ICIC Express Letters
Part B: Applications
Volume **11**, Number **6**, June **2020**

ICIC International ©2020 ISSN 2185-2766
pp. 577–586

FORECASTING CYBER THREATS THROUGH THE TEXT MINING ANALYSIS OF POLITICAL NEWS AND SECURITY LOGS

Heeyoung Choi[1], Juyoung Kang[2] and Sangun Park[3]

[1]Department of Management Information System
[2]Department of e-Business
School of Business
Ajou University
206 Worldcup-ro, Yeongtong-gu, Suwon 16499, Korea
{ h2748; jykang }@ajou.ac.kr

[3]Department of Management Information System
College of Economics and Business Administration
Kyonggi University
154-42 Gwanggyosan-ro, Yeongtong-gu, Suwon 16227, Korea
supark@kgu.ac.kr

Received December 2019; accepted March 2020

Abstract. *According to the Jan. 18, 2018, security news article, "Political and diplomatic situations have begun to flood cyberspace", cyber-threat acts by Russia and North Korea have soared significantly since 2017, indicating that hackers have taken advantage of geopolitical tensions. In other words, the number of attacks by hackers involving political and diplomatic situations has increased; to prove this phenomenon, we extracted keywords that have appeared in a number of political articles published in the central newspapers of the nation. In addition, by analyzing the relationship between political issues and security attacks through text mining analysis of political news, including the log data of typical security equipment – DDoS and IPS equipment – we were able to analyze and even predict cyberattacks based on these results. Consequently, we have found a link between certain keywords and some attacks, and based on these analyses, we want to derive a proactive response through forecasting a growing number of cyberattacks in the event of similar social issues. In particular, as security threats have characterized an increasing trend of intelligent cyberattacks against specific targets, an effective preemptive defense system can be constructed by seeking ways to enhance the security capabilities of these targets.*

Keywords: Cyberattack, Text mining, Machine learning, Keyword analysis, Security logs

1. **Introduction.** The average number of cyberattacks targeting South Korea overseas is 1.5 million per day [1], and the number of cyber-threats from Russia and North Korea, which have exploited political hostilities since 2017, is also on the rise [2]. At the time of the PyeongChang Winter Olympics, Russian military spies attempted cyberattacks by disguising themselves as North Korean military officers [3].

In overseas cases, cyberattacks, believed to be Chinese hackers targeting Finland, increased during the talks involving President Trump and President Vladimir Putin [4]. A Russian hacker attempted a malicious code attack that tricked users by creating fake apps that took advantage of the political situation in Central Asia and Russia [5]. One of the common features of this phenomenon is that political tensions serve as a pretext for increasing hacker attacks. This offers the possibility of predicting cyberattacks using the analysis of political situations. Today, typical security devices that protect against

DOI: 10.24507/icicelb.11.06.577

cyberattacks include IPS (intrusion prevention system) and antiDDoS, which apply signatures and blacklist-based detection methods. However, these existing defense systems have limitations in predicting and defending new security threats such as APT (advanced persistent threat) attacks that hackers use based on political situation [6].

Therefore, preemptively addressing potential threats, rather than engaging in passive defense has been the required mode of action. Predictive analysis is the analysis of current and historical facts to predict future or unknown events using patterns found [7]. To predict and defend against security threats in advance, large volumes of data accumulated over a long period of time should be analyzed for cyberattack patterns to prevent similar symptoms from being detected [8]. The shift from post-cyberattack to proactive responses from predictive analyses based on historical data analysis affects the security analysis and log analysis areas as well [9]. The Gartner Group [10] predicted that security analysis using big data analytics can enhance the business value of a company by discovering previously unseen accident patterns and providing insight into corporate management, including information security.

