The Adaptive SPAM Mail Detection System using Clustering based on Text Mining

Sung-Sam Hong[1], Jong-Hwan Kong[1] and Myung-Mook Han[2]*

[1]Department of Computer Engineering, Gachon University
Seongnam, South Korea
[e-mail: sungsamhong@gmail.com, ball3314@naver.com]

[2*]Department of Computer Engineering, Gachon University
Seongnam, South Korea
[e-mail : mmhan@gachon.ac.kr]

*Corresponding author: Myung-Mook Han

Received April 8, 2014; revised May 23, 2014; accepted June 9, 2014; published June 27, 2014

Abstract

Spam mail is one of the most general mail dysfunctions, which may cause psychological damage to internet users. As internet usage increases, the amount of spam mail has also gradually increased. Indiscriminate sending, in particular, occurs when spam mail is sent using smart phones or tablets connected to wireless networks. Spam mail consists of approximately 68% of mail traffic; however, it is believed that the true percentage of spam mail is at a much more severe level. In order to analyze and detect spam mail, we introduce a technique based on spam mail characteristics and text mining; in particular, spam mail is detected by extracting the linguistic analysis and language processing. Existing spam mail is analyzed, and hidden spam signatures are extracted using text clustering. Our proposed method utilizes a text mining system to improve the detection and error detection rates for existing spam mail and to respond to new spam mail types.

***Keywords*:** SPAM, Text Mining, Text Clustering, Text Classification, Detection

This research was funded by the MSIP(Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2013.
A preliminary version of this paper was presented at ICONI 2013 and was selected as an outstanding paper.

http://dx.doi.org/10.3837/tiis.2014.06.022

1. Introduction

Recent network developments and an increase in mobile device usage have made it easy to connect to the internet at any time and from anywhere. As a result, email can be easily accessed from a variety of different types of devices and quickly disseminated to a large number of people. Email is one of the most important internet applications because of its convenience, low cost, and ability to reach a vast number of people almost instantaneously. Email substitutes various roles such as written letters, telegrams, advertisements, and voice mails.

As internet utilization increases, the volume of spam mail has also increased. From a business perspective, spam mail incurs almost no overhead cost. In fact, since it can impute some of the cost to recipients, it is very attractive. It is actually possible for consumers to purchase goods by looking at spam advertisements [1].

Spam mail senders illegally send mass advertisement information from remote locations. Such advertisements may cause uncomfortable feelings and economic loss to recipients. Spam mail has recently trended toward indiscriminate sending through various sending mediums such as smart phones and tablet computers. Spam mail consists of approximately 68% of total mail volumes; however, it is thought that the true volume of spam mail is much higher.

Text Mining [2] is a mining technique that finds valuable and meaningful information from unstructured text data. This technique allows users to extract meaningful information from a vast amount of information, understand its relationship to other information, and categorize it. In order for a computer to analyze information described with human language and extract hidden information, massive language sources and complex statistical orderly algorithms need to be applied. Through the propagation of social culture, such as social networking services and blogs, text mining has been utilized for advertisements, marketing, law case analysis, information search, and trend analysis.

In order to analyze and detect spam mail, this paper introduces a technique that utilizes text mining to detect spam mail by extracting the linguistic analysis and language processing characteristics of spam mail. Existing methods that filter spam mail primarily do so by setting rules to filter according to **identity** (ID), **internet protocol** (IP), or the preset letter string of senders. These methods analyze spam mail in advance and generate spam mail signatures for filtering. Our method analyzes existing spam mail and extracts its characteristics to detect future spam mail. Spam mail classified as such by users is also analyzed through text classification, and hidden spam signatures are extracted through text clustering.

The remainder of this paper is organized as follows: Related work is discussed in Chapter 2, and the proposed system is introduced in Chapter 3. In Chapter 4, spam mail analysis that utilizes text mining is explained in detail. Finally, we provide our concluding remarks in Chapter 5.

2. Related Works

2.1 Text Classification

For text classification, the data classification algorithm, utilized by existing data mining, is broadly used. Data classification classifies input data into classes based on the results of learning. In this paper, **K-nearest neighbor** (KNN) classification is utilized. The KNN algorithm predicts the value of a new entity using previously known entities stored in memory.

This training set enables a new entity to be categorized accordingly [3].

