International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13321-13327

A Detection Method of Data Leakage by Cooperation of Insiders

Hee-Jin Shin
M.S. student, Software Convergence Department, Soongsil University, Seoul, 156-743, Korea.
Orcid : 0000-0002-9492-5033

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

Abstract

Due to the development of IT technology, many internal data leakage accidents are occurring due to changes in business environment. To prevent the leakage of internal data, companies are using security solutions based on individual leakage attempts. However, because many internal data leakage incidents are caused by multiple malicious collaborations, existing security solutions alone are difficult to detect. Therefore, in this paper, we propose a method to detect internal data leakage by creating a group using HR information, network information, and security solution log of internal users. We also use a security solution log to define a single scenario and a composite scenario for determining leakage. A single scenario represents a user's behavior that can occur in each security solution, and a combined scenario represents a pattern of users that can come together in a single scenario. The security risk of the previously created group is calculated using a single scenario and multiple scenarios and detected as an internal data leak when the score is above the baseline score. In order to verify whether the proposed method effectively detects the leakage of internal data by cooperation of users, it is compared with existing security solutions and it is confirmed that it can detect virtual scenarios that did not detect existing security solutions.

Keywords: Insider Threat, Date Leakage Prevention, Scenario of Data Leakage, Grouping.

INTRODUCTION

The development of IT technology has changed work environment of enterprise, such as cloud and Bring Your Own Device (BYOD). As a result, access to and leakage of internal data becomes easier, and the incidence of internal data leakage accidents is increasing. According to the Survey on the Level of Technology Protection by the National Intelligence Service Industry Confidentiality Protection Center, outflows by internal employees account for a large percentage of data leakage incidents, accounting for more than 80%. Internal data leakage causes not only economic deterioration of company image and economic damage, but also leakage of confidential data to foreign countries causes problems of weakening technical competitiveness of nation [1].

Companies use many security solutions such as Digital Rights Management (DRM) [2], Data Leakage Prevention (DLP) [3] and Security Information and Event Management (SIEM) [4] in order to reduce internal data leakage and minimize the damage. These security solutions are a way to detect the behavior of an individual trying to leak data, such as unlocking the security of documents or sending confidential documents by e-mail. However, internal data leakage is caused not only from the attempts by individual employees to leak, but also from the simulation in a large number of internal employees attempting a planned leakage. According to the arrest status of industrial technology leakage offense in National Police Agency, 86 cases of data leakage occurred in domestic in 2015, but only 223 offenders were arrested. This shows that one or more internal employees were involved in the outflow in an incident [5]. Typically, there is an accident in which the leakage of electronic circuit design programs of domestic small and medium enterprise occurred in 2013. In this case, four internal employees, including the vice president, leaked the electronic circuit design program and set up a same kind of company, and then again leaked the technology to a Japanese company. As in this case, the leakage of internal data caused by multiple malicious collaborations is harder to detect than if it was caused by an individual. The later the detection time is, the more damage will be to the enterprise and the nation. Therefore, even in the case of data leakage accidents caused by cooperation of several persons rather than individuals, it should be possible to minimize the damage through quick detection.

This study proposes a method to detect the leakage of internal data by creating a user group and calculating the security risk of the group in order to detect the leakage by malicious collusion of several persons. In the proposed method, a group is created and detected using user information, so that it can be detected from the viewpoint of the group even if it is not judged as a leakage from an individual viewpoint. In addition, since the group is created using a combination of the user's personal information, the personal relationship information, and the history of the use of the file or the mail, the leakage attempt that cannot be detected from a simple viewpoint such as the user's department or position can be detected.

This study is composed as follows. In Chapter 2, this study summarizes related researches to detect leakage of internal

data, and in Chapter 3 describes the process of designing and implementing a method to detect leakage by cooperation of internal users. In Chapter 4, this study evaluates the proposed method through experiments using hypothetical scenarios, and in Chapter 5 summarizes conclusions and future directions.