In this study, we want to analyze political news media articles to predict cyberattacks. As described above, many recent cyberattacks have been related to political tensions; thus, we want to perform predictive analyses of security threats by analyzing the content of political articles and their association with cyberattacks. To this end, we analyze security logs and political articles on cyberattacks to show which political issues inspire which attacks. It is expected that this will contribute to the establishment of a preemptive hacking defense system by discovering patterns of security attacks and facilitating the forecast of such attacks in the future when the same political situation occurs. The rest of this study is organized as follows. In Chapter 2, we review the related literature on cyberattacks and text mining. Chapter 3 presents the research procedure and then describes the analytical results. Chapter 4 discusses the results and contributions to the field.

2. **Literature Review.**

2.1. **Types of cyberattacks and defense systems.**

Recent cyberattacks have been characterized by scanning and monitoring target networks for intrusive ports several years before accidents occur and then attempting various attacks at the proper time. Network cyberattacks are largely classified as distributed denial-of-service (DDoS) attacks, scan attacks, hacking attempts, worm attacks, and intelligent, advanced persistent threat (APT) attacks [11].

DoS attacks have developed into a distributed form of attack at the option of the attacker, as an attack that paralyzes the target network by means of draining the resources of a particular system or draining bandwidth [12]. A DDoS attack is an attack aimed at stopping service by sending network packets of greater processing power to the target system [13].

A scan attack refers to checking the operation of a server that provides services by means of information gathering for hacking and checking the services that it provides. Scanning attacks can collect vulnerabilities in the port [14].

Hacking attempts attack the vulnerability of CGI, a program installed during Web server deployment, involving an attempt to view directory listings or perform remote commands after obtaining server permissions exploiting bugs in CGI [15]. A worm attack involves sending unrecognizable e-mails or putting malicious code into a normal file before the program is distributed, and worm codes targeting infrastructure and control systems continue to be detected [16]. An APT is a cyber threat attack that has been widely used recently as a method of continuous attacks on government agencies or certain companies for political or cyberterrorism purposes [17]. The defense system against cyberattacks

FIGURE 1. The defense system of cyberattacks

consists of DDoS defense, intrusion prevention systems (IPSs) and firewalls, as shown in Figure 1.

Table 1 shows the DDoS defense and IPS attack detection and defense methods. DDoS protection is a system that detects and defends against DDoS attacks as described above. Intrusion prevention systems (IPSs) serve in critical positions in the network configuration [18]. IPS is installed in front of the internal main servers to protect internal systems from external threats. IPS devices detect and block scan attacks, hacking attempts, and worm attacks for collecting hacking advance information [18].

TABLE 1. Security system detection rules criteria

Criteria	Intrusion Detection and Defense Method
DDoS defense device	∘ Selective defense – Unauthorized access IP blocking (External → Internal) ∘ DDoS attack alarm and defense – Pattern and threshold (CPS) based detection *CPS (connection per second): Allowed thresholds per user per second
IPS	∘ Scan, hacking attempt, worm attack alarm and defense – Intrusion attempt IP block – Pattern based detection

In security equipment, the alarm and defense criteria are based on the attack count within the acknowledged attack time; when the count is less than the threshold value, it is regarded as an alarm target attack, and when the number exceeds the threshold value, the defense is activated.

To detect APT attacks more effectively, the aim is to identify cyber threats that mimic legitimate users who are not detected on the network. By upgrading to a cyber threat prediction and analysis service model that utilizes big data technology, it is possible to respond to such external threats by identifying fragmented infringement signs and by following up on signs of threats and analyzing each phase of an APT [19].

2.2. **Intelligent security data analysis.** A variety of security technologies based on data analysis are also being studied, as attack technologies have become advanced and diversified, requiring technologies to detect attacks and threats and analyze or predict patterns of new attacks [20]. Organizational hacker groups have engaged in cyberattacks that target specific groups in a deliberate manner. Consequently, major information breaches, control system attacks, and cyber weapons become more serious, causing social confusion and threatening national security [21]. Therefore, research using big data analysis as a form of intelligent security has been increasingly conducted.

In accordance with this research trend, various studies have been conducted to analyze the event logs managed by an enterprise security manager (ESM) using machine learning algorithms to improve intrusion detection accuracy [22]. In addition, these studies proposed a model that can classify and detect attack mail from features used in existing

machine-learning-based spam detection models, and those used the analysis of documents studied in the field of text mining and in spam detection models [23].

