Privacy Level Indicating Data Leakage Prevention System

Jinhyung Kim[1], Choonsik Park[1], Jun Hwang[1] and Hyung-Jong Kim[1]
[1] Department of Computer Science, Seoul Women's University,
Seoul, Republic of Korea
[e-mail: {jinny,csp,hjun,hkim}@swu.ac.kr]
*Corresponding author: Hyung-Jong Kim

Received February 6, 2013; accepted February 19, 2013; published March 29, 2013

Abstract

The purpose of a data leakage prevention system is to protect corporate information assets. The system monitors the packet exchanges between internal systems and the Internet, filters packets according to the data security policy defined by each company, or discretionarily deletes important data included in packets in order to prevent leakage of corporate information. However, the problem arises that the system may monitor employees' personal information, thus allowing their privacy to be violated. Therefore, it is necessary to find not only a solution for detecting leakage of significant information, but also a way to minimize the leakage of internal users' personal information. In this paper, we propose two models for representing the level of personal information disclosure during data leakage detection. One model measures only the disclosure frequencies of keywords that are defined as personal data. These frequencies are used to indicate the privacy violation level. The other model represents the context of privacy violation using a private data matrix. Each row of the matrix represents the disclosure counts for personal data keywords in a given time period, and each column represents the disclosure count of a certain keyword during the entire observation interval. Using the suggested matrix model, we can represent an abstracted context of the privacy violation situation. Experiments on the privacy violation situation to demonstrate the usability of the suggested models are also presented.

Keywords: DLP system, Critical Information Protection, Privacy Protection

The preliminary version of this paper was presented in the APICIST 2012, July 4-6, Jeju, Republic of Korea. This research was supported by a research grant from by the Technology Innovation Program, 10039670(2011), funded by the Ministry of Knowledge Economy(MKE, Korea)

http://dx.doi.org/10.3837/tiis.2013.03.009

1. Introduction

A data leakage prevention system is operated to manage the flow of a company's significant internal information effectively. According to the data security policy defined by each company, the system is applied to their important information assets [1]. Government organizations and companies operate the system to prevent the important data kept in their intranet from leaking outside, and thus to protect their assets. The data leakage prevention system monitors the packets that are sent from an intranet to the Internet according to a data security policy, and sometimes blocks outgoing packets or discretionarily deletes important information contained in the packets [2]. However, during packet monitoring, it may happen that information that includes private data is seen [3][4].

This paper describes a method for representing the privacy violation status using log data. In particular, we propose two models for indicating the privacy violation status. One model shows the frequency of the monitored private keywords and the other uses matrix representation to show the context of private data disclosures. The output values of the model indicate the degree to which privacy has been violated by the DLP system; these values can be used to realize the privacy protection status of companies and organizations. In addition, the values can be used to avoid privacy violation during the detection of critical information leakage. This paper also describes the design and implementation of the suggested models and presents experimental results to show the operation of the models while the DLP system is detecting information leakage[5]. We used well sampled data to represent the characteristics of the suggested models.

The remainder of this paper is organized as follows. In Section 2, data leakage detection technology and the trade-off relationship between leakage detection and privacy violation are described. In Section 3, the models of the privacy violation level representation, which uses private keywords frequencies and a matrix of time and the keywords, are presented. In Section 4, the two models proposed in this work are examined on the basis of scenarios that are applicable to the models, and the examination results are compared and analyzed. In Section 5, conclusions are drawn.

2. Concerning the Privacy Violation Relation in DLP Process

When Data Leakage Prevention(DLP) systems are operated in organizations, privacy violations can occur during the monitoring process. A large number of DLP keywords must be reviewed, and private data is inevitably included in the keywords. For this reason, we consider that privacy violation can be an issue during the operation of a DLP system.

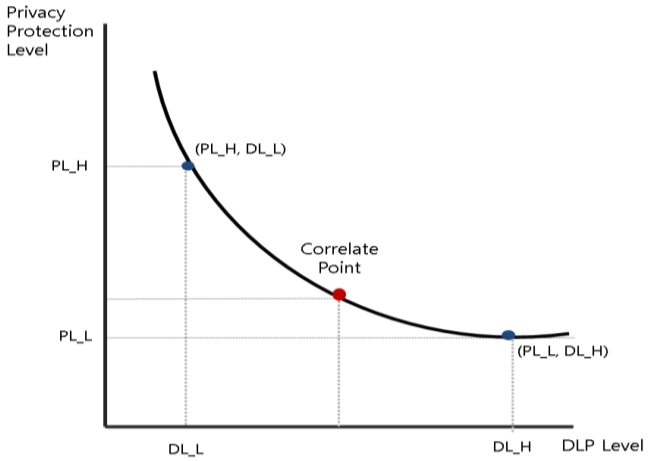


Fig. 1. Correlations between DLP and privacy protection level [2]

Fig. 1. shows the relationship between the data leakage protection level and privacy violation level. The authors' previous study discussed the trade-off between the two indexes [2]. In the private keywords portion of DLP keywords, if a part of the private keywords is excluded from the DLP keywords to protect the employees' privacy, the data leakage detection rate decreases.

3. Design of Privacy Violation Level Model Considering DLP System

3.1 Keyword-based Privacy Exposure Level Measurement Model

Our first model only counts the number of private keyword occurrences in the target information being monitored. Although it may appear to be a trivial solution to represent the privacy violation level, the private keyword occurrence frequency is very useful information for defining a more complex model that fulfills the same purpose.

