International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 21 (2017) pp. 11546-11554

An advanced data leakage detection system analyzing relations between data leak activity

Min-Ji Seo[1]
Ph. D. Student, Software Convergence Department,
Soongsil University, Seoul, 156-743, Korea.
[1]Orcid : 0000-0002-7564-9264

Myung-Ho Kim[2]
Professor, Software Convergence Department,
Soongsil University, Seoul, 156-743, Korea.

Abstract
In order to prevent the data leakage by the internal staff, the companies protect important information of the company by inputting the behavior pattern related to the data leakage into the system in advance and by defining the employee as the staff who leaked the data, whose behavior pattern is detected when such inputted behavior pattern is detected. However, in the case of the existing system, if the data is leaked according to the pattern of the security log occurrence which is not inputted into the system, whether of the data leakage cannot be properly detected. Therefore, this study proposes a system to prevent the leakage of data in a data leakage pattern that is not input to the system by defining a set of security logs that can appear simultaneously at the time of data leakage through association analysis algorithm as a data leakage judgment scenario. As a result of experimenting the function of the system suggested, this study judged whether of data leakage with higher accuracy than the data leak detection system which does not apply association analysis algorithm, also it showed lower percentage of false positive and false Negative. This suggests that the proposed system is less likely to misjudge data leakage.

Keywords: Security Log Analysis, Apriori algorithm, Generating Data Leakage Detection Scenario, Convolutional Neural Network, Data Leakage Detection.

INTRODUCTION

In recent years, the size of company‘s average damage suffered by leakage of internal information has been gradually increasing. Most of such accidents are caused by internal staffs, contractor‘s staffs or retired staffs, so the importance of research to prevent the leakage of internal information is increasing [1]. Due to these a trend, currently, a method with which the company determines the data leakage by monitoring the work activities of internal employees is being studied. An intrusion detection system exists as a typical method, and it is divided into an abnormal behavior intrusion detection method [2] and misuse intrusion detection method [3]. First, the abnormal intrusion detection technique is to detect the behavior that deviates from the scope of the specified action by specifying the range of normal action from normal business activities. However, if the range of normal behavior specified is ambiguous, normal business activities also have the risk of being regarded as data leakage in the system. Compared with this, the misuse intrusion detection method is the one that detects the behavior similar with or same as the data leakage pattern inputted, by inputting the data leakage behavior pattern into the system which appeared in data leakage accidents in the past. Although the misuse intrusion detection method can correctly detect the data leakage behavior pattern inputted to the system, it cannot detect the data leakage behavior pattern which is not inputted into the system. As a solution to this problem, a system that can cope with the data leakage behavior pattern not analyzed by the manager is needed to prevent the data leakage by the internal staff.

Therefore, this study suggests a method to determine the data leakage through convolutional neural network [4] after writing the security log collected from the work activities of internal employees as graphs according to the scenario, by creating a scenario to determine data leakage with Apriori algorithm [5] that finds the association relationship between each data. Since the data leakage judgment scenario consists of a set of behaviors that can occur with the data leakage by applying the Apriori algorithm to the security log history collected in the past data leakage accident, it can judge data leakage even if the data is leaked through a behavior pattern that has not been previously entered into the system. In addition, since the security log pattern is analyzed with convolutional neural network which shows high performance for mass data processing and image recognition, data leakage staff can be identified with high accuracy in a short period of time.

This study is the version that extended the study of data leakage detection which used the associative relation between data leakage behaviors under review. The study of data

leakage detection which uses associative relation between data leakage behaviors suggested a process of creating the data leakage judgment scenario through the association analysis algorithm based on the virtual security log set. This study suggests a method to judge data leakage by comparing the behavior patterns of the internal employees with the data leakage staff, after creating a data leakage judgment scenario by analyzing the security log collected through the security device with the association analysis algorithm.

RELATED RESEARCHES

In the past, it mainly detected and blocked the approach for data leakage through misuse intrusion detection method or abnormal behavior intrusion detection method. Each detection method can be described as follows. The misuse intrusion detection method is a method of inputting a scenario of abuse pattern which accesses the data for misuse by a legitimate user or for malicious purposes and detecting the behavior similar with or same as such pattern; it is mainly used in experts systems, rule-based systems, and pattern matching systems. Currently, ITPT (Insider Threat Prediction Tool) [6] generated by G. B. Magklaras exists as an intrusion detection method. In the ITPT tool, it judged information misuse and intentional access by first recording the behavior patterns of existing information users, and comparing the behavior of each information user with the behavior patterns stored in the database through the monitoring criteria stored in the ITPT. However, in the case of the ITPT tool using the misuse intrusion detection method, a lot of supporting data is required to input the data leakage behavior pattern, and the security manager may have to analyze the respective ground data, which can take time cost.