2.3. **Text mining.** Text mining is the process of extracting and processing meaningful information by applying natural language processing technology [24] and document processing technology to unstructured data, and discovering new knowledge from unstructured large-scale text [25]. Using the text mining technique, the social phenomenon can be observed based on the data [26]. To detect financial fraud, such as illegal lending, a methodology was proposed to analyze unstructured data in the form of a public relations letter sent to social networks, to analyze which articles are related to financial fraud. This methodology was also used for pilot testing of financial fraud prevention programs by local governments [27]. In this way, text mining is useful for extracting the main concepts shown in texts, as well as for understanding and visualizing relationships with other concepts [28]. In this study, we extract the main issues of political news articles, which are text data, with text mining and identify the relationship to security attacks.

2.4. **News media analysis.** Research that involved predictions using the news has been mainly conducted with regard to the stock market and real estate pricing [29]. One study explains that both domestic and foreign news related to North Korea has an impact on the foreign exchange market and stock market, particularly the stock market, rather than the foreign exchange market [29]. There is also a study that predicts the response variable, the house lease price, through a variation on how news keywords changed over time [30]. A study of share prices of Israeli companies on how they react after news reports of the Middle East peace process revealed that stock yields fluctuated more widely than on a day without political events [31]. By analyzing cyberattack patterns using big data engines, similar symptoms can be found and preemptively blocked based on the results of that analysis [32]. This study aims to demonstrate whether political situations can affect the course of attack through association with political articles rather than simply a pattern of cyberattacks appearing independently.

3. Prediction of Cyber Attack through News Keyword Analysis.

3.1. **Experiment design and procedure.** This study assumes that the cyberattack of the day is predicted using the collection of news from that day. Therefore, the aim of this study is to analyze and predict the link between the daily news vector and the daily cyberattack count.

Figure 2 shows the research procedure of this study. First, we collected news keywords from BIGKinds that we thought could be associated with cyberattacks and converted them into normalized keywords after a process of tokenization and normalization. For each news item, we converted the keywords into a BOW (bag of words) to generate a count vector of a certain length. Because news that is associated with a cyberattack on a certain day is supposed to be a collection of news on the same day, the next step is to add a daily news vector by counting the count vectors generated for each news. For each of these vectors, the number of attacks collected from the cyberattack log was collected on a daily basis and linked to dependent variables on the same date. The training set and test set were generated by splitting the generated data set, and the optimal model was generated using cross validation and grid search for multiple linear regression, ridge regression, and lasso regression schemes on the training set. Finally, the optimal model was evaluated using a test set.

3.2. **Data collection and preprocessing.** BIGKinds[1] is a site that collects articles from various newspapers, builds an integrated database of articles, and provides various news analysis services such as POS tagging analysis, entity name analysis, and network