RELATED RESEARCHES

Representative security solutions to prevent leakage of internal data include DRM, DLP, and SIEM. DRM (Digital Rights Management) solution prevents data leakage from occurring by granting privileges to each user, and manages history such as creation and modification of files to monitor data leakage [2]. Data Leakage Prevention (DLP) allows to block access to corporate data in advance or to grasp the history of data use [3]. Security Information and Event Management (SIEM) has been developed to compensate for data leaks by DRM and DLP [4]. The SIEM solution is used to collect and centrally integrate and analyze logs from each security solution.

In addition, there are leakage prevention systems based on insider risk analysis and data leakage scenarios to prevent internal data leakage [6, 7]. Insider risk analysis calculates risk by evaluating people, values, vulnerabilities, etc. to insider's risk level or whole institution's risk level. The leakage prevention system based on the data leakage scenario analyzes log generated by each security solution and applies it to the data leakage scenario to detect leakage of internal data.

Existing methods detect internal data leakage based on an individual's behavior pattern. However, these methods have a problem that cannot be detected when a plurality of users cooperate and attempt to leak. Therefore, this study proposes a method that can detect even when several users try to leak data.

This study proposes a method to detect leakage by making a group of users who are likely to leak data to detect leakage attempts by users' malicious cooperation, and analyzing the behavior pattern of such group. The proposed method is divided into three steps: log collection step, user grouping step, and internal data leak detection step. Figure 1 shows how data leakage is detected by the cooperation of internal users. It collects logs for each security solution, groups the users by using collected logs and user information, calculates risk and judges internal data leakage.

Collecting log and user information by security solution

The security solutions that collect logs in this study are DB security solution, File Server, DRM, DLP. The logs collected from each device are used to find associations of users grouped into personnel information and network information. Also, to detect internal data leakage, user behavior is defined as a single scenario, and it is used to calculate the risk of each behavior as a security risk. Each log should contain information, including the IP or employee number that identifies the user, and should include behavioral information that can match the behavior between the device and the scenario. Table 1 is an example of the log generated in the DB security solution, and shows the log when userA tried to return more than 10000 lines of data from the DB named as SAMPLE, but failed due to access to too much data.

The user information consists of the user's personnel information and the personal relationship information, and the personnel information is the employees information managed by the company such as the department or the position of the user, and personal relationship information is user‘s personal information such as user’s alma mater or region. It is used to identify groups of users who may possibly collaborate and leak data.