The KNN algorithm calculates a weighted property value and a similarity value. Let X and Y be data formed by K properties, and let x_i and y_i be the i^{th} property value of X and Y, respectively. Assume T is a purpose property, and $T(X,Y)$ is the similarity of X and Y to T. We define $\Omega_T(X,Y)$ as

$$\Omega_T(X,Y) = \sum_{j=1}^{k} \omega_T(j) \cdot S_T(x_i, y_i) \ , \quad (1)$$

where $\omega_T(j)$ is the i^{th} weighted property value, and $S_T(x_i, y_i)$ is the similarity of property values between x_i and y_i. Hence, the similarity calculation in the KNN algorithm is classified into two stages: $\omega_T(j)$ is the weighted value for each property, and $S_T(x_i, y_i)$ is the similarity between property values.

2.2 Text Clustering

Text clustering differs from existing classification standards in that it analyzes given data and groups similar data according to similarity characteristics. Therefore, it can express characteristics of given data and extract unrecognized hidden data. Moreover, utilizing text clustering also makes it possible to predict data. In this paper, the K-means clustering algorithm [4] and hierarchical clustering [14] have been used. K-means is a type of data mining algorithm that measures the distance of entities for clustering. Data is converted to numerical values in order to measure distance, and a value of K is chosen to determine the number of groups the data should be divided into. Once K is set, data is divided into K clusters based on the numerical information of each entity. All entities are divided into K groups centered at the point nearest to themselves. If clustering is only performed once, K clusters are sufficient for finding the center point of each group for re-clustering; however, clustering can be repeated to generate K new centers. The process of re-clustering is repeated until a random critical value, set by the user, is satisfied. In this way, groups can be formed of the most ideal type.

Three factors must be taken into account for the K-means algorithm to be successfully applied [5]. First is the value of K, or the number of clusters chosen, second is how text information is converted to numerical values for the distance calculation, and third is the critical value that determines when clustering should be stopped. Each of these factors is very important because they affect computing time and the overall effectiveness of the system.

3. Spam Mail Detection System using Hybrid Text Analysis

Most spam mail is unstructured data formed by text; therefore, text mining is one of the most effective methods for detecting and analyzing spam. Similar to existing methods based on IP information, strings of letters can be utilized to analyze existing spam mail.

The system proposed in this paper automatically analyzes existing spam mail and updates spam mail detection rules automatically. The system can be largely classified as a text mining system. In the mail service system, detected spam mail is filtered, and warning messages are