[1]https://www.bigkinds.or.kr/

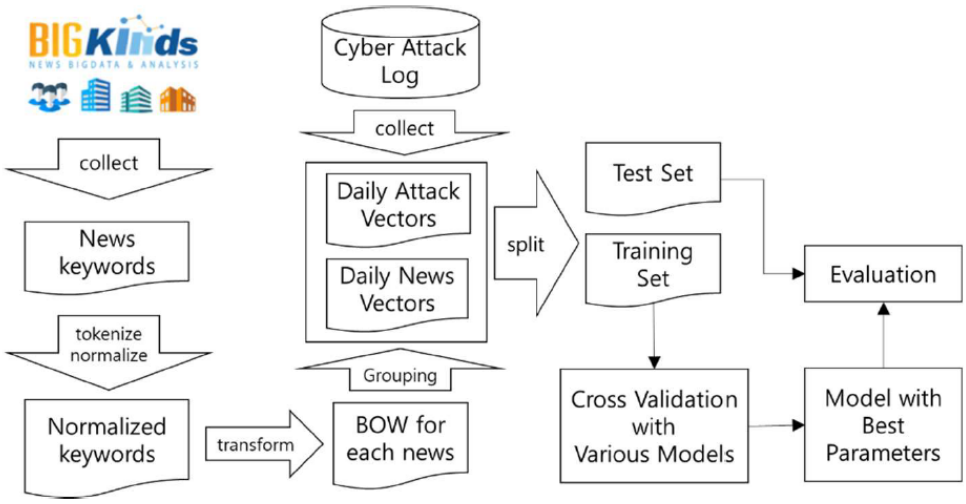

FIGURE 2. The overall research procedure

analysis. In particular, keyword analysis is performed for each article to provide important keywords for the source. In this study, we collected keywords from the political news section from July 1, 2017 to June 30, 2018 to analyze the link between political articles and cyberattacks. The total number of political news items collected was 124,990.

Because the collected news keywords consist of one document for each news item, we performed tokenization just as in the first step of general natural language processing. As each keyword is already validated by BIGKinds, we did not perform normalization separately. Tokenized keywords were converted into count vectors for each news item using the Scikit-learn CountVectorizer class, except the keywords that appeared in less than 1% of the news. Consequently, 1,809 keywords were used. This means that the dimension of the count vector is 1,809; thus, when regression is performed using the news, the number of independent variables is 1,809.

Predictions of cyberattacks are performed on a daily basis; thus, count vectors for each news item were added for the same day news. Therefore, 1,809 independent variables were created in the form of fixed-length vectors per day and there were 318 days of related news during the year. The security log was collected from the security system of public corporation A. Cyberattack data were extracted from the logs of the period July 1, 2017 to June 30, 2018, as in the news, from the daily suspected attack detection count of cyberattacks and the number of automatic blocking of harmful attacks. Table 2 summarizes the daily average number of incidents classified by cyberattack described in the literature review. The most frequent attacks were DDoS alarms with an average of 114,515 daily attacks. Meanwhile, the average number of scan defenses was six cases per day. DDos, scan, hacking, and worm attacks all had fewer defenses than alarms. The news vector generated above was an independent variable, and the number of cyberattacks counted above was generated by the dependent variable, which was then separated into a training set and a test set for further analysis. Thus, 286 out of 318 days were assigned to the training set and 32 days were assigned to the test set.

3.3. **Preliminary multiple linear regression results.** The first model used in the experiment is a simple multiple linear regression model. Table 3 shows the results of regression for eight dependent variables. While R^2 for training sets are all equal to 1, for the test set, all values except those for "worm alarm", "DDoS alarm", and "scan alarm" were meaningless. As shown in Table 2, the daily average number of incidences of scan, hacking, and worm attacks is very small compared to other values. It seems that the pattern of values is not trained properly. Meanwhile, "DDOS defense" was not well trained despite a sufficient daily average. The larger difference in the R^2 score between

TABLE 2. Daily average number of cases (unit: case)

Criteria			Daily average number of cases (unit: case)
DDoS	Alarm		114,515
	Defense		7,156
IPS	scan	Alarm	2,255
		Defense	6
	hacking	Alarm	3,737
		Defense	56
	worm	Alarm	10,248
		Defense	302

TABLE 3. Multiple regression results

Criteria			R^2 for training	R^2 for test
DDoS	Alarm		1	0.496
	Defense		1	−0.804
IPS	scan	Alarm	1	0.545
		Defense	1	−74.708
	hacking	Alarm	1	−2.587
		Defense	1	−10.153
	worm	Alarm	1	0.279
		Defense	1	−0.379

the training set and the test set indicates that the regression model is overfitting for the training set. For example, in the case of "scan alarm", the R^2 score in the training set is 1, which explains all the training set values but only 54.5% for the test set. To address this overfitting problem, L2 regularization with no coefficients is performed in ridge regression [33]. Therefore, in the next step, ridge regression was performed to reduce the overfitting and improve the predictive power on only three meaningful results.