Analysis and design

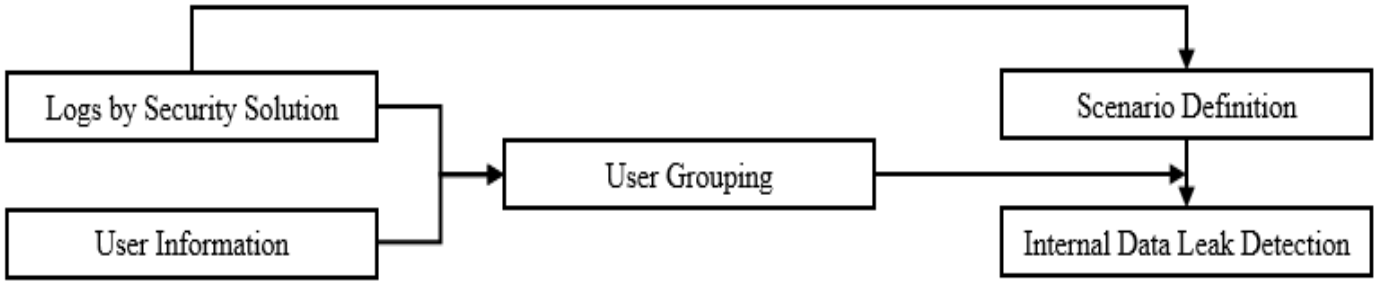

Figure 1: A Method of Data Leakage Detection by Cooperation Internal Users

Table 1: DB Secure Solution Log

Column name	Example
ALERT_TYPE	DBSECU_ALERT
SQL_TYPE	SELECT
TIME	2017.07.14. 15:03:27
EVENT	Too Much Return Data
SERVER_IP	203.253.xxx.xxx
SERVICE_NAME	DSSDB
SERVICE_PORT	22
SESSION_ID	10
CLIENT IP	203.253.xxx.xxx
DB_SERVER_IP	203.253.xxx.xxx
DB_NAME	SAMPLE
DB_USER	userA
STMT_ID	30
RETURN_ROWS	10000
USER_ID	userA

User Grouping

A user who decides to leak internal data proposes a leak of data to a common or similar person based on a network, such as department, position, or origin. After that, they cooperate with each other to collect and modify the data, share it with each other, and attempt to leak it. Therefore, this study classifies the users who can leak data based on the risky users who have data leakage motivation, and analyzes the history that the users collected and modified the data and the history of communication with each other. Users based on grouping include retirees, retirement prospects, person dissatisfied with position treatment, and core job managers. In the proposed method, not only employee information such as department or position, but also user relationship information such as user's school, living area, etc. are used to classify users. After that, the group is determined by judging which information has been accessed through the user's file, mail history and contacted persons.

Figure 2 shows the pseudo-code of the user grouping. In the first step, grouping is performed using the personal information and the relationship information of the user based on the users having the data leakage motive. Grouping uses personnel information such as departments, positions, and date joining the company, as well as personal information such as alma mater and the area of origin. The reference user and the users having one or more common points are to be collected and created as one group.

In the two-step grouping, grouping is performed using the history of the user's file usage. This study groups the people who used the same files as the users in Step 1 or who use related files into a group. Relevant files include when files exist in the same file server, or when a person uses a file and modifies it to create a new file. For example, it is the same case when employee A downloads 'confidential document .doc' from the file server and then saves it as 'general document .doc', and employee B sends 'general document .doc' by mail.

Step 3 grouping creates a group of three levels by using the history of users' mail or messenger usage. When someone contacted anyone in the group, add such person as a group and exclude the user who has no contact history from the group. If the person who is communicating with is using an external mail or is not an internally identifiable employee, consider such person as a dangerous person and add him/her to the group.

```
Algorithm UserGrouping
Input Users, DRMLog, DLPMailLog
//Users : id, name, department, position, address, school(high school, university)
//DRMLog : id, sourcefile, destinationfile, path, state
//DLPMailLog : id, usermail, receiver, receivermail
Output UserGroups
//UserGroups: Grouping of those related for each user
//UserGroup_i : ist user's group list
Begin
  for each user_i in Users
    /*risk status: Retiree, destined retiree,
    dissatisfied with position, core manager*/
    if user_i is in risk,
      //b_i: Judger if user is in risk, if it is 1, he is risky user
      set b_i = 1
  for each user_i in Users
    if b_i == 0
      UserGroup_i += user_i
    else if b_i == 1
      //UserGroup_i : user_i is reference user group
      UserGroup_i += user_i
      //Grouping using user network
      for each user_j in Users
        if user_i and department, position, residence,
          when one is the same among alma mater(High school, college)
            UserGroup_i += user_j
      //Grouping by file use history
      for each drmlog_k in DRMLog
        if not used the same file as user_i
          && if it is not id included in UserGroup_i,
            UserGroup_i += id of drmlog_k
          if not used the same file as user_i
            && if it is id included in UserGroup_i,
            UserGroup_i -= id of drmlog_k
      //Grouping using mail use history
      for each dlplog_1 in DLPMailLog
        if exchange mail with user_i,
          && it not id included in UserGroup_i,
            UserGroup_i += id of dlplog_1
          if not exchanged mail with user_i
            && if it is id included in UserGroup_i,
            UserGroup_i -= id of dlplog_1
End
```