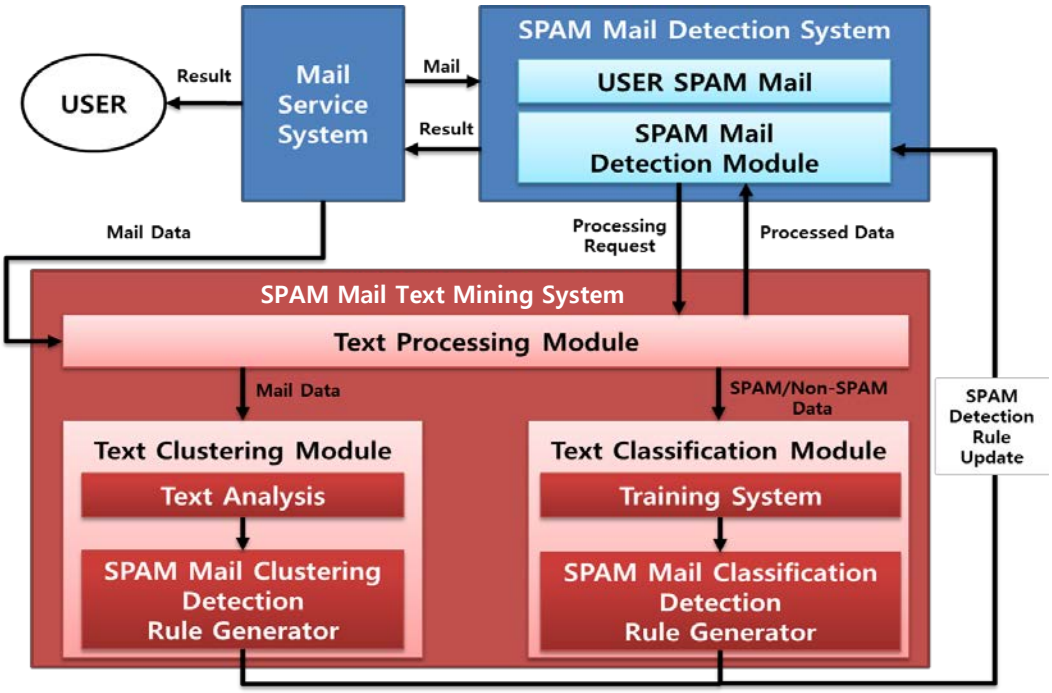

Fig. 1. The Proposed SPAM Mail Detection System

provided to users. The advantage of this system is that it can be improved by raising the detection rate of spam mail and by utilizing characteristics of spam mail generated by text mining methods. The system uses clustering to extract characteristics of new spam mail and rules are automatically updated when a new spam message is received; that is, spam mail detection rules are adaptively generated. **Fig. 1** shows the proposed system structure.

3.1 SPAM Mail Text Mining System

The SPAM Mail Text Mining System is classified into the Text Clustering Module, Text Classification Module, and Text Processing Module. The Text Processing Module conducts text processing such as stopword remove, stemming, and feature selection for pre-processing.

The Classification Module consists of a Training System and Classification Detection Rule Generator. The Training System learns spam/non-spam mail classified beforehand. The Classification Detection Rule Generator generates detection rules for classifying the resulting data. By learning and analyzing existing mail data in the Classification Module, the precision of spam mail detection is increased. Moreover, learning spam mail classified as such by a user aids the detection of spam mail.

The Clustering Module clusters data using Text Analysis as input. By analyzing all data, a rule is generated to detect new types of spam mail by extracting hidden knowledge. This module also supports existing detection rules to improve the detection rate and error detection rate.

As already mentioned, the detection rate of spam in this system can be improved by analyzing existing spam mail information. Through text clustering, the characteristics of unknown spam mail can be extracted from the whole mail data, and the characteristics of individual mail can be classified.

3.2 SPAM Mail Detection System

The SPAM Mail Detection System updates rules based on information supplied from the Text Mining System. Using a predetermined rule and filtering signature of IP, ID, or letter, mail data sent from the Mail Service System is screened for possible spam mail. The SPAM Mail Detection System also plays a role in sending spam data to each user. The spam mail detection process is as follows:

① Received mail is sent from the Mail Service System to the SPAM Mail Detection System,

② Delivered mail is sent to the Text Processing Module,

③ Processed data is sent to the SPAM Mail Detection Module,

④ Spam mail is identified,

⑤ Results are reported to the Mail System.

By following this procedure, a filtering method, based on existing letter string, IP, and ID information can be utilized. The Text Mining Detection Rule can be used to identify similar documents and spam mail by using characteristics of each word. Eventually, a system is established that can easily be applied to an existing spam mail filtering environment. As a result, the detection rate of spam mail is increased, and spam mail trends can be used to identify new spam mail.

4. SPAM Mail Analysis using Text Mining

In this chapter, spam mail is analyzed using text mining by utilizing R [6] of the Text Mining Framework. The tm package of R [7] is provided for text mining.

4.1 Data

The data used for analysis is provided by spam and Ham [8] mail data and consists of 500 spam and 2500 non-spams messages (Ham). The text document data set is called the Text Collection and uses a unit called a Corpus.

4.2 Pre-Processing

A basic text pre-processing was conducted for the text analysis. To begin, text is extracted from input data and all punctuation is removed. All characters are changed to lowercase and stopword is removed. The stopword is removed in accordance with the English stopword list appointed by the tm package.

For future performance improvements, keywords and stopwords should be determined by the environment, which will improve the speed and quality of processing by reducing overall dimensionality.

4.3 Term-Document Matrix

Analyzing unstructured text requires a standardized mathematical model. The

Term-Document Matrix (TDM) is the matrix that presents the frequency of each document by term. The TDM is obtained from the corpus after pre-processing is complete. The characteristics of the corpus can be analyzed through TDM.