The abnormal behavior intrusion detection method is the one of analyzing the behavior pattern of the user in the usual time and recording the same in the system, and notifying the administrator of the intrusion when a pattern causing a relatively rapid change is detected to prevent the leakage of data. At this time, the behavior pattern of the user at a normal time can be quantitatively defined by the administrator or can be defined by calculating the statistical value of the behavior pattern normally displayed.

In recent years, a method of predicting behaviors and commands to be taken next by each user after studying a behavior pattern of a user in usual time by introducing machine learning into an abnormal behavior intrusion detection method is being studied. Cha Byeongryae [7] analyzed the system calls generated by the daemon program or the root privilege program through the Bayesian network, and then notifies the administrator of the abnormal behavior when is different from the existing system call pattern. However, if the wrong data is used to create a behavior pattern of a user in the usual time, there is a risk that the data leakage behavior may be erroneously determined as a normal behavior pattern.

PROPOSED METHOD

Recently, accidents in which confidential information of company is leaked by internal employees happen frequently. Accordingly, companies analyze the data leakage pattern that has appeared in the past, inputs it to the system, and prevents the data leakage by using the misuse intrusion detection method which notifies the administrator when the same behavior as the inputted pattern is detected in the system. However, in the case of the misuse intrusion detection method, there is a problem that the system cannot properly detect the leakage when data is leaked in a behavior pattern not inputted to the system. Therefore, there is a need for a system that can prevent data leakage through behavior patterns that are not input to the system.

In this study, a method to create a data leakage judgement scenario is suggested by analyzing the security log collected from past data leakage incidents with Apriori algorithm. As an algorithm that finds association rules between data based on the frequency of occurrence of data, this study finds a set of security logs with relevance at data leakage and defines them as data leakage decision scenarios through Apriori algorithm. Then, the security log is collected from the work activities of the internal employees, and the security log generation graph is generated according to each scenario and a system for determining whether of data leakage is determined through convolutional neural network.

The suggested system consists of data collection stage, scenario generation stage, and data leakage detection stage as shown in Figure 1. In the data collection phase, the security log that can be generated by the internal employee's work activity is collected in each security device, and in the scenario generation stage, associative relation between the security logs collected in the past data leakage accident is analyzed and the behavior in deep associative relation is defined as data leakage judgment scenario. In the data leakage detection stage, it judges the data leakage employee by detecting the employee who shows high similarity with the security log generation pattern of the actual data leakage employees through convolutional neural network based on the data leakage judgment scenario defined at the scenario generation level.