3.1.1 Static Privacy Violation Level (SPVL)

To estimate the level of privacy, we need to measure the degree of privacy violation. In this study, we defined two important measures that can be used to control the privacy violation level of a DLP system. The first is$\mathbf{PVL_{static}}$, which represents the current privacy violation level, calculated using only the number of keywords:

$$PVL_{static} = \frac{n(Keyword_{private})}{n(Keyword_{dlp})} \tag{1}$$

where,

$$Keyword_{dlp} = \{\mu | \mu \in DLP\ System's\ Keywords\}$$
$$Keyword_{private} = \{p | p \in Private\ keyword\ of\ DLP\ System's\ Keywords\}$$

The PVL_{static} is represented only as a portion of the private keywords in all the DLP keywords. PVL_{static} represents the attitude of the DLP system to private data. If the PVL_{static} value is 0.5, half the keywords are private. The value can also imply the system's perspective on handling the privacy.

3.1.2 Dynamic Privacy Protection Level (DPVL)

It is easily understood that the DLP system cannot meet the private data if no private keyword is included in the target information of the inspection. For this reason, we need a second factor, called $PVL_{Dynamic}$, which indicates the frequency of private keyword occurrences in the DLP system.

The level of $PVL_{Dynamic}$ is determined by a function of time, and it represents the number of private keywords monitored by the DLP system as shown in (2). The expression implies that when the DLP system's monitoring target currently contains more private data than at a previous time t, the value of $PVL_{Dynamic}$ increases. In other words, the $PVL_{Dynamic}$ shows the current privacy violation level of the DLP system. Using this information, the administrator can adjust the $PVL_{Dynamic}$, e.g., by decreasing it. If the administrator removes a certain private keyword that is frequently monitored but is not very critical from the DLP viewpoint, the $PVL_{Dynamic}$ value can be significantly decreased.

$$PVL_{Dynamic} = \frac{\int_{t=t_0}^{t1} KeywordNum_{private}(t)dt}{n(Keyword_{dlp})} \tag{2}$$

where,

$$Keyword_{dlp} = \{\mu | \mu \in DLP\ System's\ Keywords\}$$
$$KeywordNum_{private}(t) = Numberof\ private\ keywords\ detected\ by\ DLP\ System\ into\ given\ time\ t$$

3.2 Context-based Privacy Exposure Level Measurement Model

3.2.1 Pattern Definition

A model for detecting more than two kinds of personal information and a certain pattern of information is proposed. **Figs. 2** and **3** present the concept of patterns of personal information.

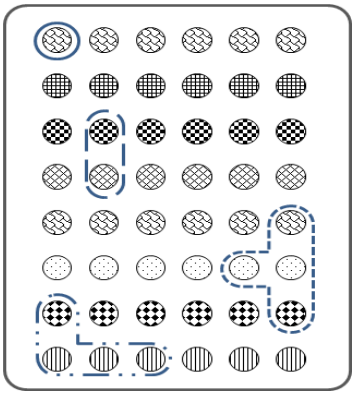


Fig. 2. Patterns in monitoring packet

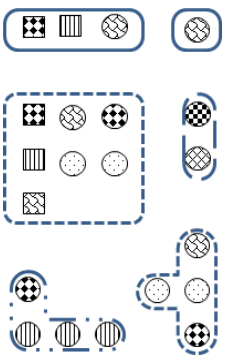


Fig. 3. Defined patterns of personal information

The concepts displayed in **Figs. 2** and **3** are applied to the personal information area to define personal information patterns, as shown in **Tables 1** and **2**.

Personal information patterns are defined on the basis of the importance of the personal information and its exposure frequency in time. Using the defined patterns, it is possible to detect the exposure level of the personal information that is included in the packets to be monitored in the data leakage detection process by time unit, data type, and data pattern. Personal information patterns are divided into the representative types as follows[6]:

o Pattern 1: Representativeness + Distinctive data (1 or more than 2 data)
Resident registration number + Financial information
(Bank account number, Credit card number)
o Pattern 2: Specific time zone + Specific data (1 data)
(Within work hours + Resident registration number)
o Pattern 3: Specific time zone + Bundle of specific data (more than 2 data)
(Within work hours + Resident registration number + Credit card number)

The personal information patterns used in this study were based on the 12 data defined in the personal information pattern-based detection mechanism, and reflected the characteristics of the personal information.

The personal information in **Tables 1** and **2** is based on the meaningful data among the data that serve the function of identifying an individual, and is defined as the data that are expected to violate one's privacy critically when more than two data are detected simultaneously.

Table 1. Dual-P with two personal information data

	Data 1	**Data 2**
Dual-P1	Name	Resident registration no.
Dual-P2	Name	Mobile phone no.
Dual-P3	Name	E-mail addr.
Dual-P4	Company ID	E-mail addr.
Dual-P5	Company ID	Resident registration no.
Dual-P6	Company ID	Mobile phone no.

Table 2. Triple-P with three personal information data

	Data 1	**Data 2**	**Data 3**
Triple-P1	Name	Credit card no.	Company ID
Triple-P2	Name	Bank account no.	Company ID
Triple-P3	Name	Resident registration no.	Credit card no.
Triple-P4	Company ID	Credit card no.	Telephone no.