Table 1. Top 20 Terms of Spam Mail

Term	Frequency	Term	Frequency
Widthd	1272	Can	624
Font	1190	free	554
Email	1038	facearial	538
Table	995	div	536
Will	904	please	495
Sized	878	facearial	451
helvetica	830	height	432
Width	757	html	431
sanserif	725	arial	426
size	686	faceverdana	393

Table 2. Top 20 Terms of Ham Mail

Term	Frequency	Term	Frequency
can	1469	people	831
list	1282	date	827
will	1264	mailing	822
just	1141	time	770
one	1111	now	741
get	1057	email	727
use	1045	message	683
like	991	url	664
wrote	920	also	656
new	842	said	629

TDM is used as the Feature Vector for the Clustering Algorithm. For example :

	can	network	…	visit
Document 1	20	2	…	0
Document 2	2	0	…	11
….	…	…	…	…
Document 3	1	13	…	32

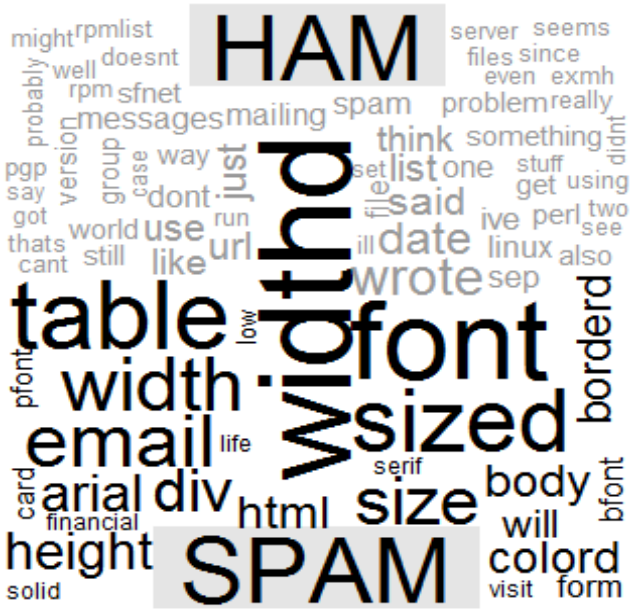

Fig. 2. The Visualization of Ham and Spam Term Analysis Results

4.4 Term Frequency Analysis

A basic text analysis examined the frequency of terms appearing in the corpus. The spam mail document data consisted of a total of 22,170 terms, and among those, 41 terms appeared more than 300 times. The non-spam mail data set included 29,146 terms, 76 of which appeared more than 300 times. Many HTML codes were found since they were not deleted prior to the pre-processing step. In future experiments, we plan delete HTML codes beforehand.

Table 1 shows the top 20 terms and their associated frequencies in the spam mail data set. The higher frequency terms in the table are very important for detection rule generation. Similar results for non-spam are shown in **Table 2**.

4.5 Visualization

By visualizing the analysis results, we are able to better understand their implications. Through visualization, our results can be seen at a glance, and the characteristics of the data set can easily be seen. **Fig. 2** is a visualization screen that uses wordcloud [9]. The top 200 words with the highest frequencies in both spam and non-spam mail sets are visualized. Bold font terms correspond to words appearing in spam mail, and gray font terms are those appearing in non-spam mail.

4.6 Spam Mail Detection Experiment using Text Clustering

A clustering experiment was conducted to evaluate the text mining method of detecting spam mail. The potential for the proposed system to detect unknown spam mail types is demonstrated in this experiment.

Table 3. The Number of Spam and Ham Mails

Document	Ham	Spam
300	252	48
600	497	103

4.6.1 Data Set

Two sets of data, the 'Ling Spam Data Set [10]' and 'Enron Spam Data Set [11],' were used in the experiment (HTML code deleted), and the experiment was carried out using 300 and 600 documents for each set, respectively. **Table 3** shows the numbers of Spam and Ham mails for each number of documents.

4.6.2 Differences of Text Preprocessing

In order to compare the performance of text pre-processing methods, documents pre-processed by stemming and stopword were compared using text mining and text clustering, respectively.

Table 4. Methods of Measurement

Measure	Formula	Meaning
Precision	$\frac{TP}{TP+FP}$	The percentage of positive predictions that are correct.
Recall/ Sensitivity	$\frac{TP}{TP+FN}$	The percentage of positive labeled instances that were predicted as positive.
Specificity	$\frac{TN}{TN+FP}$	The percentage of negative labeled instances that were predicted as negative.
Accuracy	$\frac{TP}{TP+TN+FP+FN}$	The percentage of predictions that are correct.