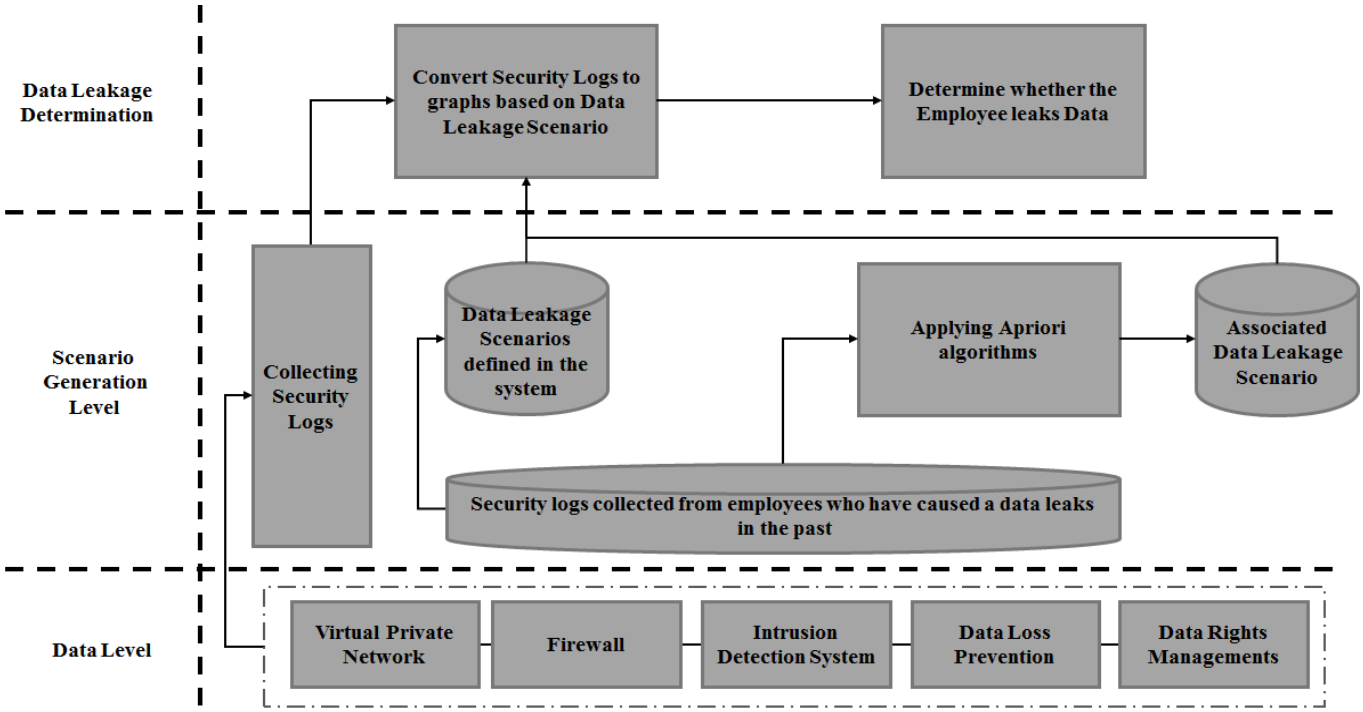

Figure 1 : Security control system using association analysis

Collecting Security Logs

First, the security logs output from each security system are collected as the basis data for analyzing the work activities of internal employees and defining the data leakage judgment scenario. The security log shows the content of the internal employee's business activities and can be collected from the security system defined in Table 1. The security log shows the leakage behavior, the list of leaked data, and the security log creation time. For example, if the leakage of data is caused by intrusion through the network or through the network transmission protocol, we can grasp whether the network traffic currently being transmitted through the security log collected by the web firewall, virtual private network, and network intrusion detection system is an authorized IP address, or whether the currently used network protocol is an authorized protocol.

Therefore, in the system suggested, the security log is analyzed and the pattern of behavior that can appear in data leakage is defined as a scenario. Through the data leakage judgment scenario defined in this study, the administrator can analyze each scenario when data leakage is determined and appropriately deal with it according to the data leakage behavior pattern.

Table 1: List of systems collecting security logs

Security equipment	Function
Virtual Private Network(VPN)	Keep mutual authentication and communication in communication between computers or computer and network
Patch Management System	Manage software update records for responding to the vulnerabilities found in operating systems or applications
Data Loss Prevention(DLP)	Monitor attempts to access non-business sites during work hours or attempts to leak specific information outside
Firewall	Store all access records with software that controls access to the PC using an external network
Data Rights Management(DRM)	Grant authorized users access to data only within the scope of authorized permissions
Database	Manage confidential information and keep access records

Creating the Data Leakage Judgement Scenario

This study consisted of the scenario in which the leakage behavior mainly appeared in data leakage accident and the scenario analyzed with connectivity between security logs. First, the scenarios are defined by analyzing the behavior pattern that appears mainly in the data leakage accident for data leakage judgment. In general, certain behaviors, such as 'downloading documents' and Document release using USB, are the ones that can occur frequently in data leakage behavior and can be the one having the risk of data leakage. Therefore, it is necessary to input a specific behavior pattern mainly appearing in the data leakage accident such as 'document download' and Document release using USB into the system as a dangerous behavior, and to identify an internal employee who shows a similar behavior pattern as a data leakage employee. In this study, the following data leakage judgment scenarios are defined based on the data leakage behavior pattern, which was mainly observed in past leakage accidents:

1) An unauthorized employee tries to bypass the VPN to access the data, accesses the internal private network, and leaks data.

2) Data was leaked as internal employee accessed to the site unauthorized by the company, and such data was infected by malicious code, although there was no intention of the internal staff to leak data:

3) Internal employee access shared folders to access documents managed by the company, and copy the files to their PC for leakage purposes, or delete document files that affect their HR scores.