Triple-P5	Company ID	Bank account no.	Telephone no.

3.2.2 Pattern-based Privacy Detection Model

The pattern-based personal information detection mechanism is defined as a method for understanding and analyzing the personal information patterns included in packets, and accurately detecting personal information using not only the distributional map of one of the data, but also the assignment and detection frequency of more than two data.

In this study, we used Kronecker's delta in the process of defining a matrix to operate using the insertion of 0 and 1 in the course of drawing a matrix model to fit each condition. In Kronecker's delta, the two variables i and j with a positive number in linear algebra are defined as follows:

$$\delta_{ij} = \begin{bmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1j} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ \delta_{i1} & \delta_{i2} & \cdots & \delta_{ij} \end{bmatrix} \text{ (Where, } \delta_{ij} = \{0,1\} \text{)} \tag{3}$$

The Kronecker's delta is used to arrange data by type, data pattern, and time. Thus, it is possible to find each mapping value of personal information patterns and operate them. The formula to calculate PIEt, which refers to the exposure level of personal information in time unit t, is presented as follows. The exposure level of personal information during time T or in a specific time unit is calculated depending on the following conditions.

o Detection level by data type: $PIED_{Type}$
In the case of j=n, use the matrix model δij with the insertion of 1

$$PIED_{Type} = \begin{bmatrix} PIE_{A_1,t_1} & PIE_{A_2,t_1} & \cdots & PIE_{A_{12},t_1} \\ PIE_{A_1,t_2} & PIE_{A_2,t_2} & \cdots & PIE_{A_{12},t_2} \\ \vdots & \vdots & \vdots & \vdots \\ PIE_{A_1,t_n} & PIE_{A_2,t_n} & \cdots & PIE_{A_{12},t_{1n}} \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} = \text{PIED}_{A_1} \tag{4}$$

(However, PIE_t is the exposure level of personal information in time t)

o Detection level by data type: $PIED_{Time}$
In the case of $i=n$, use δ_{ij} with the insertion of 1

$$PIED_{Time} = \begin{bmatrix} PIE_{A_1,t_1} & PIE_{A_2,t_1} & \cdots & PIE_{A_{12},t_1} \\ PIE_{A_1,t_2} & PIE_{A_2,t_2} & \cdots & PIE_{A_{12},t_2} \\ \vdots & \vdots & \vdots & \vdots \\ PIE_{A_1,t_n} & PIE_{A_2,t_n} & \cdots & PIE_{A_{12},t_{1n}} \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix}$$
$$= \text{PIED}_{t_1} \quad (5)$$

(However, PIE_t is the exposure level of personal information in time t)

4. Implementation and Experiments

4.1 Scenario Definition

4.1.1 Scenario of the Keyword-based Model

In the keyword-based mechanism that measures the personal information exposure level, when μ , which is the number of personal information data in packets, is 100, and the defined p , which is the number of checked privacy data, is 30, the $\text{PVL}_{\text{Static}}$ value becomes 0.3, which means that 30% of privacy data are checked in the monitoring process. In addition, $\text{PVL}_{\text{Dynamic}}$ requires information on the time unit; when $\text{n}(\text{KeywordNum}_{\text{private}})$, the mapped value in the time unit, is 10, $\text{PVL}_{\text{Dynamic}}$ becomes 0.1.

Table 3. Number of defined data

$n(Keyword_{dlp})$	100
$n(Keyword_{private})$	12
$n(KeywordNum_{private})$	10

Table 4. Defined $\boldsymbol{n(p)}$

Privacy Keyword in Detect Keyword		
Social number	Phone	Email
Name	Card number	Account number
Address	id	Password
Car number	Driver's license number	Passport number

When the number of personal information data among the total 100 data checked in a company is 12, as shown in **Table 3**,

o SPVL is 12/100 = 0.12.

$$PVL_{static} = \frac{n(Keyword_{private})}{n(Keyword_{dlp})} = \frac{12}{100} = 0.12 \quad (6)$$

PVL_{Static} is calculated using the mapping value of keywords and defined data within a time unit. DPVL refers to the mapped value using the monitored packets among $Keyword_{private}$, and through the entire packet mapping process, $n(KeywordNum_{private})$, the number of actually mapped data, is calculated. For example, when the number of defined $n(Keyword_{dlp})$is 100, the number of $n(Keyword_{private})$ is 30, but when the number of $n(KeywordNum_{private})$ matching the actual packets is 10, the actual DPVL becomes 0.1.

$$PVL_{dynamic} = \frac{\int_{t=t_0}^{t_1} KeywordNum_{\text{Private}}(t)dt}{n(Keyword_{\text{dlp}})} = \frac{10}{100} = 0.1 \quad (7)$$

4.1.2 Scenario of the Pattern-Based Model

In this section, the scenario of the pattern-based mechanism proposed in Section 5, which measures the privacy exposure level, is applied and examined.

First, the packet data used for a day in a company are analyzed, and then the mapped value is calculated using the defined personal information data, and thereby the risk level of personal information leakage is calculated. When a personal information administrator checks whether there is a strong possibility of personal information being exposed in the process of controlling the information flow in a company, when personal information is frequently checked in the data flow, the quantity of personal information exposed to the personal information administrator increases, and therefore the risk of personal information leakage becomes high. The data detection pattern is presented as shown in the following. The time unit t is set to one hour, and 12 personal information data detected per hour are shown. The accumulated value from t0 to t22 is displayed.

