzzy Clustering Ensemble Algorithm " , International Journal of Advanced Computer Research (IJACR), Volume-3, Issue-13, December-2013 ,pp.1-6.

[3]. Ashutosh Kumar Dubey, Animesh Kumar Dubey, Vipul Agarwal, Yogeshver Khandagre,

[4]. “Knowledge Discovery with a Subset-Superset Approach for Mining Heterogeneous Data with Dynamic Support”,Conseg-2012.

[5]. Preeti Khare, Hitesh Gupta, “Finding Frequent Pattern with Transaction and Occurrences based on Density Minimum Support Distribution”,

[6]. IEEE 2011. Dr. Bhavani Thuraisingham,” Data Mining for Malicious Code Detection and Security Applications”, EuropeanIntelligence and

[7]. Security Informatics Conference, 2011.

[8]. Sherly K.K,” A Comparative Assessment Of Supervised Data Mining Techniques For Fraud Prevention”, TIST. Int.J. Sci.Tech.Res., Vol.1 (2012), 1-6.

[9]. Clifton Phua, Kate Smith-Miles,Vincent Cheng- Siong Lee, And Ross Gayler,” Resilient Identity Crime Detection”, IEEE Transactions On

[10]. Knowledge And Data Engineering, Vol. 24, No. 3, March 2012.

[11]. Syed Imran Ahmed Qadri, Kiran Pandey, “Tag Based Client Side Detection of Content Sniffing Attacks with File Encryption and File Splitter

[12]. Technique”, International Journal of Advanced Computer Research (IJACR), Volume-2, Number-3, Issue-5, September-2012.

[13]. V.Priyadharshini, G.Adiline Macriga,” An Efficient Data Mining for Credit Card Fraud Detection using Finger Print Recognition”,

[14]. International Journal of Advanced Computer Research (IJACR),Volume-2 Number-4 Issue-7 December-2012.

[15]. Shashi Sharma, Ram Lal Yadav, “Comparative Study of K-means and Robust Clustering” , International Journal of Advanced Computer

[16]. Research (IJACR), Volume-3, Issue-12, September-2013 ,pp.207-210.

[17]. Namrata Shukla, Shweta Pandey,” Document

[18]. Fraud Detection with the help of Data Mining and Secure Substitution Method with Frequency

[19]. Analysis”, International Journal of Advanced Computer Research (IJACR) ,Volume 2 Number 2, June 2012.

[20]. Animesh Dubey, Ravindra Gupta, Gajendra

[21]. Singh Chandel, “An Efficient Partition Technique to reduce the Attack Detection Time with Web based Text and PDF files”, (IJACR),Volume-3 Number-1 Issue-9 March-2013.

Cite this article as :

Ashifa.T, Sathya.R, " Document Fraud Detection", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 182-186, March-April 2019. Available

at doi : https://doi.org/10.32628/CSEIT1951135
Journal URL : http://ijsrcseit.com/CSEIT1951135