3.4. **Enhancement with ridge regression.** Ridge regression can be used to find the most suitable model by adjusting the ratio of the L2 regulatory term added to the objective function, using alpha as a model to prevent the coefficient from overfitting using L2 regulation. To find the most appropriate alpha, we performed a grid search for the ridge regression analysis and selected an alpha for each model as shown in Table 4. In the grid search, a suitable alpha was searched on a log scale; then a secondary grid search was performed on the peripheral values of the selected alpha. The table shows the R^2 values for the validation set and the test set in the selected optimal model. The worm alarm was improved from 0.279 to 0.598 and the DDoS alarm was improved from 0.496 to 0.759. The scan alarm was a little off the score in the training set; however, the score of the test set showed higher results. We believe that the above results show that it is meaningful to predict the number of cyberattacks through news only for the three models with R^2 score greater than 0.5 in the test set. Therefore, a third model, lasso regression, was performed to analyze the main keywords that actually affect these three models.

TABLE 4. Ridge regression result

Criteria			Alpha for best model	R^2 for test
DDoS	Alarm		400	0.759
IPS	scan	Alarm	500	0.538
	worm	Alarm	700	0.598

3.5. **Keyword analysis with lasso.** Lasso regression is similar to ridge regression in the sense that regularization is performed; however, it differs in that covariates are selected through soft thresholding on coefficients [34]. In this step, we use this characteristic of lasso regression to select keywords that have an important influence on prediction. Table 5 shows the results of lasso regression for the three independent variables selected in the previous step. Likewise, a grid search was performed to find the optimal alpha. As a result, the validation score of "worm alarm" dropped but the test score remained at 0.5. Next, we performed a lasso regression analysis on all the data using the retrieved alpha values to find keywords that have a major association with cyberattacks. Table 6 shows the number of reduced keywords for each independent variable. The R^2 score shows that the selected words account for more than 80% of cyberattacks. For example, in the case of DDos alarm, only 46 keywords (2.5% of all keywords) account for 86.5% of cyberattacks.

Table 5. Lasso regression result

Target	alpha for best model	R^2 for test
DDoS alarm	3,000	0.557
scan alarm	40	0.518
worm alarm	100	0.499

Table 6. Lasso score on the entire set

Target	R^2	Non-zero coefficients
DDoS alarm	0.865	46
scan alarm	0.841	66
worm alarm	0.891	93

Table 7 shows the top 30 keywords with the highest coefficients for the three models. Among the keywords associated with DDoS detection attack, scan detection attack and worm detection attack, the keywords of "Kim Jong Il", "against South Korea", "Communist Party", and "North Korea policy" are common keywords.

Table 7. Top 30 keywords

Target	Keywords
DDoS alarm	Settlement, route, the very day, China visit, step by step, Kim Ki Sik, the first trial, primary race, agenda, boundary, report, performance, government officer, origin, birth, research lab, mind, freezing, 30th, May, Kim Jong Il, broadcasting, protocol, solution, tertiary, feeling, reorganization, revision, overseas, solidarity
scan alarm	May, route, active, Kim Sung Tae, six years, Kim Ki Sik, day after day, supreme council member, rebuttal, trade, meal, competition, government office building, meeting, Tokyo, place, 2014, against South Korea, under, performance, Internet, Chosun, tourism, entrance, protocol, message, setting, group, pursuit, pointing
worm alarm	Address, special activity cost, congress, general assembly, public service, next year, 2nd, Rex, system, trial, opening, participants, radio, security guard, effect, left side, working level talks, battle, July, 2015, settlement, Communist Party, North Korea policy, political retaliation, recovery structure, summit, human right, consciousness, two countries, violation

Cyberattacks from North Korea have diversified in recent years and have been characterized by purpose-oriented attacks focused on information gathering from institutions and individuals rather than the unspecified majority, and inter-Korean relations are influencing the flow of cyberattacks [35]. North Korea has conducted six nuclear tests so far and there have been cases of cyber terrorism attacks since the nuclear test. After the 6th nuclear test (September 3, 2017) that took place during the analysis period, government officials, major security companies and telecommunication businesses gathered two days later on September 5, 2017, to prepare for such cyberattacks [36]. The analysis shows that these attacks, which were influenced by political issues including inter-Korean relations, are relatively weak in intensity and are often carried out as detection attacks.