- Horizontal axis: 12 personal information data to be protected;
- Vertical axis: 24 hours by one-hour unit from T0 to T22 (1 day);
- Value: each accumulated value of 12 data included in the packets analyzed every one hour

On the basis of the defined personal information patterns, the results of data detection every hour are presented. Based on the detected value, each item of personal information data value is counted. Considering the simultaneous detection based on the defined patterns, detection results by pattern are drawn. Dual-Pn represents the result of detection using a two-data-based pattern, and Triple-Pn represents the result of detection using a three-data-based pattern. Three scenarios to measure the exposure level of personal information by time, data type, and data pattern are defined as shown in Table 5. Based on the scenarios, the exposure level of privacy was measured[4].

Table 5. Scenarios to measure the pattern-based personal information exposure level

Type	Scenario
Case 1	Measure the exposure level of the highly important specific personal information data included in the monitored packets
Case 2	Measure the exposure level of personal information defined in a specific time zone
Case 3	Measure the exposure level of specific information (combination of two or three data) in a specific time zone

4.2 Implementation and analysis of examination

4.2.1 Implementation

To examine the mechanism proposed in this work, we designed a data leakage detection system with a function that measures the exposure level of personal information, as shown in **Fig. 4**.

In the level-based detection mechanism, it is possible to measure each level by applying the leakage value calculated by the conventional data leakage detection system, and calculate each privacy exposure level by adding the indexing function into the conventional system.

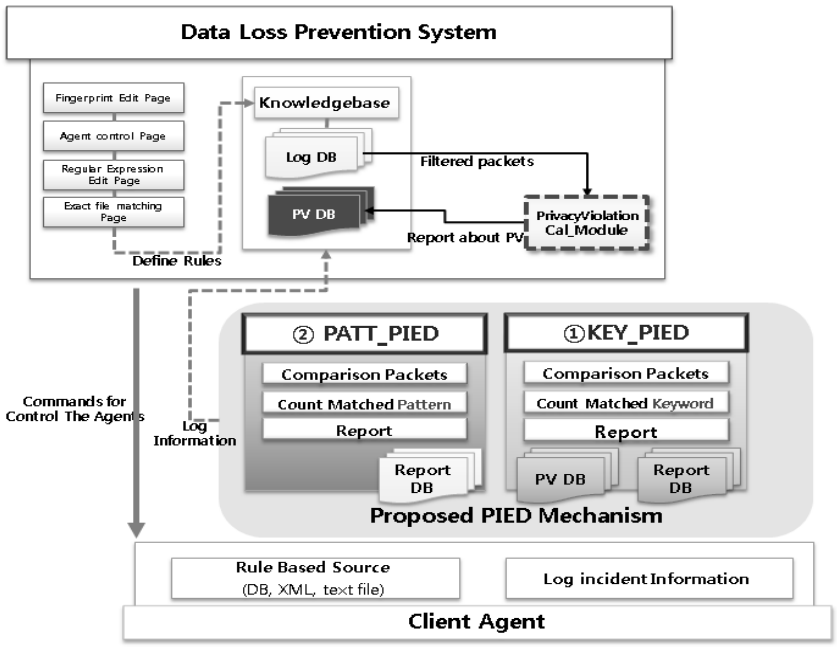


Fig. 4. Architecture of privacy exposure level measurement system

In the keyword-based detection mechanism, on the basis of the keywords defined by the KEY_PIED module, the keyword data included in packets are found and are indexed to measure the privacy exposure level. In the pattern-based detection mechanism, on the basis of the pattern defined by the PATT_PIED module, the data simultaneously detected in packets are found and indexed to measure the privacy exposure level.

4.2.2 Results

In this section, the screen images of the system that was implemented using these research results are presented and described.

Fig. 5 shows the main page of the proposed system. The page consists of Notice, Alert, Data Search, and a Detected Keyword in System graph. Notice provides general information for the administrator and Alert contains the important detection results of the DLP system. Data Search enables the administrator to find information using the keyword and period of time. The graph provides intuitive recognition of the DLP system's detection result.