4) Leak the document file containing the information of the company to the outside via the messenger, and upload the document to the web hard with the malicious purpose.

5) Leak the document file containing the information of the company to the outside via the messenger, and upload the document to the web hard with the malicious purpose.

6) Attempt to log out from the DRM Agent in order not to be affected by the DRM security device managing the document access authority, or leak it to outside through a mobile device such as USB or output DRM document using printers after downloading DRM documents.

7) Leak the confidential information through e-mails after inquiring the query with access to the database that controls the information of customer and human resources of the company.

However, when the data leakage is judged only by the behavior pattern frequently appearing at the time of data leakage, there is a problem that it cannot be determined whether the data is leaked properly if the data is leaked through a path not yet inputted into the system. Therefore, in the system suggested, the association analysis algorithm is applied to the security log history that occurred in the past data leakage accident, and the data leakage of each internal employee is judged through the data leakage judgment scenario with a set of security logs that can be displayed together when data leakage is leaked. The association analysis algorithm can represent a data leakage behavior pattern that the security administrator does not understand, so that the data leakage behavior pattern can be identified more flexibly than the existing data leakage judgment scenario.

The association analysis algorithm used in the system suggested is an Apriori algorithm, which searches for the rules that can cause simultaneous occurrence of items that affect the occurrence of a specific event by calculating the support and confidence. What each of the numerical values represents is as follows. First, the support level is the number of times the security log corresponding to the behavior A and the behavior B occurred in the total number of occurrence times of the security log. A high value of support means that the ratio of security log A and security log B is high among internal employees' work activities. On the other hand, the confidence level is the rate of occurrence of the security log for B, which is another behavior when the security log for a specific action A occurs. The Apriori algorithm can analyze the support and confidence values of each security log and define a highly relevant set of each security log as a set of behaviors likely to be associated with each other at the time of data leakage.

Table 2 shows the security logs collected from three leakage employees collected from past data leakage incidents. When the association analysis is performed based on only the activities performed during the work of each employee, after determining whether or not each employee has leaked data, the system cannot grasp which folder the employee leaks the data and which document file is leaked do. Therefore, it is necessary for the manager to inspect each employee's work activities in detail, which makes the efficiency of data leakage discrimination low. In the proposed system, after the data leakage staff is identified, the file and folder that is accessed at the time of business activity is specified in the security log list to which the Apriori algorithm is applied so as to efficiently determine the file and folder accessed by the employee, Through the association analysis, the administrator can efficiently identify the behavior scenarios and the files accessed by the employee. Therefore, the security log list shown in Table 2 is also composed of each employee's work activities and access folders and files, and shows the behavior scenarios and access files and folders of the employees through association analysis.

Table 2: Item-1 set focused on business activities and access folders

Transaction	Security system	Folders and files	Behavior
	Host-based Intrusion Detection System	/var/log/secure	Access to server
		/var/log/secure	Access to root account
		/var/log/secure	Password error
	Database	/usr/libexec/mysqld	Access to root account
		/usr/libexec/mysqld	Query the schema Table
		/usr/libexec/mysqld	Department information inquiry
		/usr/libexec/mysqld	Fix your own risk
	Data Rights Management	C:\Users\home\Documents	Access to DRM
			Download confidential documents
			Create personal documents for users
			Save personal documents for users
			DRM logout
	Data Loss Prevention	C:\Users\home\Documents\logicaldoc_mydlp.xlsx	Copy confidential documents to USB
		C:\Users\home\Documents\logicaldoc_mydlp2.xlsx	Copy confidential documents to USB
		C:\Users\home\Documents\logicaldoc_mydlp3.xlsx	Copy confidential documents to USB
	Messenger	Include Confidential Documents	File transfer
		Include Confidential Information	Send message
T2	Host-based Intrusion Detection System	/var/log/secure	Access to server
			Attempt to access Secure Shell
			Succeed to access Secure Shell
			Access to root account
	Data Rights Management	C:\Users\home\Documents\logicaldoc_mydlp2.xlsx	Access to DRM
			Download confidential documents
			DRM logout
	Data Loss Prevention	C:\Users\home\Documents\logicaldoc_mydlp2.xlsx	Connect a USB device
			USB transmission
T3	Host-based Intrusion Detection System	/var/log/secure	Access to server
			Access to root account
			Attempt to access Secure Shell
			Succeed to access Secure Shell
			Change the Host Security log file
			Change the mail log file
	Data Rights Management	C:\Users\home\Documents\logicaldoc_mydlp2.xlsx	Access to DRM
			Download confidential documents
			DRM logout
	Mail	Include Confidential Information	Data transmission

