ISSN No:-2456-2165

Information Leakage Prevention Model using Multi-Agent Architecture in a Distributed Environment

Alese Boniface Kayode (Federal University of Technology, Akure, Nigeria)
Adewale Olumide Sunday (Federal University of Technology, Akure, Nigeria)
Alowolodu Olufunso Dayo (Federal University of Technology, Akure, Nigeria)
Adekunle Adewale Uthman (Federal Polytechnic, Ede, Nigeria, Nigeria)

Abstract:- Various cases of leakage of sensitive files such as confidential report and privacy documents of customers and staff have been reported mistakenly sent via email, leaked through unprotected USB Sticks and mobile devices. Multi-agent systems have experienced growing acceptance and importance as distributed systems become more widespread. The most common issues of this system are security in term of vulnerability of documents. This research address the security issues of a multi-agent in a distributed environment providing a data leakage prevention model that effectively control data leakage, data loss and data theft from an insider.

Keywords:- *Information, Leakage Prevention, Multi-Agent, Distributed System.*

I. INTRODUCTION

The appearance of smart devices in our daily lives was due to various developments in information technology which in turn lead to greater problem of data leakage in distributed systems [1]. There can be accidental or intentional distribution of sensitive information to an unauthorized entity. The spread of confidential information that has been leaked may be practically impossible to stop once it gets out [2].

In information security, prevention of data disclosure by an insider and unauthorized entities is very critical, and it has continuously and constantly drives major sectors to investigate, design and develop different solutions to take mitigate the risk of data leakage [3]. Total prevention of data leakage is not always feasible in a distributed system because users of the system need to access share and use information, which at some points lead to release of confidential data [4].

[5] reported on the growing data leakage, and indicate growing concerns in government and business sectors, and according to DataLoss DB (2015), around 15% of recorded data leakage occurred in the government sector, around 60% occurred in the business sector and around 25% occurred in the education and health sectors. Other users (private users) are also affected by data leakages, but calculating the exact amount and their severity if really difficult to know but not all reported leaks are detrimental but many have them caused high damage [5].

In order to address these serious issues, security experts around the world have developed various security measures such as firewalls, intrusion detection systems and virtual private networks have been introduced over the past decades. If the data to be protected are well defined, constant and structured [6], these systems developed over the years performed satisfactorily. Protecting confidential data can be to be naïve using these methods because a firewall can block access to a confidential data segment using simple centralized rules while the same data segment can be access using email attachment or other means [3].

To find solution to this deficiencies, a new direction for data protection was considered, which led to the introduction of Information leakage prevention model (ILPM). ILPM have the ability to identify, monitor and protect confidential data and detect misuse based on predefined rules. In this research, a decision tree technique is used to determine the semantics of confidential data, that is, semantics text classification for document feature selection, extraction and transformation. This provides accurate document classification with minimum preprocessing of data and provides clarity of data. Group communications data leakage is discovered using some analysis.

II. RELATED WORKS

According to the data leakage detection system proposed by [7] for improving probability of identifying leakages, in order to detect faulty agent, it made a change in data allocation. The system was able to detect faulty party without tempering integrity of the real data.

[8] researched on the significance of data leakage which gave the ideas leading to social network analysis and clustering of text. The work emphasized on different data leakage preventions methods and their related problems.

[9] designed a system that makes use of data allocation methods to increase detection of leakages. The robust mail filtering and information leakage system emanated from other applications. The system makes use of fingerprints of messages bodies and email addresses. The research work also emphasized that distributor need to calculate the aspects that make open records corresponding data leakages from various agents.

[10] authored a Fast Detection of Transformed Data Leaks which focused on unpremeditated data leak detection. Detecting various loopholes for the exposure of important data was difficult due to data transformation. In the model designed, two types of sequences where analyzed i.e. sensitive data sequence that requires to be protected from unauthorized parties and content steps which is to be examined for various leaks. The content data may include records extracted from distributed system, or personal computers from supervised network channels. The sensitive data sequences are known to the analysis system. The sensitive data sequences make use of sequence alignment approaches for tracking the patterns in complex data leak which are known to the analysis system.

[11] presented an approach to detect data leakages in a very secured communication in any ad-hoc networks. A data provider can owns vulnerable data of trusted agents, and some data can be revealed at such permitted place. There was extra security with encryption after the introduction of the fake objects. The system proposed described data loss and reduces performance degradation in a multi-agent environment; it is made up of cryptography and routing protocol execution at each strategic stage during the data transfer.

[12] proposed a hybrid detection leakage framework that make use of both signature and anomaly based solutions thereby leading to both detection and prevention. The system illustrates the challenges in data loss detection and prevention through a running example within the healthcare domain and presents a framework to address these challenges. The aim of the system is to develop a framework for detection leakage programme that employs an anomaly-based engine to detect anomalous transactions. The research work fails to show in detail the anomaly-based techniques adopted.