3.6. **Discussion on the experiment results.** Table 2 shows that the daily average occurrences for scan, hacking, and worm defenses are very small compared to other values. This can be interpreted as a failure to properly learn the pattern of values. However, the DDoS defense was not well trained even when the average number of daily occurrences was sufficient. Based on this result, prediction in a news-based defense appears to be difficult to conduct, because of the lack of relevant political articles compared with the alarm. In addition, among alarms, hacking efforts were difficult to predict using political news. By contrast, DDoS, scan, and worm alarms exhibited greater association with political articles and showed R^2 scores greater than 0.5 in forecasts. These results demonstrate that some of the cyberattacks might be predictable based on political articles.

4. **Conclusion and Contribution.** This study analyzed the relationship between cyberattacks and political news in the central newspapers of the nation to demonstrate the influence of political issues on cyberattacks. For this purpose, the association between them was analyzed by extracting keywords that are associated with the presence of cyberattacks using text mining in news articles. An analysis of eight kinds of cyberattack found that the association was high for DDoS alarm, scan alarm, and warm alarm, and the forecast performance was also good. This shows the possibility of predicting cyberattacks through the analysis of political news. In the analysis period of this study, it was noticed that the top 30 keywords associated with cyberattack include keywords related to North Korea such as the 6th nuclear test by North Korea.

This study is based on one year of news articles and cyber-attack logs; thus, a limitation of this study is that the target period is relatively short. In further studies, it will be possible to analyze the impact of various political issues on cyberattacks by extending the data by more than two years. From a methodology perspective, classification with the labels of the increase, decrease, and retention of cyberattacks instead of the number of direct cyberattacks is expected to help improvements in the prediction performance by applying more diverse classification algorithms, such as SVM, Decision Tree, and XGBoost. In addition, it is expected that higher performance will be obtained by considering a larger variety of news and adding other variables, rather than limiting the data to political articles.

Acknowledgement. This work was supported by the Ajou University research fund.

REFERENCES

[1] BoanNews, *National Security Threat Hacking Issues*, 2018.
[2] BoanNews, *The Political and Diplomatic Situation Has Begun to Flood into Cyberspace*, 2018.
[3] BoanNews, *U.S. Intelligent Agency Says Russia Is Responsible for the Cyberattack on the PyeongChang Winter Olympics*, 2018.
[4] BoanNews, *Trump, Putin, During the Talks, Has Seen a Surge in Cyberattacks Targeting Finland*, 2018.
[5] BoanNews, *Octopus Malware That Exploits Political Situations in Russia and Central Asia*, 2018.