Figure 2: Pseudo Code of User Grouping

Internal Data Leak Detection

Single Scenario Definition and Security Risk Setting

Table 2: Single Scenario and Security risk for DB Security Solution

Security Solutions	Single Scenario	Security risk
DB security	DB access permission	1
	DB login failed	3
	DB access denied	3
	Using Select Query	3
	Data return successful	3
	Return time greater than 10 seconds	5
	Number of returned data exceeds 10000	5
	Using confidential data access queries	10
	Using Drop queries	10

It is a step of analyzing the log collected by each security solution, defining user actions that can be judged through the log, as a single scenario, and determining the risk level for each scenario. That is, a single scenario represents what the user has done in a security solution, and the security risk of a single scenario represents how dangerous the action is. In the proposed method, the security risk is set at 1 point when it is not the most dangerous, and 10 when it is judged to be the most dangerous. For example, in Table 2, a single scenario that can occur in a DB security solution is authentication to connect to a DB, and access is permitted or denied depending on the result of authentication. Alternatively, when a user uses a DB, it can be expressed as a single scenario using a general query such as Select, Update, or a query having a large effect on data such as Drop or Alter. The security risk of the DB security solution may be accidentally caused by authentication failure or denial of access to the DB system. However, since it is dangerous if it occurs several times, the security risk level is set to 3 points. In addition, it was judged to be very dangerous when using a query that had a large effect on data, so it was judged to be 10 points. If the number of data or the return time is exceeded by attempting to return other data too much, it may appear as a result of dangerous action, but since there is no returned data, it is set to 5 points. Successful data return can be included in the general business, but this can also be dangerous if it occurs many times occur and the security risk is 3 points.

Defining complex scenarios and setting security weights

A compound scenario is a scenario created by combining two or more single scenarios. For example, when one or more user actions occur, such as '10 connections' for a single scenario, 'DB connection' such as 'userA accesses DB 10 times', it is judged as a combined scenario. The security weights of the combined scenarios are used to determine the risk of each action with a value between 0 and 1.

In the proposed method, the security weights of the compound scenarios are divided into three steps. Each step includes an data collecting step to collect data for the leakage, an data processing step to release or modify security of the collected data, an data leaking step to leak data outside, such as storing the file into the USB or outputting the file. In the data collecting step, DB security solution and File Server solution are used, and in data processing step, DRM solution is used, and in data leak step, DLP solution is used. The security weights of the compound scenarios consist of three steps and risks are divided into Good, Fair, and Poor by number of times. By making the most dangerous case Good, its security weight is 1, and Fair is the security weight of 0.5, and the security weight of the Poor is 0.1. Table 3 shows the combined scenarios and security weights of a single scenario of data return success among DB security solutions. The security weights of 1, 0.7, 0.3, and 0.1 were assigned to the four scenarios grades in Good, Mediate Good, Mediate Poor, and Poor according to the number of single scenarios cases.

Table 3: Multiple Scenarios and Security Weights for a Single Scenario

Single Scenario	Compound scenario	Security Weights (Grades)
Data return successful	Less than 10 occurrences	0.1 (Poor)
	More than 10 and less than 30 cases	0.3 (Mediate Poor)
	More than 30 and less than 50 cases	0.7 (Mediate Good)
	More than 50 cases	1 (Good)

Security risk calculation and internal data leakage group detection

In order to detect user groups cooperating with the leakage of internal data, the security risk of the group is calculated and compared with the reference score to determine whether or not the internal data is leaked. Figure 3 shows how to calculate the security risk to detect an internal data leakage group.

First, the security risk of each complex scenario step is obtained. The step-by-step security risk is the value which multiplied the security weight by the total of security risks incurred in the log of the security solution belonging to each of the compound scenario step. For example, in a DB security solution belonging to the data collecting step, if the data return success log occurred 20 times, the security risk of the data collecting step is 3*20*0.3 = 18 points.