[13] presented a work that seeks to explain the objectives and properties of the mobile agents in currently used architecture and platform of mobile world from the currently used approach RPC (Remote Procedure Calling) and new approach RP (Remote Programming) of the mobile network. There were little practical introduction to mobile agent technology and surveys the state of the art in mobile agent research.

[13] explained the currently used approach for remote procedure calling and the new approach for remote programming of mobile network. For most mobile agent research, there were little practical introduction to technological application and surveys. [13] proceeds to develop a mobile app using The Aglet mobile-agent Model by gathering of relevant journal articles and categorizing them and describing the fundamental operations of Aglet mobile agent. The result of the research shows that it is purely descriptive.

III. ILPM MODEL

This section is based on the Support Vector Machine (SVM) and the C4.5 decision tree classification methods for document classification to improve classification accuracy. A set of text are automatically classified into different categories from a predefined set.

➢ *C4.5 Decision Trees*

Using a set H of cases, C4.5 grows an initial tree using the divide-and-conquer algorithm. If all the cases in H belong to the same class, the leaf label with the most frequent class in H is the tree. Else, a test based on a single attribute with outcomes more than two is selected. Applying the same procedure to each subsets, H is divided into subsets H1,…..,Hn based on the outcome for each case. C4.5 uses of information gain and default gain ratio criteria to rank tests. The information gain is used for minimizing the total entropy of the subsets while the default gain ratio helps divides information gain by the information provided by the test outcomes.

➢ *Support Vector Machine (SVM)*

The support vector machine (SVM) is a supervised machine learning model for group classification problems using classification algorithms. The SVM model are able to categorize new text after training the model using some sets of labeled training data.

An SVM model is basically like different classes in a hyperplane in multidimensional space. In order to reduce error, the hyperplane is generated in an iterative manner by SVM. The SVM divideS the datasets into classes to find a maximum marginal hyperplane.

➢ *C4.5 Decision Trees and SVM*

In this section, an information leakage detection model was proposed which accurately classify document by cascading SVM and C4.5 algorithm. The method is divided into training phrase and detection phrase.

- *Training Phase*

The training phrase is divided into three stages, the first is to for clustering, the second is graph building and the third stage is the pruning stage, where Hi represents the training data set.

During the training phrase, the SVM model is applied to split the training space into various clusters C1,….,Ck. where, C4.5 model is trained with the instances in each SVM cluster. The SVM model ensures that only one cluster is associated with each training instance. The confidential scores are calculated during the detection phrase by matching the documents to the graphs of clusters. A document is considered as confidential only when the confidential score is more than the predefined threshold. The testing phase is subdivided in two phases, the Selection Stage and the Classification Stage. In selection stage, Euclidean distance is computed for every testing instance to find the closest cluster. The classification stage applies the test instances (Hi) over the C4.5 decision tree of the

computed closest cluster. The test instances (Hi) are then classify as normal or anomaly. Table 1 shows the C4.5 and SVM Algorithm.

C4.5 AND SVM ALGORITHM

Selection Stage
Input: Test Instances H_i, i=1,2,3,…..,N (Confidential Documents set and Non-Confidential Documents set)
Output: Computed closet cluster to the test Instances H_i

Start
Procedure Selection
Begin
Step 1: for each H_i instance
- Calculate the threshold $S(H_i,r_j)$, j=1,2,3,….,k and determine the closest cluster
- Calculate the closest cluster to C4.5 Decision tree

End // End Procedure

Classification
Input: Instance (H_i) test
Output: Classified Instance (H_i,) test
Start
Procedure Classification
Begin:
- Apply the test instance H_i over the C4.5 decision tree of the computed closest cluster
- Classify the test instance H_i as normal or anomaly
- include it in the cluster
- Update the cluster

End //End Procedure

Table 1:- C4.5 and SVM Algorithm.

IV. EXPERIMENT RESULTS AND DISCUSSIONS

The experiments have been performed using C4.5 and SVM which help prevent data leakage with emphasis on data classification using the preventive approach.

➢ *Dataset*

The data that needs to be protected against information leakage are categorized as either confidential or non-confidential data. This could be any sensitive data. The dataset used consist of sensitive information such as Bank Verification Number, Account Number, Salary Details, Username, Password, Card Expiration Date, and other personal information collected and exported into Excel formats for use in the modeling stage.

➢ *Evaluation Matrix*

Table 2 shows the confusion matrix for a two class classifier that would commonly be used in IDS. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class

	Predicted Class	
Class	Normal	Anomaly
Normal	True Negative (TN)	False Positive (FP)
Anomaly	False Negative (FN)	True Positive (TP)

Table 2: Confusion Matrix for Evaluation Purpose

The evaluation is accessed based on the confusion measurement as follows:

True Positive Rate (TPR): It is referred to sensitivity or recall, and used to measures the percentage of actual positive which are correctly identified. TPR is defined as:

$$TPR = \frac{TP}{TP+FN} \quad (1)$$

False Positive Rate (FPR): helps get the quantitative relation between the normal cases detected as anomaly and the overall number of normal cases. FPR is calculated as:

$$FPR = \frac{FP}{FP+TN} \quad (2)$$

False Negative Rate (FNR): In FNR, anomaly test are classified as normal. The FNR is calculated as:

$$FNR = \frac{FN}{FN+TP} \quad (3)$$

Accuracy: The accuracy is the total accuracy in classifying both normal and anomaly and is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision: Precision is the percentage of total true positives (TP) instances divided by total number of true positives (TP) and false positives (FP) instances:

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (5)$$

F-measure: The F-measure is the mean of the precision and recall.

