Privacy Level Indicating Data Leakage Prevention System

Jinhyung Kim, Jun Hwang and Hyung-Jong Kim*

Department of Computer Science, Seoul Women's University
{jinny, hjun, hkim}@swu.ac.kr*

Abstract

As private information can be contained in the DLP (Data Leakage Prevention) system's target of monitoring, the monitoring process inevitably violates privacy of the internal employees. Currently, existing DLP systems do not consider the privacy violation during the monitoring process. In this work, we are proposing a DLP system considering privacy violation level. The privacy violation level of our system has static and dynamic characteristics. The static privacy level just indicates the monitoring target's portion of private data. The dynamic privacy level indicates the portion of private data which are disclosed by DLP system. The privacy level of our proposing DLP system can be used to control the private violation by removing specific monitoring targets in DLP system. The contribution of this work is defining the privacy level in DLP system and implementing the proposed idea.

Keywords: *DLP system, Critical Information Protection, Privacy Protection*

1. Introduction

For the protection of corporate assets, the Data Leakage Prevention (DLP) system is used by many corporations. DLP system is the monitoring system that uses a network packet to monitoring the system's information. To protect the company's information, corporations use systems that are developed to prevent the loss or leakage of information that checking the packet included sensitive data or related assets [4]. However, on the process of using these systems, administrators may monitor employees' private data. Generally, the private information of employees of an organization should be protected and laws and regulations require this as a basic right of people [3]. However, the DLP system may monitor some part of private information and it could cause a privacy violation. So, in this respect, we need the DLP system considering privacy protection. In this paper, we proposed a privacy violation level indicating DLP system.

The remainder of the paper is organized as follows. Section 2 explains the concepts of the trade-off relation of DLP and privacy protection. Section 3 describes the estimation model for the degree of privacy violation on the monitoring for detecting the critical data. Especially, we considered two cases of privacy violation level. One is static privacy violation level which can be calculated simply using private data portion of monitoring target. The other is dynamic privacy violation level which is calculated using currently monitored private data portion. Especially the dynamic privacy violation level is displayed through graphical user interface for showing the DLP system's privacy violation level. Finally, section 4 concluding remarks and outline of future work are shown.

2. Concerning the Privacy Violation Relation in DLP Process

When Data Leakage Prevention systems are operated in organizations, privacy violations can be occurred during the monitoring process. The large number of DLP keywords must be reviewed and private data is inevitably included in the keywords. That is why we considering the privacy violation can be an issue in DLP system operation.

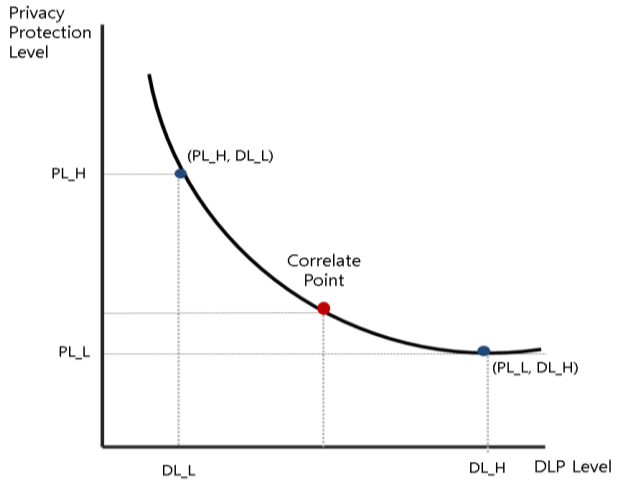

Figure 1. Correlation of Privacy Protection and DLP Level[1]

The figure 1 shows the relationship between data leakage protection level and privacy violation level. Authors' previous paper shows the trade-off between the two indexes [1]. When it comes to private keywords portion of DLP keywords, if a part of private keywords are excluded from DLP keywords to protect employees' privacy, the detecting rate of data leakage is getting lower than before.

3. Design and Implementation of Privacy Violation Level Considering DLP System

3.1. Private Keywords and DLP Keywords

As we have reviewed in section 2, there is a trade-off relationship between the DLP level and privacy level. In the DLP system's design implementation viewpoint, the trade-off relationship can be modeled as number of keywords for each category. In other words, because the DLP keyword set contains private keyword set, if the number of private keywords in DLP keyword set is large, the detection process monitors many private data. That's our basic idea of representing privacy level in DLP system.

3.2. Static and Dynamic Privacy Violation Level

To estimation the level of privacy, we should measure the degree of privacy violation. In this work we defined two important measures which can be used to control the privacy violation level of DLP system. First one is PVL_{Static} and it represents the current privacy violation level which is calculated using just the number of keywords.

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

where,

$Keyword_{dlp} = \{u \mid u \in \text{DLP system's keyword}\}$

$Keyword_{private} = \{p \mid p \in \text{Private keywords of DLP system's keywords}\}$

(1)

As we can see (1), the level represents the portion of private keywords of DLP keywords. Although this level can be derived in a very easy way, it effectively reveals how much the organization's DLP process violates.