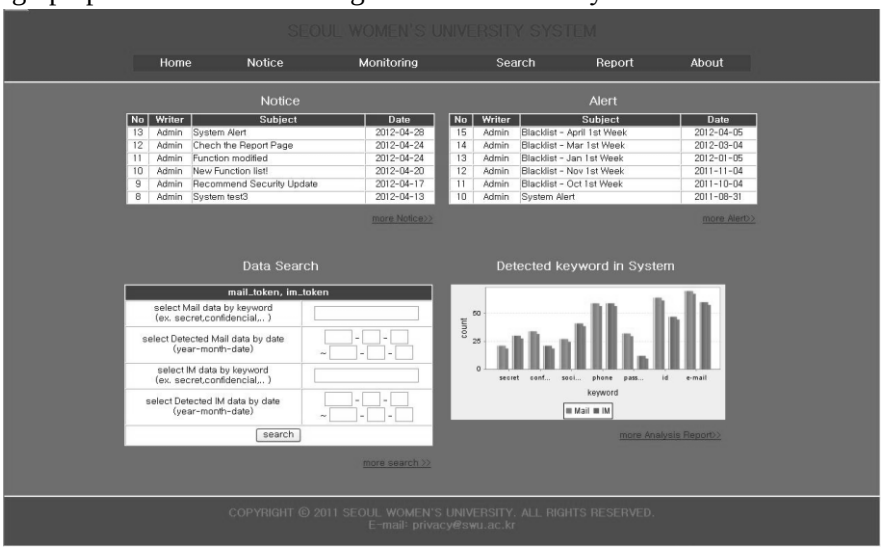


Fig. 5. Main page

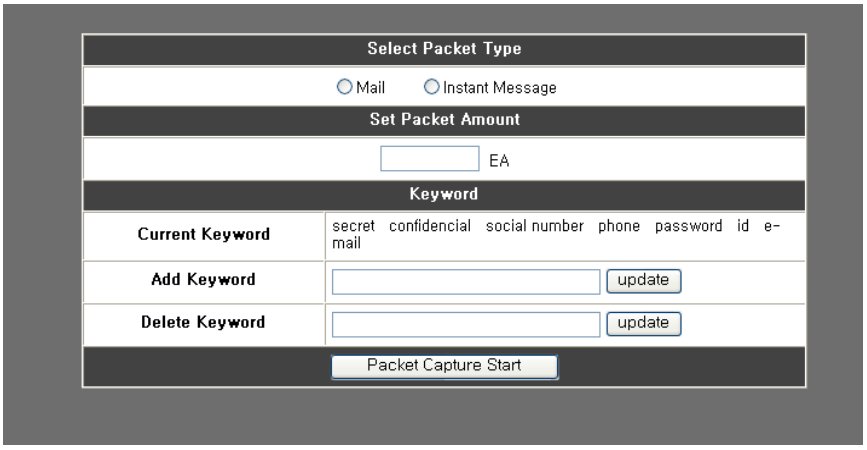
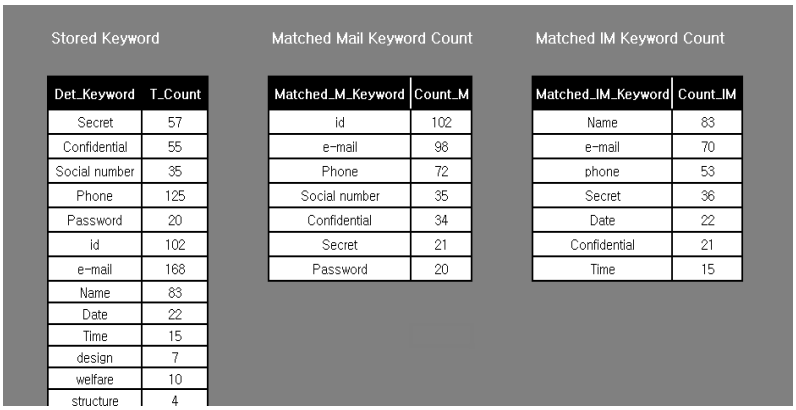


Fig. 6. Keyword management page

Using the page shown in **Fig. 6**, administrators can add or delete the keyword of e-mail and instant messenger

Fig. 7 shows the number of detections for each keyword. These data are used to calculate the PVLDynamic value. In our DLP system, there are 100 keywords for critical data protection and 12 private keywords. Therefore, the current PVLstatic value is 0.12. Since the PVLDynamic value varies over time, the database in which the detection result is stored should include a time field. Even though the database table has the time field, **Fig. 7** shows only the keyword and the detection count, because the purpose of the page is to manage the PVLDynamic value. In this case, if the private keyword "ID" of e-mail and the "name" of messenger are removed, the PVLDynamic value will decrease sharply[7].


Stored Keyword

Det_Keyword	T_Count
Secret	57
Confidential	55
Social number	35
Phone	125
Password	20
id	102
e-mail	168
Name	83
Date	22
Time	15
design	7
welfare	10
structure	4

Matched Mail Keyword Count

Matched_M_Keyword	Count_M
id	102
e-mail	98
Phone	72
Social number	35
Confidential	34
Secret	21
Password	20

Matched IM Keyword Count

Matched_IM_Keyword	Count_IM
Name	83
e-mail	70
phone	53
Secret	36
Date	22
Confidential	21
Time	15

Fig. 7. Result of the keyword based detection

When the patterns defined in this study were included in the Mail and IM packets that were collected and analyzed to prevent the leakage of important information, the system that calculates the exposure frequency of privacy information was implemented. **Table 6** presents the pseudo code of the algorithm for measuring the pattern-based privacy exposure, which is aimed at measuring the appearance frequency of each of the defined patterns. The frequency is measured in Email packets and IM packets using the mapped data with the defined patterns. Based on each pattern of two and three personal information data, the exposure level of the detected packets was measured and displayed.

Table 6. Pseudo Code of Personal Information Pattern Based Detection System

Matched_Triple_Pattern_Count() *{* *i=get_Pattern();* *switch(i) {* *case Triple-P1:* *Count1++;* *break;* *case Triple-P2:* *Count2++;* *break;* *case Triple-P3:* *Count3++;* *break;* *case Triple-P4:* *Count4++;* *break;* *case Triple-P5:* *Count5++;* *break;* *}* *print_num_of_count(**Count**$_k$);* *}*	*Matched_Dual_Pattern_Count()* *{* *i=get_Pattern();* *switch(i) {* *case Dual-P1:* *Count1++;* *break;* *case Dual-P2:* *Count2++;* *break;* *case Dual-P3:* *Count3++;* *break;* *case Dual-P4:* *Count4++;* *break;* *case Dual-P5:* *Count5++;* *break;* *case Dual-P6:* *Count5++;* *break;* *}* *print_num_of_count(**Count**$_k$);* *}*

The results of the pattern-based measurement of the personal information exposure level proposed in this work are indexed and presented as graphs as shown in **Figs. 8** and **9**. Each detection value of the pattern of two personal information data and the pattern of three personal information data was measured at a specific point; the results are presented as graphs.