[6] D.-K. Kim, S.-B. Pyo and C.-H. Kim, Study on APT attack response techniques based on big data analysis, *The Society of Convergence Knowledge Transactions*, vol.4, no.1, pp.29-34, 2016.
[7] wikipedia, *Predictive Analytics*, www.wikipedia.org, Accessed on May 27 2015.
[8] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. H. Byers, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, Mckinsey Global Institute (Mgi), 2011.
[9] S.-J. Lee and D.-H. Lee, Real time predictive analytic system design and implementation using bigdata-log, *Journal of the Korea Institute of Information Security & Cryptology*, vol.25, no.6, pp.1399-1410, 2015.
[10] N. MacDonald, *Information Security Is Becoming a Big Data Analytic Problem*, https://blogs.gartner.com/neil_macdonald/2011/04/12/information-security-is-becoming-big-data-problem/, Accessed on Jan.27, 2019.
[11] S.-H. Im, J.-S. Kim, J.-K. Yang and L. Chae-ho, Current status of APT and countermeasures for new malicious code, *Journal of the Korea Institute of Information Security & Cryptology*, vol.24, no.2, pp.63-72, 2014.
[12] S. Cho, T. Lee and S. Yi, Analysis of DoS attack vulnerabilities and DoS attack cases in TCP/IP network protocol, *Journal of the Korea Institute of Information Security & Cryptology*, vol.24, no.1, pp.45-52, 2014.
[13] R. Sandeep, A study of DOS & DDOS – Smurf attack and preventive measures, *International Journal of Computer Science and Information Technology Research*, vol.2, no.1, pp.1-6, 2014.
[14] Y. E. Kang, *A Study about Intrusion Examples Analysis and Security Threat Countermeasure through Construction of Intrusion Prevention System*, Master Thesis, DanKook University, 2004.
[15] E. K. Hong, *A Study on Hacking and Invasion Prevention Technique*, Master Thesis, Dongguk University, 2003.
[16] S. Hong, Analysis and countermeasure of malicious code, *Journal of Convergence for Information Technology*, vol.4, no.2, pp.13-18, 2014.
[17] J. Kim, A case analysis of targeted cyber attack and a study its countermeasure, *The Korea Association of National Intelligence Studies*, vol.9, no.2, pp.119-160, 2016.
[18] N. Y. Cho, *Security Audit Checklist and Performance Evaluation of Intrusion Prevention System (IPS)*, Master Thesis, Konkuk University, 2008.
[19] D. Jeon and D.-G. Park, Analysis model for prediction of cyber threats by utilizing big data technology, *The Journal of Korean Institute of Information Technology*, vol.12, no.5, pp.81-100, 2014.
[20] S.-S. Hong, K. Y. Shin and M. M. Han, A classification model for attack mail detection based on the authorship analysis, *Journal of Internet Computing and Services*, vol.18, no.6, pp.35-46, 2017.
[21] J. H. Kim, S. H. Lim, I. K. Kim, H. S. Cho and B. K. Noh, Technical trends of cyber security with big data, *ETRI Journal*, vol.28, no.3, pp.19-29, 2013.
[22] S. G. Choi, *A Study on the Prediction of Intrusion Types Using a Support Vector Machine*, Master Thesis, Yonsei University, 2015.
[23] S.-S. Hong, *The Model of Attack Detection Based on Intelligent Security Data Analysis*, Master Thesis, Gachon University, 2016
[24] T. S. An, H. G. Seo and G. I. Lee, Text ming based precision search system, *Korea Information Processing Society Review*, vol.11, no.2, pp.88-97, 2004.
[25] S.-Y. Kim and Y.-M. Chung, An experimental study on selecting association terms using text mining techniques, *Journal of the Korean Society for Information Management*, vol.23, no.3, pp.147-165, 2006.
[26] J. H. Noh, *A Study on the Changes in Topic and Crisis Communication in Corporate Issue: Topic Modeling Approach*, Master Thesis, Yonsei University, 2018.
[27] S. Choi, J. Lee and O. B. Kwon, Financial fraud detection using text mining analysis against municipal cybercriminality, *Journal of Intelligent Information System*, vol.23, no.3, pp.119-138, 2017.
[28] D. Paranyushkin, Identifying the pathways for meaning circulation using text network analysis, *Nodus Labs*, vol.26, 2011.
[29] W. Kim, *An Exploratory Study of Feature of Companies That React to North Korea-Related News*, Master Thesis, Industrial Engineering, SungKyungKwan University, 2016.
[30] J. M. Lee, J. A. Lee and J. H. Jung, Prediction of lease price using news big data, *Journal of Real Estate*, vol.69, no.1, pp.43-57, 2017.
[31] T. Zach, Political events and the stock market: Evidence from Israel, *International Journal of Business*, vol.8, no.3, 2003.
[32] J.-H. Kang, A study on the measures to reinforce preemptive cyber warfare using big data, *Journal of Security Engineering*, vol.13, no.3, pp.195-204, 2016.

[33] A. E. Hoerl and R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, vol.12, no.1, pp.55-67, 1970.
[34] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol.58, no.1, pp.267-288, 1996.
[35] BoanNews, *14 Days before Liberation Day, North Korea's Assumed Security Program False Attack Detected*, 2018.
[36] ETNews, *North Korea, Which Ended Its Nuclear Test, This Time Cyber Terrorism?*, 2017.