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

where,

$Keyword_{dlp} = \{u \mid u \in \text{DLP system's keyword}\}$

$KeywordNum_{\text{Private}}(t)$: Number of private keywords detected by DLP system in given time t

(2)

The dynamic privacy violation level $PVL_{Dynamic}$ is determined by a function of time which represents the number of private keywords detected by DLP system as shown in (2). The expression implies when the DLP system's monitoring target contains more private data than previous time t, the value of $PVL_{Dynamic}$ is getting larger. The $PVL_{Dynamic}$ shows current violation level of the system and during the monitoring process the detected keywords and the number of detection for each keyword are shown through graphical user interfaces. Using the information, administrator can have the insight how the make the $PVL_{Dynamic}$ lower. In other words, if the administrator removes a certain private keyword which is frequently detected but not that critical in DLP viewpoint, the $PVL_{Dynamic}$ value can be highly decreased.

3.3. Implementation

Above mentioned concepts are implemented as a DLP system. Our DLP systems categorize the monitoring target keyword and manage the private keyword set. Whenever the private keyword is monitored, the keyword's monitoring number increases. Even though this process looks very simple, through this simple process administrator can imagine the level of privacy protection of their DLP system.

Figure 2. shows the system architecture of proposed DLP system. The system has a component named PrivacyViolation_Cal Module which is conducting the mentioned calculation. The calculation is proceeded in a given period of time and the result is stored in PV database table.

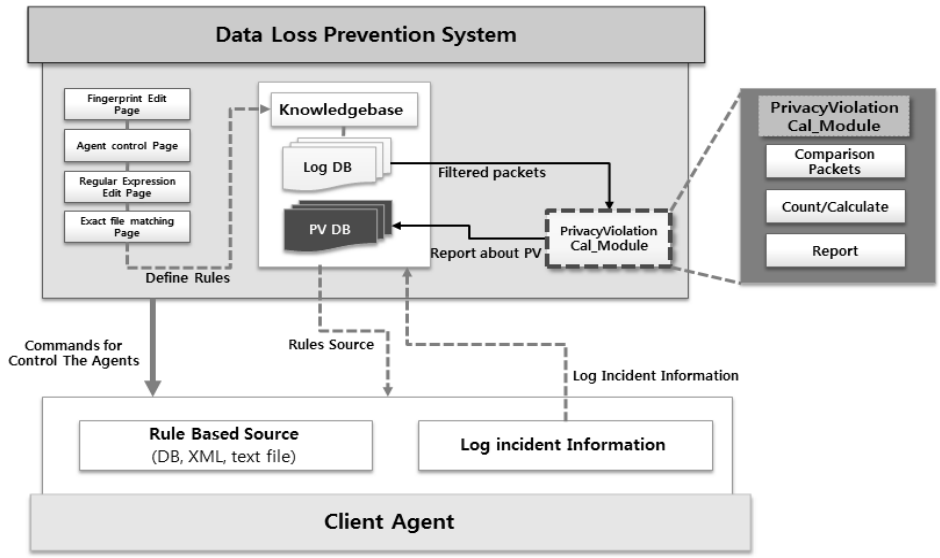

Figure 2. The DLP System Structure considering privacy protection

As shown in below part of Figure 2, our system's monitoring is proceeded by agent software installed in employee's computer. The detection rules which are contains the DLP keyword are downloaded from server and the rules are applied to each packet going out from the employee's computer. Usually, the detection target would be the packets of e-mail and instant messenger. The detection result is reported to the server and the calculation is conducted.

Figure 3 is Main Page of the proposed system. The page consists of Notice, Alert, Data Search and Detected keyword graph. Notice is for general information for administrator and alert contains important detection result of DLP system. The Data Search enables administrator to find some information using keyword and period of time. The detected graph is for intuitive recognition of DLP system's detection result.