The process of applying the association analysis algorithm to the security log history of Table 2 is as follows. First, each security log shown in Table 2 is defined as a set of security log item-1 for analyzing associative relation, security system names such as 'server connection' and 'password error', folders and files. In this case, if the item-1 set indicates a value smaller than the minimum support, the item set should be removed. To create a set of security logs with relevance, item-1 sets satisfying the minimum support are linked to other item sets to generate item-2 sets. From the set of item-2 confidence is calculated and the association between the elements constituting each item set is analyzed, and if the minimum confidence is not satisfied, such item set is removed. The minimum support and minimum confidence used in the association analysis algorithm is the number specified by the security administrator to reduce the number of items to be computed in the system suggested. Since the security log entries that do not satisfy each minimum support and confidence are removed and the association is analyzed, the association between security log entries can be detected more efficiently when analyzing all security log entries. In the same way, the calculation of confidence and support are repeated until the association is analyzed for all security log entry sets.

After analyzing the security log history in Table 2, ["Server logout", "Root account access", '/var/log/secure", 'DRM access', 'Download confidential document', 'Data transmission via mail'] is finally generated. Through this, the data leakage judgment scenario of 'after the internal employee accesses the DRM system to download the confidential document, the data leakage using mail' can be newly defined.

In addition, it is possible to cope with data leakage more efficiently at data leakage with availability for identifying the type and contents of documents accessed by employees who leaked data to the outside, through a set of 'action' and 'access file and category' items that can be derived by analyzing the association between security logs.

Determining Data Leakage Scenario Graphs

When the definition of the data leakage discrimination scenario is completed, the system suggested analyzes the security log collected with convolutional neural network through the process in Figure 2. Convolutional neural network is a deep learning algorithm that extracts features of an image through convolution layer and pooling layer, and then learns the features extracted by combining neural networks or judges the image. In the convolution layer, a feature map representing the characteristics of the image is created, shifting a specific filter to the original image and integrating the result of multiplying the image. In this study, a 5 by 5 size filter was used and the filter was applied by skipping one space at a time. In the pooling layer, the space of the feature map created in the convolution layer is reduced through the Max-Pooling method, which represents the largest value in each image area. Then, the neural network layer and the softmax classifier learn the corresponding image through the image feature map created in the convolution layer and the pooling layer, or classify the input image by finding an image having a similar feature map, and let the user know the result of what kind of image it is.

In the proposed method, in order to determine the data leakage based on the security log collected from each internal employee's business activity, the security log generated by the internal staff during the business activities is drawn as a graph in accordance with the data leakage judgment scenario defined in this study. The security log constituting each data leakage judgment scenario is created as a graph of the number of occurrences of the security log according to time. For example, the scenario ['Server logout', 'Root account access', 'DRM access', 'Confidential document download', 'DRM logout', and 'Data transmission via e-mail'] that was defined by applying Apriori algorithm to the set of security logs in Table 2 should consist of a set of graphs of the times of occurrences of the security log corresponding to each action.

The graph of the times of occurrence of the security log constituting each scenario is classified into a data leakage behavior graph of a normal employee and a data leakage judgment scenario of a data leakage employee, and is studied by convolutional neural network. Based on the graph data learned in convolutional neural network, it is judged whether or not the graph obtained by collecting the security log from the internal employee's work shows a pattern similar to the graph of data leakage employee, and if the similarity is shown as high, the employee is identified as data leakage employee.

When an employee who is suspicious of having a data leakage is identified through the notification from the system, the administrator can identify the data leakage behavior path based on the security log occurrence pattern of such employee, so that the data leakage can be responded appropriately. First, it is judged whether or not data leakage is suspected in the graph based on the data leakage detection scenario defined in the system, and data leakage behavior of the internal employee is judged. If the data leakage suspicion is not detected in the graph created based on the data leakage judgment scenario defined in the system, you can effectively respond to data leakage as soon as possible, as you can grasp data leakage behavior that each scenario represents when data leakage suspicion is detected in the corresponding scenario graph, by judging whether or not data leakage suspicion is determined in the data leakage judgment scenario graph created by the security log association analysis.