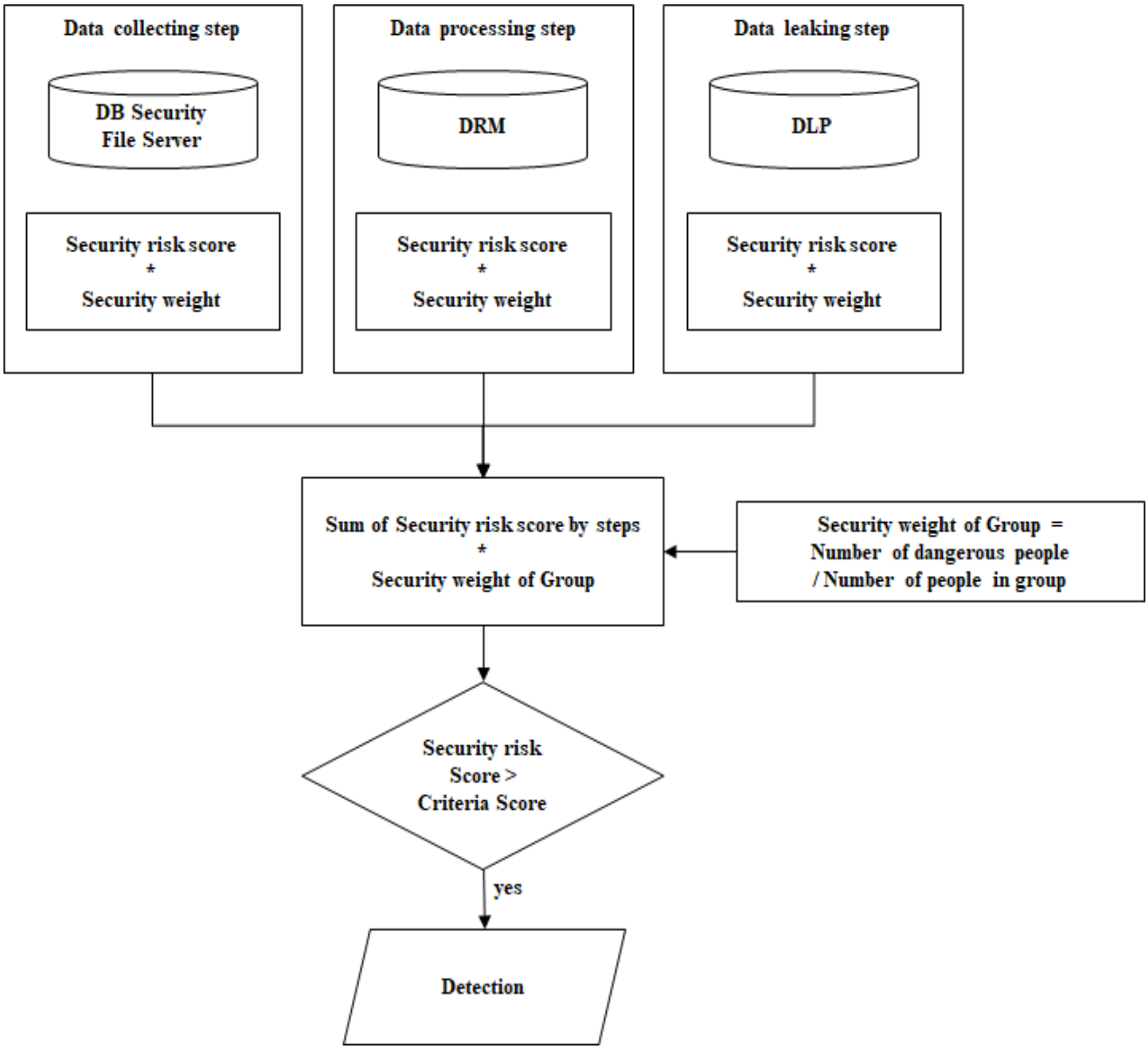

Figure 3: Calculation of Security Risk Score for Detecting Internal Data Leakage

Combine the security risks for each step and multiply this number by the security rating of the group. The security level of a group is calculated by dividing the number of dangerous persons belonging to each group by the total number of persons in the group. The dangerous persons consist of retirees or destined retirees, persons suspicious of alleged leakage, key employees, employees of partner companies, non-cooperators in the company, and prospective employees. For example, if the total number of people in the group is 5 and the number of destined retirees are two, the security rating of the group will be 0.4. In the proposed system, the security level of the group is non-zero between 0 and 1 because the group is based on the dangerous person. If the sum of the security risks at each step is 50 and the security level of the group is 0.4, the security risk of the group is 20. If it is higher than the reference value compared with the reference score, it is detected as a leakage. For example, if the reference score is 10, the above example is detected as a leakage.

RESULTS AND PERFORMANCE EVALUATION

To verify the performance, this study sets up a virtual scenario and describes the implementation result according to the design process. The purpose of the experiment is to group users in the company and detect if there are any groups attempting to leak internal data.

Describe the user grouping steps in order. First, the baseline user for grouping the user is a dangerous user having the motivation for internal data leakage as described earlier. The baseline user of the virtual scenario is composed of retiree

'user35', destined retiree 'user2', 'user3', 'user8', and personnel dissatisfied with position 'user18', 'user31', and 'user32'.

Among seven baseline users, 'uesr2' is selected as the baseline user and the grouping step is described.

Table 4: Result by Grouping Step

Grouping step	Result of grouping by 'user2' reference
1st Step	'user2', 'user8', 'user14', 'user18', 'user19', 'user29', 'user30', 'user34', 'user35'
2nd Step	'user2', 'user8', 'user18'
3rd step	'user2', 'user8', 'user18', 'nonuser1', 'nonuser2'