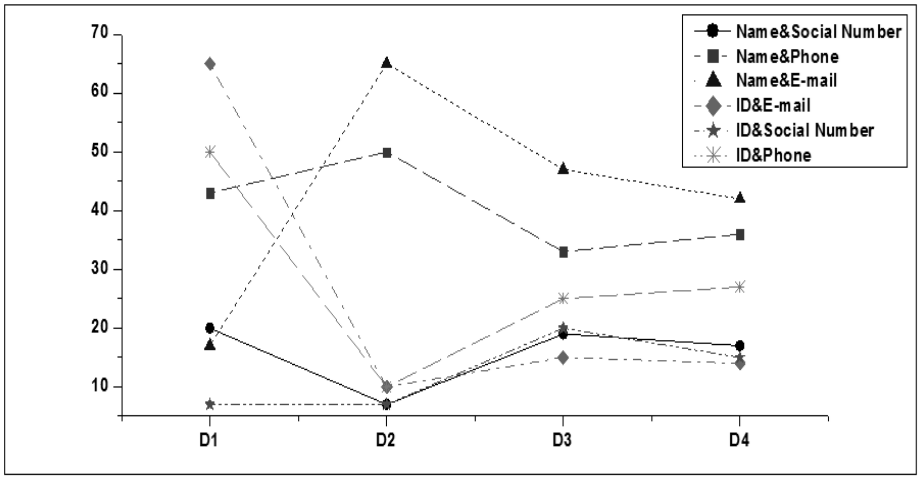


Fig. 8. Dual-P pattern matching result at a specific measurement point

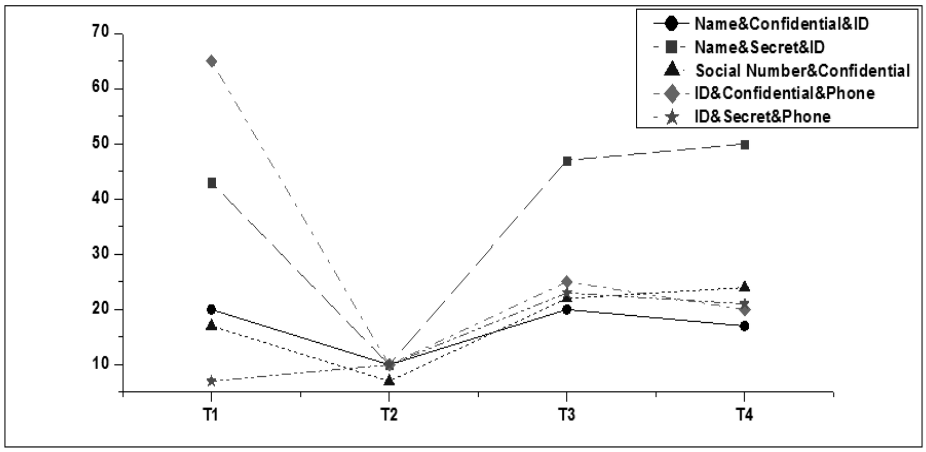


Fig. 9. Triple-P pattern matching result at a specific measurement point

Fig. 10 and **11** present the graphs showing the amount of the defined patterns detected in the packets that were monitored for a day. Each detection value of the dual and triple pattern had been measured between 0 and 22 hours, and is presented in the graphs.

