The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.7.1173
29(7), 1173–1183

Developing the information security risk index using network gathering data

Jin Woo Park[a] · Seokhoon Yun[a] · Jinheum Kim[a] · Hyeong Chul Jeong[a,1]

[a]Department of Applied Statistics, University of Suwon

(Received May 25, 2016; Revised July 6, 2016; Accepted August 10, 2016)

Abstract

In this paper, we proposed an information security risk index to diagnose users' malware infection situations (such as computer virus and adware) by gathering data from KT network systems. To develop the information security risk index, we used the analytic hierarchy process methodology and estimated the risk weights of malware code types using the judgments of experts. The control chart could be used effectively to forecast the information security risk for the proposed information security risk index data.

Keywords: malware, analytic hierarchy process, information security risk index, control chart

1. 서론

정보화의 급격한 진전과 더불어 인터넷 침해사고, 개인정보 노출, 악성코드 감염 등 정보화의 역기능 역시 급격히 증가하고 있다. 특히 2009년 7월 7일 발생한 DDoS 공격과 같은 인터넷 침해사고가 사회에 미치는 파급효과는 무시할 수 없는 수준으로, 정보화 역기능에 대한 인터넷 사용자들의 불안감 역시 날로 커져가고 있는 실정이다 (KISA, 2010, 2011, 2012). 최근 사회 전반적으로 정보화 역기능 현상을 사전에 미리 예측하고 대비할 필요성이 증가하였으며, 시스템 개발 업체들이 정보화 역기능을 대비하는데 많은 투자를 해야 한다는 압력이 강해지고 있는 실정이다. 현재, 시스템 개발 업체나 통신업체 등에서 운영하는 서버에는 악성코드와 관련된 빅데이터들이 매일 쌓이고 있다. 이에 따라 인터넷 운영업체들이 서버에 쌓여있는 악성코드 감염 및 치료에 대한 빅데이터를 분석하고, 이용자들의 악성코드 감염 및 치료패턴을 파악하여 정보보호 위험도를 예보하는 활동을 할 수 있다면, 이는 해당 시스템을 사용하는 고객들에게 큰 신뢰를 줌과 동시에 사회적 요구를 충족하는 소임을 다하는 행위라 하겠다. 즉, 고객만족 차원에서 악성코드 감염에 대해 그동안 축적된 방대한 자료를 체계적이고 과학적으로 분석하여 정보보호 대응책을 마련하는 것은 시급하고도 마땅히 이루어져야 할 일이라 생각된다.

현재, 인터넷 운영업체는 사회전반의 정보보호 현황을 예보할 수 있는 기술이나 방법을 개발할 수 있는가에 관심을 두고 있다. '예보'란 예측(prediction)의 의미를 지닌다. 예측이란 불확실한 미래를 알아보고자 하는 것이다. 현재, 예측이나 예보의 기법이 가장 널리 사용되는 곳은 기상학이나 경제학 분야라 할 수 있다. 기상학의 일기예보는 주로 물리적 법칙에 대한 예측이라 할 수 있고, 경기예보는 사회적 법칙에 대한 예측이라 할 수 있다. 그런데, 정보보호 위험도 예측은 일기예측이나 경제예측과는 또

[1]Corresponding author: Department of Applied Statistics, University of Suwon, 17 Wauan-gil, Bongdam-eup, Hwaseong-si, Gyeonggi 18323, Korea. E-mail: jhc@suwon.ac.kr