In the first step, 'user2' and total of 9 users having a common point in department, position, and alma mater are defined as a group. In the 2nd step, 3 users out of 9 internal employees that were determined in the first grouping, are grouped after removing six users who did not use the associated file. In the 3rd grouping, grouping is done through the mail usage log of the DLP solution. The three users who were determined in the second grouping all had mail communicating with each other, and two other users who were in contact with them were added to the group. These two users are non-identifiable users who use external mail that cannot be identified internally by the company, and risk points are calculated judging them as dangerous persons.

With three times of groupings, 'user2' was the baseline and this study was able to find a group of users including two internal employees and two outsiders. Table 4 shows the result of each grouping step. The security risk is calculated based on this grouping, and if the point is above the threshold, it is detected as an internal data leakage group.

The security risk of the group is calculated by using each user-specific scenarios and security risks of the created group, the security weights of the scenarios, and the security level of the group. According to the virtual scenario, for example, the sum of the security risk point of the group in the DRM solution is 149, the occurrence frequency is more than 50, and the security weight is 1, therefore, the security risk point in the data processing step is 149*1, which is 149 points. After calculating the sum of the security risks generated in each step, the average value of security risk is given to outsiders when there is an outsider using external mail. Multiplying the security risk value obtained above by the security level of the group is a security risk of the group. Considering 3 internal dangerous persons and 2 unidentified outsiders as dangerous persons, the group's risk level is 5/5, meaning 1. As a result of the experiment, the security risk of the group was 735, which is higher than the standard score of 500, so it was detected as internal data leakage.

Table 5: Risk Score and Detection of group by Baseline User

Group by baseline user	Security risk	Detection
Group 'user2'	735	o
Group 'user3'	115	x
Group 'user8'	735	o
Group 'user18'	735	o
Group 'user31'	189	x
Group 'user32'	144.	x
Group 'user35'	0	x

Table 5 shows the results of detecting the internal data leakage group including the group of 'user2' as the baseline user. The group in which 'user3', 'user31', 'user32', and 'user35' are baseline users was not determined to be a leakage. In particular, 'user35' is a retiree and there is no separate log, so the result of user grouping is 'user35'. Only groups in which were 'user2', 'user8', and 'user18' were baseline were judged to be leakages, and these three employees were included in one group.