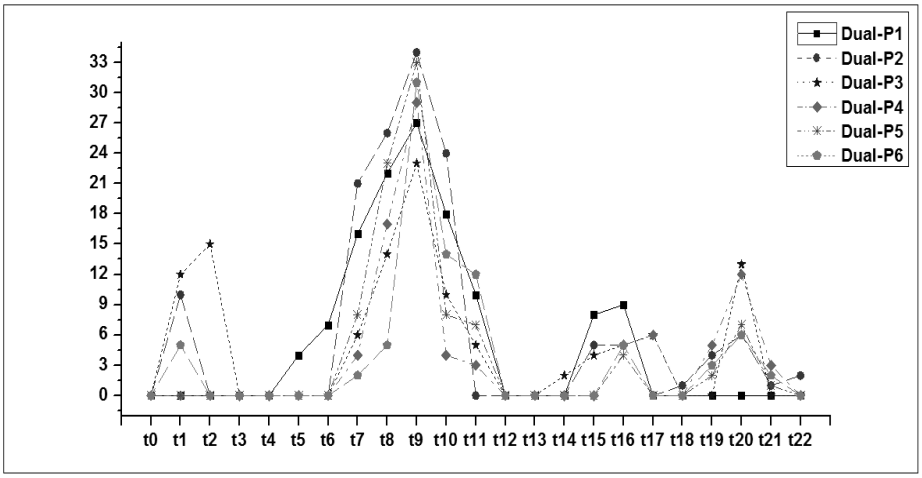


Fig. 10. One-day Dual-P pattern matching result

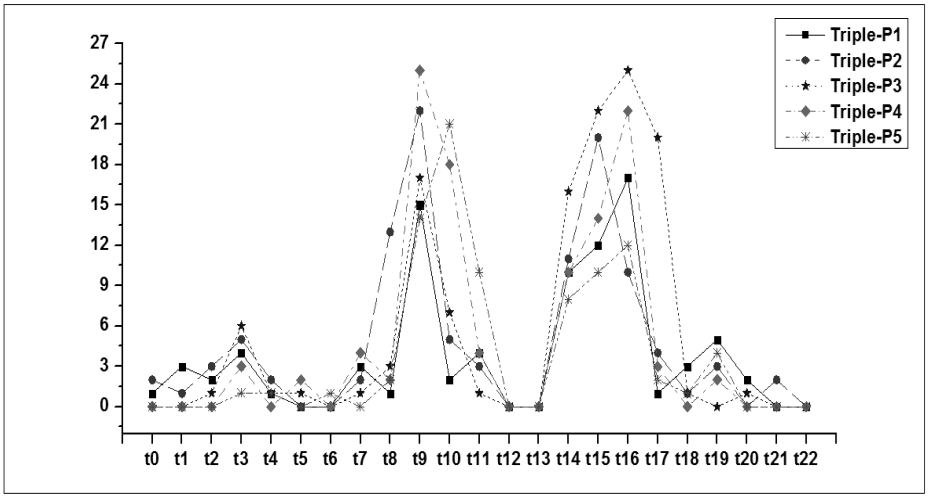


Fig. 11. One-day Triple-P pattern matching result

4.2.3 Comparison Analysis of Scenario Based Examination Results

Given that the result of the static calculation method of formula (1) is compared with that of formula (2) in the keyword-based mechanism for indexing personal information exposure, the privacy violation is checked on the basis of $n(Keyword_{dlp})$, which is the number of the keywords defined in formula (1); the result is different from the result of formula (2), which uses the keywords detected in actual packets. Based on this, the following problem is identified: there is a case where an unnecessarily large amount of p included in the keyword μ to be detected in the monitoring process is defined. In fact, it is difficult to check them through one-time monitoring. However, if the undetected p value is defined as μ through the comparison of detection results during a time unit, this means that the relevant data are monitored, although they do not constitute the company's important internal information. Therefore, by excluding p from μ, it is possible to protect employees' privacy.

In the pattern-based detection mechanism described in Section 3.2, the scenarios are based on actual cases that could occur in a company. Therefore, patterns were defined on this basis. The mechanism was based on the patterns of the detectable data by time unit, and the patterns detected for a certain period of time, or the detection time zones by data were identified. In the first scenario, in the process of finding whether important information is included, the mechanism checks the exposure level of specific information during a certain period of time to investigate the personal information leakage risk level for the specific information. According to the results of the measurement of leakage risk, through each exposure level, of resident registration number, bank account number, and credit card number, which are described in Case 1, in packets, in Triple-P3, the measurement value at 9 o'clock in the morning was 17, and at 4 o'clock in the afternoon, 25. Therefore, based on the daily average use value of 5.35, a relatively large amount of data was detected, showing that resident registration numbers, bank account numbers, and credit card numbers were frequently used. The fact that such important information is used frequently in the time zone implies that the

risk of exposure and leakage of the relatively important information is high. In particular, given that the combination of resident registration number, bank account number, and credit card number is generally used for payment, the leakage of all this information could result in secondary financial damage. Therefore, based on the use pattern, it is possible to prepare a policy to prevent data leakage.

If the mapping patterns for specific data in a specific time zone, which are described in Case 2, are checked, it is possible to find the occurrences of personal information leakage through hacking and other attacks, and respond to them. If personal information is used frequently during a specific time, an administrator checks a large amount of personal information data in the monitoring process during that time, and therefore such information is exposed to the administrator.

If patterns are detected on the basis of a specific time zone and specific data, which are described in Case 3, in **Figs. 8** and **9**, it is possible to identify the use state of specific data within work hours, the use state of personal information outside work hours, and the use state of specific data in a specific time zone. In the case of the data use pattern within work hours, it is predicted that most personal information data are used immediately after the employees' arrival at their office, and that these data are used frequently because of their log-on, corporate authentication, and other job-related actions. In addition, it is predicted that the reason for personal information use increasing slightly from 3 to 5 o'clock in the afternoon is that employees process banking and credit card operations until 4 o'clock, when the banks close to the public

The conclusions of this study, based on the results of the proposed keyword-based mechanism for detecting personal information and the pattern-based mechanism for detecting personal information, as shown in **Table 7**.

Table 7. Comparison analysis of the results of the proposed mechanisms

		KEY_PIED	PATT_PIED
Detection Method		The personal information that should be protected in the corporate monitoring process aimed at data leakage detection is defined as keywords; the keyword value measured in the monitoring process is calculated as a personal information data value.	The keywords to be protected by the keyword-based detection mechanism are used. Each combination of two and three data is defined as a pattern. The pattern value detected in the monitoring process is calculated as personal information pattern value.
Proposed mechanisms	Advantages	The defined keywords can be used as the foundation for establishing a relevant detection policy.	Based on the defined patterns, it is possible to define a new pattern, and apply it to various cases.
	Disadvantages	All of the personal information cannot be defined as keywords, and the exposure level of undefined keywords is not be calculated.	It is difficult to calculate undefined patterns.