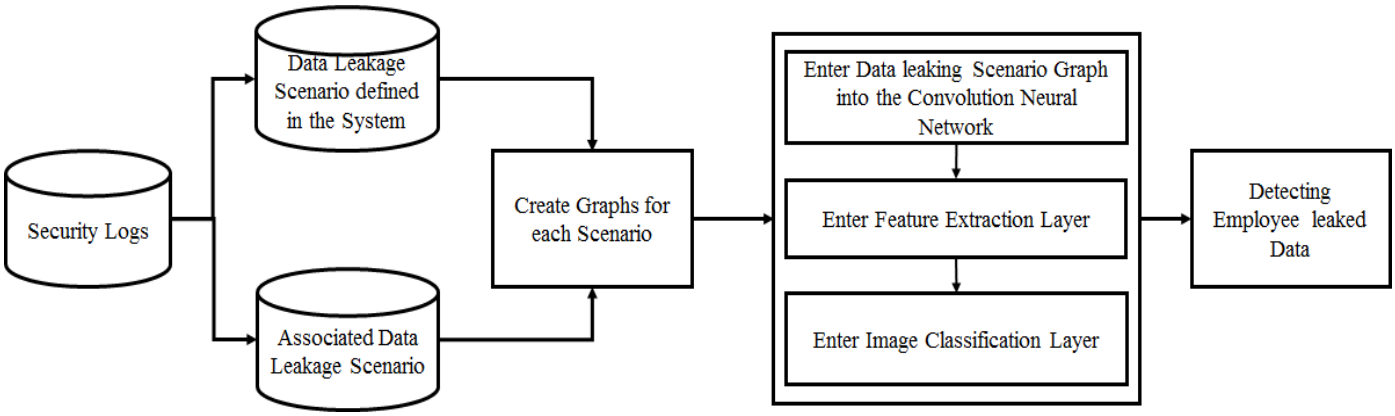

Figure 2: The process of determining data leakage employees with convolutional neural network

Performance Evaluation

In this study, a data leakage decision scenario is defined using the association analysis algorithm in the security log, and a method to improve the detection performance of data leakage is suggested. In the previous study, it was determined whether data leakage was similar to the behavior pattern of data leakage employee by only the data leakage judgment scenario entered in the system. However, the existing data leakage discrimination system has a disadvantage that it cannot correctly judge data leakage if the data is leaked in a behavior pattern not inputted into the system.

Therefore, in this study, a system to judge data leakage by creating a new type of data leakage judgment scenario is suggested by analyzing the association between security logs of employee appeared in previous accident history by using association analysis. Data leakage is determined by analyzing if the graph of security log generation for each scenario shows a pattern similar to data leakage employee through convolutional neural network.

The system suggested is implemented in Docker based Tensorflow framework environment, and whether it can detect data leakage is tested based on the security log collected from the actual internal staff. After collecting the security log from each internal employee and creating the security log graph according to each data leakage judgment scenario, it verified that it can detect whether data leakage is detected through the judgment result output from convolutional neural network. The security log is collected from a syslog that records events that occur on the operating system, a network-based intrusion detection system that records traffic occurring on the network, a data loss prevention and data rights management solution that records overall business activities, such as access logs of the database holding important information of the company, and document download, and the names and functions of the solution are shown in Table 3.

Table 3: Security log collection Open source software

Security log collection software	Function
OSSEC	System log analysis and file integrity check occurring on each operating system
MYSQL	Access, modification, history detection to the database storing confidential information, corporate information, customer information
MYDLP	Check and block file transfer paths in CD, memory, external disk, mail, P2P, web hard, instant messenger, and printer
LOGICALDOC	Encrypt the file itself that contains sensitive information and allow it only within the allowed scope of permissions
BRO	After collecting network packets, it outputs the pattern of intrusion and network usage through packet analysis.
SNORT	After assigning network analysis rules to the system, it records traffic analysis and packets in real time.