Finally, this study compared the performance with the existing solutions to see if the proposed method based on the previous scenarios can be effectively detected. To create an environment similar to the existing security solution, this study used three detection objects: a group based on a single user and a department, and a group based on a position. In the remaining case, this study used grouping in the proposed method. For comparison, the baseline user was experimented with 'user2' in all cases.

Table 6 shows the results for each detection target, and the score is 49 points for 'user2 individual', 187 points for 'user2's department group', 162 points for 'user2's position group', and 735 points when following 'proposed method'. It indicates that the detected object in the corresponding scenario has only one result group according to the proposed grouping method. In this way, unlike existing security solutions, it can be seen that the proposed method can efficiently detect leakage of internal data by cooperation of several persons.

Table 6: Security Risks and Detection by Target for comparison with Existing Security Solution

Target of detection	Security risk	Detection
Individual	49	x
Group by department	187	x
Group by position	162	x
Proposed group	735	o

CONCLUSION

With the development of IT technology, the business environment of the company changed, and as the documents became digitalized, it became easier and more convenient to access data of company and to leak data, so the importance of data management became bigger. Internal data leakage accidents are continuously increasing and the damage of the accident causes not only the economic damage of the enterprise but also the problem of deteriorating the national competitiveness.

Many companies use a variety of security solutions to prevent such data leakage. However, current security solutions detect data leakage based on individual behavior, or only detect groups from one viewpoint such as department or rank. This is because it is only possible to detect individuals or simple groups, and it is difficult to detect cases where a plurality of people try to leak out internal data. Therefore, in many cases, it is not discovered at that time, but is discovered after major damage happening.

Therefore, this study proposes a detection method of data leakage by cooperation of internal users. In order to find users who collaborate, grouping was done by using personal relationship information, file usage history data, and mail history data in a step-by-step manner, and a single and multiple scenarios were defined through log analysis of the security solution and security risks and security weights were set depending on risk of each behavior. The security risk of the group is calculated using the security risk of each user and the security weights of the combined scenarios, and the internal data leakage is determined when the calculated result is above the reference value.

In order to verify whether the proposed method detects the leakage of internal data by cooperation of users, this study experimented according to a virtual scenario. In order to compare with existing security solution, detection was divided into individual, department group, position group, and proposal group, and as a result, only the group according to the proposed grouping method was detected. Through this, this study could found that the proposed method can efficiently detect the leakage of internal data by cooperation of users. For future research this study will include social network service (SNS) data such as Facebook or Twitter to increase the reliability of user network information.

REFERENCES

[1] “2016 Survey on Information Security(Business)”. National Intelligence Service, Korea, 2017

[2] “Police Statistical Yearbook 2015”. Korean National Police Agency, 2016

[3] S. Liu, and R. Kunn, “data Loss Prevention,” Journal of IT Professional, Vol. 12, No. 2, pp. 10-13, 2010

[4] S. R. Subramanya, and B. K. Yi, "Digital rights management". Journal of IEEE Potentials, Vol. 25, No. 2, pp. 31-34, 2006

[5] S. Bhatt, P. K. Manadhata, and L. Zomlot, "The Operational Role of Security Information and Event Management Systems". Journal of IEEE Security & Privacy, Vol. 12, No. 5, pp. 35-41, 2014

[6] H. W. Shin, "Methodology to analyze insider risk for the prevention of corporate data leakage". Korea Information Science Society, Vol. 39, No. 1, pp. 295-297, 2012.

[7] J. S. Park, and I. Y. Lee, "Log Analysis Method of Separate Security Solution using Single Data Leakage Scenario". Journal of KIPS Tr. Comp. and Comm. Sys., Vol. 4, No. 2, pp. 66-72, 2015.