When the mapped patterns are identified on the basis of a specific data pattern at a specific time, depending on its result values, it is possible to use them as the foundation for establishing a policy to prevent the leakage of personal information, and in particular, for establishing a policy concerning the issues to which an administrator of a data leakage detection system should pay attention in terms of personal information exposure in the monitoring process. In addition, it is easy to conceive of a foundation for investigating an administrator's improper actions.

5. Conclusion

In this paper, we have described a study on a data leakage detection model that reflects employees' privacy protection in the process of data leakage detection, which is conducted by companies to prevent leakage of their corporate information. A model for measuring the level of the inevitable privacy breach in the data leakage detection process was proposed, and a system based on the proposed model was designed and implemented. Further, the system was evaluated through scenarios. In the process of monitoring packets flowing from an intranet to the Internet using a conventional data leakage detection system, the exposure of internal users' personal information that is included in the packets to be monitored can lead to the violation of the internal users' privacy. Recognizing this problem, we investigated an approach for

minimizing the exposure of internal users' personal information. As a result, we proposed the personal information keyword-based and pattern-based privacy exposure level measurement model as an approach for data leakage detection that reflects the protection of internal users' privacy.

The proposed mechanism can measure the exposure level of personal information in the data leakage detection process. The data leakage detection system in which the mechanism is implemented can measure how much personal information is included in its monitored packets. Based on the measurement, it is possible to establish a privacy policy in companies. In addition, the measured results can be used as a fundamental basis for privacy protection. In addition, companies that develop data leakage detection solutions can develop and sell the solutions that reflect the personal information exposure level measurement model, and by so doing, can contribute to realizing a data leakage detection system that reflects internal users' privacy protection and achieves safe and efficient Internet environments. In addition, it is expected that the keyword-based and pattern-based detection mechanism proposed in this work will be used to allow an administrator to monitor packets taking internal users' privacy into consideration in the process of detecting data leakage.

References

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "On abnormality detection in spuriously populated data streams," *ACM Computing Surveys* (CSUR), vol. 41, issue 3, July 2009. Article (CrossRef Link)

[2] Jinhyung Kim and Hyung-Jong Kim, "Design and implementation of data leakage prevention system considering the level of privacy protection and violation," *Information - An International Interdisciplinary Journal*, vol. 14, no. 5, November, 2011. Article (CrossRef Link)

[3] Salvatore J. Stolfo, Shlomo Hershkop, Chia-Wei Hu, Wei-Jen Li, Olivier Nimeskern, and Ke Wang, "Behavior-based modeling and its application to email analysis," *ACM Transactions on Internet Technology*, vol. 6, no. 2, pp. 187–221, May 2006. Article (CrossRef Link)

[4] Balachander Krishnamurthy, Delfina Malandrino, and Craig E. Wills, "Measuring privacy loss and the impact of privacy protection in Web browsing," *Proceedings of SOUPS (Symposium On Usable Privacy and Security)*, July, 2007. Article (CrossRef Link)

[5] Sakaki Hiroshi, Yanoo Kazuo, Ogawa Ryuichi and Hosomi Itaru, "An information leakage risk evaluation method based on security configuration validation," *IEICE Technical Report*, vol. 105, no. 398, pp.15–22, 2005. Article (CrossRef Link)

[6] K. Das and J. Schneider, "Detecting anomalous records in categorical datasets," KDD, 2007. Article (CrossRef Link)

[7] Daeseon Choi, Seunghun Jin, and Hyunsoo Yoon, *A Personal Information Leakage Prevention Method on the Internet*, 3rd edition, Springer-Verlag, Berlin Heidelberg, New York, 1996. Article (CrossRef Link)

Jinhyung Kim received the B.S degree in the Information Security, Seoul Women's University, Seoul, Korea in 2006. She received the M.S degree in 2008 and Ph.D. degree in 2013 in the computer science at the same university. Her researches interests include protect techniques and policies in Privacy protection and Cloud Computing Security.

Choonsik Park received the B.S degrees in department of Wireless Communication Engineering from Kwangwoon University, Korea, in 1981 and M.S degrees in Electrical Engineering department from Hanyang University in 1983, Korea, and Ph.D. degrees in Electrical Engineering department from Tokyo Institute of Technology, Japan in 1995. Since 2009, he has been an assistant professor at the Department of Information Security in Seoul Women's University. His current interests are Cryptography, Information Security, and Privacy Protection.

Jun Hwang received the B.S and M.S. and Ph.D. degrees in Computer Science from Chung-Ang University, Korea, in 1985, 1987, and 1991 respectively. Since 1992, he has been a professor at the Department of Multimedia in Seoul Women's University. His current interests are IPTV, convergence computing, and digital broadcasting.

Hyung-Jong Kim received his B.S. degree in Information Engineering from the Sungkyunkwan University in 1996 and his M.S and Ph.D. degrees in Electrical Computer Engineering department of Sungkyunkwan University. He worked as a principal researcher of Korea Information Security Agency (KISA) from 2001 to 2007. Also, he worked in the CyLab at CMU(Carnegie Mellon University), Pittsburgh, PA, USA as a visiting scholar from 2004 to 2006. Currently, he is with the Seoul Women's University in Seoul, Korea as an associate professor in department of Information Security since March, 2007. His research interests include cloud computing security, VoIP security, privacy protection and simulation modeling methodology.