The security log collected from each security device is created as a graph of security log occurrence count based on the data leakage judgment scenario defined in this study. The data leakage judgment scenario consists of scenarios defined as behaviors that can frequently appear in data leakage and scenarios defined as behaviors that can appear in relation to data leakage in association analysis algorithm. The graph created according to each scenario is input to convolutional neural network to determine whether it is data leakage, and if the pattern is similar to the graph of data leakage employee, the system defines him/her as the data leakage employee and notifies the manager of such employee.

In order to verify the function of the system suggested, the accuracy of data leakage discrimination was compared according to the number of employees who analyzed the existing data leakage judgment system and the security log. The experimental results are shown in Figure 3. In the existing data leakage detection system, it can be seen that the accuracy of the existing data leakage detection system is significantly lowered as the number of employees to judge data leakage increases, As the system suggested improves the function of the data leakage decision scenario using association analysis, which means that data leakage detection accuracy is 95% or more regardless of the number of internal employees.

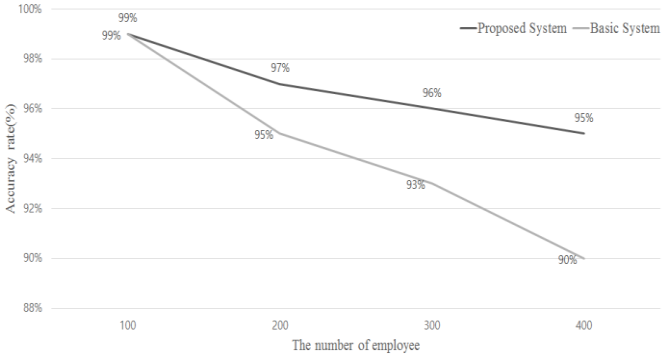

Figure 3: Data leakage discrimination accuracy

In addition, it is also tested to see if data leakage can be correctly discriminated when data is leaked by a behavior pattern that is not input to the system, and the data leakage behavior pattern is defined as shown in Table 4.

Table 4: Data leakage judgment scenario not defined in the system

Data leakage Judgment scenario number	Data leakage Judgment scenario
A	Copy document after connecting with Secure Shell
B	Download document using provided an account
C	Change the file format to copy
D	Capture all databases information and document files to screen shots and copy them to USB

In order to evaluate the function of the system suggested, this study compared the values of false positive rate and false negative rate with the system that did not update the scenario with association analysis whether or not it can identify the employee who leaked the data through the scenarios in the Table 4. Although the system judged false positive as data leakage, it represents the number of cases when the corresponding employee did not actually leak the data. Also, False Negative is the number of cases where the employee who leaked data actually does not detect the corresponding employee in the system. Through the measurement of false positive and false negative, it is possible to determine whether the system accurately detects data leakage of the internal employee.

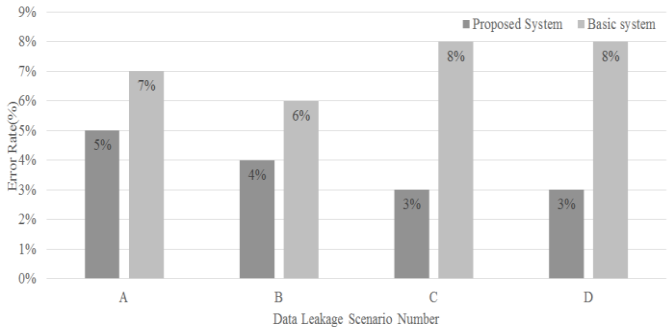

Figure 4: Comparison of False Positive Rate with Existing Systems

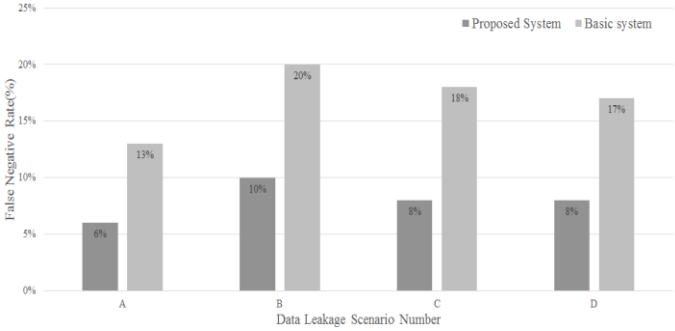

Figure 5: Comparison of False Negative Rate with Existing Systems

Figures 4 and 5 show the rate of false positive rates and false negative based on the data leakage behavior pattern in Table 4. Based on the experimental results, This study could find that data leakage is discriminated more correctly in the case that data leakage judgment scenario is created by applying association analysis algorithm to security log with lower value in false negative, false positive than the case of discriminating data leakage by using only data leakage judgment scenario provided by administrator.