4.6.3 Clustering Algorithm and Measure Method

The clustering algorithms used for the experiment were K-means clustering and **hierarchical clustering methods** (Hclust). Clustering was performed in two groups corresponding to two clusters. In order to evaluate the performance of clustering, four measurements were used. These measurements are as shown in **Table 4** [12]. The explanation of each measurement is as follows:

- True Positive (TP) : The number of spam documents correctly classified as spam.
- True Negative (TN) : The number of non-spam documents correctly classified as non-spam.
- False Positive (FP) : The number spam documents classified as non-spam.
- False Negative (FN) : The number of non-spam documents classified as spam.

Table 5. Experiment of Spam Mail Clustering

Document (Ling Spam)	**Pre-processing**	**Algorithm**	**Precision (%)**	**Recall/ Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**
300	None	K-mean	**100**	58.33	**100**	93.33
		Hclust	**100**	6.25	**100**	85.00
	stemming + stopword	K-mean	85.41	**85.41**	97.22	**95.33**
		Hclust	**100**	6.25	**100**	85.00
600	None	K-mean	41.52	68.93	79.80	78.00
		Hclust	**100**	33.00	**100**	88.50
	stemming + stopword	K-mean	97.72	**84.90**	99.59	**96.83**
		Hclust	95.29	78.64	99.19	95.66
Document (Enron Spam)	**Pre-processing**	**Algorithm**	**Precision (%)**	**Recall/ Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**
300	None	K-mean	**100**	64.58	**100**	94.33
		Hclust	**100**	0.04	**100**	84.66
	stemming + stopword	K-mean	**100**	18.75	**100**	87.00
		Hclust	**100**	64.58	**100**	94.33
600	None	K-mean	**100**	56.31	**100**	92.50
		Hclust	**100**	69.90	**100**	94.83
	stemming + stopword	K-mean	**100**	50.00	**100**	93.98
		Hclust	**100**	71.84	**100**	95.16

4.6.4 Clustering of the Ling Spam Data Set

The first experiment was conducted by clustering 300 and 600 documents of the Ling Spam Data Set. Experimental results are shown in **Table 5**. While the documents pre-processed showed good clustering results overall, when the number of documents was reduced to 300, the recall/sensitivity of the documents preprocessed by Hclust was very low (6.25%). The performance and accuracy of K-means was found to be very stable.

The overall average performance of K-means when the number of documents was 600 was high, with 92.08% accuracy on average. When the documents were preprocessed using the Hclust method, overall performance rapidly improved.

4.6.5 Clustering of the Enron Data Set

The second experiment was conducted by clustering 300 and 600 documents of the Enron Spam Data Set. Overall, results were poor for K-means when the number of the documents was 300. On the other hand, in the case of Hclust, the recall/sensitivity and accuracy increased from 0.04% to 64.58% and from 84.66% to 94.33%, respectively, when pre-processed

documents were clustered; overall performance of Hclust was high. Average performance of Hclust was also high when the number of the documents was 600.

4.6.6 Clustering Result Analysis

When the results of the overall clustering experiment were analyzed, the best rates were 96.83% for accuracy, 100% for precision, 100% for specificity, and 85.41% for recall/sensitivity. In fact, accuracy and recall/sensitivity of pre-processed documents improved by approximately 4.26% and 8.08%, respectively, compared to documents not preprocessed; thus, pre-processing documents improves clustering performance. The K-means clustering method, in general, exhibited stable performance. Experimental results indicate that the clustering method can accurately detect spam mails, with a 90.9% accuracy rate for the entire experiment. Therefore, not only can existing types of spam mail be detected, but unknown types can be detected as well.

5. Conclusion

Much research has been devoted to the classification and analysis of spam mail [13]. Unlike other proposed mail detection systems, this paper proposed a hybrid spam mail detection system, which utilizes text clustering and text classification. This proposed method utilizes the Text Mining System to improve the detection rate and error detection rate of existing spam mail and to respond to new spam mail types. The clustering method for detecting spam mail was also verified experimentally using two clustering methods.