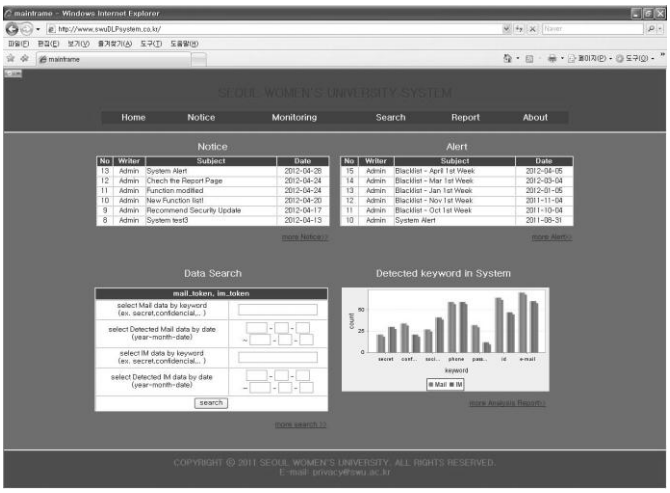

Figure 3 Main Page

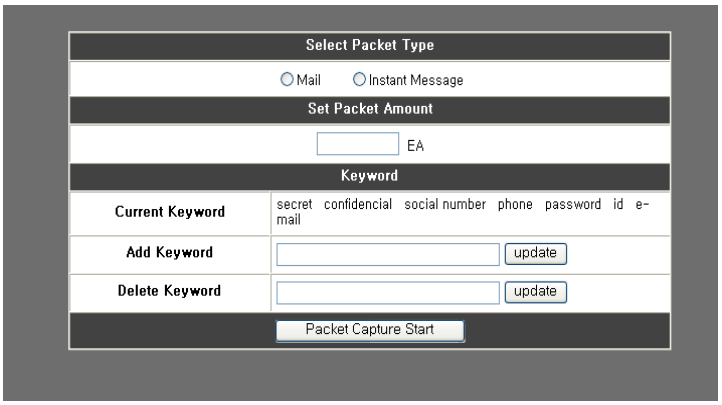

Figure 4. Page for Setting Keyword and Type of Packet

Through the page of Figure 4, administrators can add or delete the keyword of e-mail and instant messenger.

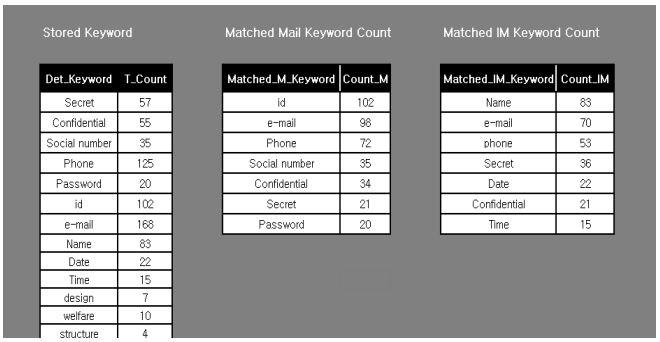

Figure 5 Pages for Showing Frequency of Private Keywords

Figure 5 shows the number of detection for each keyword. Those data are used to calculate the $PVL_{Dynamic}$ value. In our DLP system, there are 100 keywords for critical data protection and there are 13 private keywords. Therefore, the current PVL_{static} is 0.12. Since $PVL_{Dynamic}$ value varies as the time goes on, the database should have the time filed to store the detection result. Even though the database table has the time field, the Figure 5 just shows the keyword and the detection count. That is because the page is for managing the value of $PVL_{Dynamic}$. In this case, if the private keyword "id" of e-mail and the "name" of messenger are removed, the $PVL_{Dynamic}$ value is going to be lower steeply.

4. Conclusion and Future Work

In this paper, we suggested the privacy violation estimating model of DLP system. Our suggestion has very simple concept of differentiating private keywords from DLP keywords. In addition, by counting the number of private keyword, we derived the static and dynamic privacy violation level of DLP system. The proposed concept is implemented using web-based DLP system and we have presented the user interfaces showing the detection number of private keywords. Through our suggesting concept and system, administrators can speculate how much their DLP system violates the privacy during monitoring the leakage. In addition, by removing some part of private keywords, they can increase the privacy violation level.

Acknowledgements

This work was supported by the Industrial Strategic technology development program, 10039670**(2011)**, funded by the Ministry of Knowledge Economy(MKE, Korea)

References

[1] J. H. Kim and H. J. Kim, "The Data Modeling considered Correlation of Information Leakage Detection and Privacy Violation", ACIIDS 2011 : 3rd Asian Conference on Intelligent Information and Database Systems, **(2011)**, LNAI 6592, pp. 165-170.
[2] S. Hiroshi, Y. Kazuo, O. Ryuichi and H. Itaru, "An Information Leakage Risk Evaluation Method Based on Security Configuration Validation", IEICE Technical Report, vol. 105, no. 398, **(2005)**, pp.15-22.
[3] D. Choi, S. Jin and H. Yoon, "A Personal Information Leakage Prevention Method on the Internet", 3rd edn. Springer-Verlag, Berlin Heidelberg New York **(1996)**.
[4] V. Chandola, A. Banerjee and V. Kumar, "On Abnormality Detection in Spuriously Populated Data Streams", ACM Computing Surveys (CSUR), vol. 41, Issue 3, **(2009)** July.
[5] K. Das and J. Schneider, "Detecting anomalous records in categorical datasets", **(2007)**, KDD 2007.

Authors

Jinhyung Kim

She 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 is currently working toward Ph.D. degree in the computer science at the same university. Her researches interests include protect techniques and policies in Privacy protection and Cloud Computing Security.

Jun Hwang

He 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 College of Information & Media of Seoul Women's University. His current interests are IPTV, convergence computing, and digital broadcasting.

Hyung-Jong Kim

He 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 assistant professor since March, 2007. His research interests include cloud computing security, VoIP security, privacy protection and simulation modeling methodology.