In addition, unlike the existing system which was able to grasp the sequence of data leakage behavior by analyzing each security log through the created data leakage judgment scenario, in the system suggested, it is possible to judge data leakage more effectively as it makes us able to judge through what behavior path the employee suspicious of data leakage accessed to data and leaked the data to outside, through the data leakage judgment scenario graph.

CONCLUSION

Recently, confidential information held by a company is leaked by an internal employee, a partner company employee, or a retired employee. Accordingly, a system is being studied that discriminates data leakage of internal employees with a method that in advance inputs the behavior patterns, which frequently occurred in the data leakage accident, into the system, and notifies the administrator of such behavior pattern when the behavior pattern previously entered in the system is detected. However, the existing system has a disadvantage in that it cannot detect whether the data leakage is correctly detected when the data is leaked to the data leakage path which is not analyzed by the administrator.

This study suggests a system to detect data leakage by writing a data leakage scenario based on the security log collected from each employee's work activity to cope with data leakage, and creating a graph according to the scenario. The data leakage judgment scenario consists of scenarios defined as behaviors that can frequently appear in data leakage and scenarios defined as behaviors that can appear in relation to data leakage in Apriori algorithm as association analysis algorithm. Therefore, this study first extracts the behavior patterns that appear frequently in the data leakage event, and writes the scenarios through the analysis of the administrator and inputs them into the system. Then, this study defines the pattern of behavior likely to appear together with data leakage by applying the association analysis algorithm to the security log collected in the past data leakage accident and create it as a scenario. In the system suggested, it is possible to easily understand through what action the employee leaked data, even when data is leaked through a new leakage path, using the data leakage judgment scenario defined through the association analysis algorithm. In addition, a security log graph was created for each scenario through convolutional neural network, and data leakage suspicion of each internal employee was judged. Since the convolutional neural network, which is deep learning algorithm that shows high function in image recognition, is used, it is possible to discriminate the data leak more accurately than when the data leakage is detected through other security solutions.

In order to judge the function of the system suggested, this study experimented whether it can detect data leakage based on data leakage decision scenario defined by analyzing associative relation of security log when data was leaked in behavior pattern that was not entered into system. As a result of experiment, In the existing data leakage detection system, it can be seen that the accuracy of the existing data leakage detection system is significantly lowered as the number of employees to judge data leakage increases, but, the system suggested improves the function of the data leakage decision scenario using association analysis, which means that data leakage detection accuracy is 95% or more regardless of the number of internal employees. This study could find that data leakage is discriminated more correctly in the case that data leakage judgment scenario is created by applying association analysis algorithm to security log with lower value in false negative, false positive than the case of discriminating data leakage by using only data leakage judgment scenario provided by administrator.

For the future research, we will study a plan to create a data leakage judgment scenario more efficiently by improving the association analysis algorithm.

ACKNOWLEDGEMENT

This paper was supported by the Small and Medium Business Administration's First Step Technology Development Project (C0394819).

REFERENCES

[1] "Global Data Leakage Report", InfoWatch Analytical Center, Moscow, 2017.

[2] Y. Yu, "A survey of anomaly intrusion detection techniques", *Journal of Computing Sciences in Colleges*, vol. 28(1), pp 9-17. 2012.

[3] S. Kumar, E. H. Spafford, "A pattern matching model for misuse intrusion detection", in *the Proceedings of the 17th National Computer Security Conference*, Baltimore, Maryland, 1994.

[4] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella and J. Schmidhuber, "Flexible, High Performance Convolutional Neural Networks for Image Classification", in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, Barcelona, Spain, 2011.

[5] M. Kantardzic, "Data mining: concepts, models, methods, and algorithms", USA: John Wiley & Sons, 2011.

[6] G. B. Magklaras, S. M. Furnell, "Insider threat prediction tool: Evaluating the probability of IT misuse", *Computers & Security*, vol. 21(1), pp. 62-73, 2001.

[7] B. R. Cha, K. W. Park, and J. H. Seo, "Modificated Intrusion Pattern Classification Technique based on Bayesian Network", *Journal of Internet Computing and Services*, vol. 4(2), pp. 69-80, 2003.