For future research, we plan to study classification and clustering algorithms that conform to the English mail environment. Algorithms used to generate spam mail detection rules will also be studied. Furthermore, to process big data, we plan to study feature selection and dimensionality reduction methods, which are utilized during text pre-processing.

References

[1] Ho-Sub Lee, Jae-Ik Cho, Man-Hyun Jung and Jong-Sub Moon, “An Approach to Detect Spam E-mail with Abnormal Character,” in *Journel of Korea Institute of Information Security & Cryptology*, Vol.8, No.6, pp 129-137, 2008. Article (CrossRef Link)

[2] Hearst and Marti A. “Untangling text data mining,” in *Proc. of Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 3–10, 1999. Article (CrossRef Link)

[3] Altman, N. S. “An introduction to kernel and nearest-neighbor nonparametric regression,” *The American Statistician*, Vol.46, No.3, pp 175–185, 1992. Article (CrossRef Link)

[4] MacQueen, J. B., “Some Methods for classification and Analysis of Multivariate Observations,” in *Proc. of Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967. Article (CrossRef Link)

[5] Ki-joon Lee, Jin Myung Lee, Woo Ju Lee, “The Search Method of Blog using K-means, ” *The Proceeding of Korea Intelligent Information System Society*, pp 269-275, 2009. Article (CrossRef Link)

[6] http://www.r-project.org/. Article (CrossRef Link)

[7] http://cran.r-project.org/web/packages/tm/index.html. Article (CrossRef Link)

[8] Drew Conway and John Myles White, *Machine Learning for Hackers*, O'Reilly Media, 2012 Article (CrossRef Link)

[9] http://cran.r-project.org/web/packages/wordcloud/. Article (CrossRef Link)
[10] Androutsopoulos, J. Koutsias, K.V. Chandrinos, George Paliouras, and C.D. Spyropoulos, "An Evaluation of Naive Bayesian Anti-Spam Filtering," in *Proc. of 11th European Conference on Machine Learning (ECML 2000)*, pp. 9-17, 2000. Article (CrossRef Link)
[11] V. Metsis, I. Androutsopoulos and G. Paliouras, "Spam Filtering with Naive Bayes - Which Naive Bayes?," in *Proc. of Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS 2006)*, 2006. Article (CrossRef Link)
[12] M. Basavaraju and Dr. R. Prabhakar, "A Novel Method of Spam Mail Detection using Text Based Clustering Approach," *International Journal of Computer Application*, vol 5, no 4, 2010. Article (CrossRef Link)
[13] Alaa El-Halees, "Filtering Spam E-mail from Mixed Arabic and English Messages: A Comparison of Machine Learning Techniques," *The International Arab Journal of Information Technology*, Vol. 6, No. 1, pp 52-59, 2007. Article (CrossRef Link)
[14] R. Sibson, "SLINK: an optimally efficient algorithm for the single-link cluster method," *The Computer Journal (British Computer Society)*, Vol 16, No.1, pp 30–34, 1973. Article (CrossRef Link)

Sung-Sam Hong was born in Seoul, Korea, in 1983. He received the Bachelor degree in Computer Science from Kyungwon University, Korea in 2009 and Master degree Computer Science from Kyungwon University, Korea in 2011. He is currently a researcher(Ph.D candidate) for Big Data, Data Mining and Information Security in Information Security Lab of Gachon University. His research interests include Multimedia Security, Information Security, Mobile Security, Cryptology, Data Mining, Big Data.

Jong-Hwan Kong received the Bachelor degree in Computer Software from Kyungwon University, Korea in 2012 and Master degree Computer Engineering from Gachon University, Korea in 2014. He is currently a Ph.D. candidate in the Department of Computer Engineering, Gachon University, Korea. His research interests include Network Security, Information Security, Data Mining, Internet of Things Security.

Myung-Mook Han received MS degree in computer science from New York Institute of Technology in 1987 and Ph.D. degree in information engineering from Osaka City University in 1997, respectively. From 2004 to 2005, he was a visiting professor at Georgia Tech Information Security Center(GTISC), Georgia Institute of Technology. Currently, he is a professor in the Department of Computer Engineering, Gachon University, Korea. His research interests include Information Security, Intelligent System, Data Mining, Big Data. He is a member of IEEE and IEICE.