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (6)$$

$$F-measure = \frac{2*Recall*Precision}{Recall+Precision} \quad (7)$$

- *Experiment Evaluation*

The experiments were performed on a windows workstation having a configuration Intel® Pentium® Corei7 @ 3.20GHz, 3.19GHz, 16GB of RAM, and the operating system is Windows 10 Pro. An Open Source machine learning framework WEKA was used. WEKA is a collection of machine learning algorithm for data mining applications. The tool was used for performance comparison of our algorithm with the related other classification algorithms. Table 3 shows the performance of Naïve Bayes, K-NN, SVM, ID3 decision tree and the cascading SVM and C4.5 algorithms using 10-fold validation method. SVM and C4.5 is an excellent classifier

with high accuracy and ILPM performs well in the condition where confidential contents are embedded in non-confidential documents.

Classifier	Performance (%)			
	TPR	FPR	Precision	F-measure
Naïve Bayes	95.3	4.5	95.5	95.3
K-NN	94.5	5.3	94.7	94.6
SVM	99.2	3.0	99.4	99.2
ID3 Decision Tree	93.7	6.4	93.9	93.7
C4.5	99.2	2.5	99.4	99.2
C4.5+SVM	99.6	0.1	99.8	99.6

Table 3:- Performance Evaluation

V. CONCLUSION

In this paper, multi-agent based information leakage prevention model was proposed using C4.5 and SVM classification model to determine the semantic text classification of document thereby providing accurate document classification. This gives clarity of data which in turn help against communication data leakage.

REFERENCES

[1]. X. Yu, Z. Tian, J. Qiu, and F. Jiang (2018). A data leakage prevention method based on the reduction of confidential and context terms for smart mobile devices. Wireless Communications and Mobile Computing, 2018.
[2]. J. H. Ziegeldorf, O. G. Morchon, and K. Wehrle (2014). Privacy in the Internet of Things: threats and challenges. Security and Communication Networks, 7(12), 2728-2742.
[3]. S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy (2016). A survey on data leakage prevention systems. Journal of Network and Computer Applications, 62, 137-152.
[4]. C. Gibler, J. Crussell, J. Erickson, and H. Chen (2012, June). AndroidLeaks: automatically detecting potential privacy leaks in android applications on a large scale. In International Conference on Trust and Trustworthy Computing (pp. 291-307). Springer, Berlin, Heidelberg.
[5]. L. Cheng, F. Liu, and D. Yao (2017). Enterprise data breach: causes, challenges, prevention, and future directions. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(5), e1211.
[6]. R. Bace, and P. Mell (2001). NIST special publication on intrusion detection systems. Booz-Allen And Hamilton Inc Mclean Va
[7]. P. Papadimitriou and H. G. Molina, "Data Leakage Detection," IEEE Transaction on Knowledge and Data Engineering, vol. 23, no. 1, pp. 51-63, January 2011.
[8]. P. Raman, H. G. Kayacık, and A. Somayaji, "Understanding Data Leak Prevention," in 6th Annual Symposium on Information Assurance(ASIA'11), Albany, NewYork, USA, 2011, pp. 27-31.
[9]. Agarwal, M. Gaikwad, K. Garg, and V. Inamdar, "Robust Data leakage and Email Filtering System," in International Conference on Computing, Electronics and Electrical Technologies (ICCEET), IEEE, 2012, pp. 1032-1035.
[10]. X. Shu, J. Zhang, D. Yao, and W. C. Feng, "Fast Detection of Transformed Data Leaks," IEEE Transactions on Information Forensics and Security, vol. 11, no. 3, pp. 528-542, March 2016.
[11]. S. Sodagudi and R. R. Kurra, "An Approach to Identify Data Leakage in Secure Communication," in 2nd International Conference on Intelligent Computing and Applications, vol. 467, Singapore, 2016, pp. 31-43.
[12]. Elisa Costante, Davide Fauri, Sandro Etalle, Jerry den Hartog, Nicola Zannone (2016), "A Hybrid Framework for Data Loss Prevention and Detection", Security and Privacy Workshops (SPW), IEEEE
[13]. Sandeep Srivastava and Meenakshi Arora (2016) , "Mobile Agents: Objective, Platforms and Architecture (The Cutting Edge of Wireless Technology)", International Journal of Science and Research (IJSR) Volume 5 Issue 10, ,www.ijsr.net