다른 측면을 지니고 있다. 그 이유는 정보보호 현상에는 물리적, 기술적 요인 및 사회학적 요인이 함께 존재하기 때문이다. 즉, 정보보호 자료는 기술적 지표와 사회적, 심리적 지표가 혼합되어 있는 매우 복잡한 구조를 띄고 있으며, 투입변수와 결과변수 간에 일종의 게임현상이 존재하고 있어서, 정보보호 현상을 예측한다는 것은 그동안 전혀 다루어지지 않은 생소한 영역으로 간주된다. 현재, 정보보호 수준 측정과 관련되어서는 다소 연구가 진행된 바 있지만, 주로 특정 조직의 정보보호 상태를 진단하는 지표(indicator)들을 개발하는 것이 주가 되어 왔다. 한편, 개발 혹은 측정된 지표를 기초로 특정 현상에 대한 지수(index)를 만들고자 하는 연구 역시 간헐적으로 진행이 된 바 있다. 정보보호와 관련된 대표적 지수는 국가정보보호지수, 개인정보보호 신뢰수준지수, 개인정보보호지수, 스팸체감지수, 정보통신윤리지수, 사이버폭력지수, 인터넷중독지수, 개인정보보호 수준진단지수, 사이버안전지수 등이 있다 (KISA, 2010). 하지만, 기존의 많은 정보보호 수준 측정은 기술적 현상에 대한 연구가 아닌 사회적 현상에 대한 연구에 치중되었으며, 월별, 혹은 분기별 자료에 기초하고 있기 때문에 매일의 정보보호 위험도 상태를 진단하기 위한 지표나 지수로 활용하기에는 많은 한계점을 지니고 있었다. 또한 대부분의 기존 정보보호 지수들이 설문조사에 의존하고 있는데, 설문조사 자료는 기술적 관측 개념이 우선 시 되어야 하는 정보보호 수준 진단에는 적합하지 못한 면이 있을 수 있다. 그러므로 본 연구에서는 설문조사가 아닌 기계적으로 수집되는 악성코드 빅데이터에 근거하여 위험도를 측정하는 방법을 적용하기로 한다. 이를 위해, 특정 네트워크(KT) 가입자들의 악성코드 감염에 대한 운영 실태를 수집 및 분석하였고, 자료에 근거하여 악성코드 감염 위험 수준을 측정하는 문제를 다루고자 한다. 그리고 위험도 예측 지수를 정보보호 현상을 예보하는데 활용할 수 있는가의 가능성 여부를 타진하기로 한다.

본 연구를 위해 특정 네트워크 서버에 축적되어 있는 데이터 스키마 및 메타데이터를 파악하여 유용한 변수들을 추출하는 작업을 우선적으로 선행하였음을 밝힌다. 본 논문의 2절에서는 위험도 지수를 개발하는데 기초가 된 원시 악성코드 자료 구조를 제시하였다. 그리고, 3절에서 위험도 지수를 제안하고, 4절에서는 관리도를 사용하여 제안한 위험도 지수의 활용성을 살펴보았다.

2. 자료구조

본 연구에서 사용한 관계형데이터베이스의 릴레이션(relation)은 모두 27개이며, 이들 릴레이션은 취합통계 테이블 8개, 사용자 정보 테이블 5개, 운영관리 테이블 2개, 에이젼트 관리 테이블 2개, 홈페이지 관리 테이블 8개, 고객요구관리(VOC) 테이블 1개, IP 대역 관리 테이블 1개로 구성되어 있다. 사용자가 특정 인터넷 망을 사용하여 인터넷에 접속하면, 악성코드(바이러스) 탐지 프로그램이 PC 백그라운드에서 자동으로 구동되며, 사용자 컴퓨터 상에 존재하는 악성코드 현황이 서버의 탐지/치료 테이블에 전송된다. 이들 바이러스 정보는 다시 코드에 따라 9개로 재분류 되는 과정으로 자료가 수집된다. 여기서, 사용자 IP는 데이터베이스의 주요 식별자(primary key)로 간주된다.

본 연구에서는 악성코드 감염률과 감염자 1인당 감염회수 및 감염파일수에 대한 정보를 핵심정보로 간주하여 관련된 릴레이션들을 결합하였다. 분석 타플(tuple)은 일별 자료이며, 분석 대상 일자는 KT 네트워크 악성코드 측정시작 시점으로부터 4개년 자료를 분석대상으로 하였다(2007년 8월부터 2010년 4월). Table 2.1은 특정 일에 해당 시스템의 유효 접속자가 10명인 경우를 가정했을 때(서로 다른 IP를 지닌 10대의 PC가 해당 통신망에 접속했는데, 바이러스가 탐지된 4개 PC만의 악성코드 감염 현황이 서버에 전송됨), 악성코드 유형 A, B, C, D(대분류)에 감염된 4명의 상태를 보여주고 있다. Table 2.1에서 악성코드별 감염자수와 악성코드 감염회수, 감염된 파일수를 Table 2.2와 같이 가공하여, 일별로 악성코드별 감염률, 감염자 1인당 감염회수 및 감염 파일수 등을 유도하였다. Table 2.3은 악성코드 종류별 일평균 감염률을 보여주고 있다. 여기서, 분석에 사용한 악성코드 종류는 모두 9종(대분류)이며, 해당 기간 동안 Trojan 코드의 감염률이 가장 높았음을 볼 수 있다.

Table 2.1. Example: raw data type with 10 users

User	Malware name	Malware type	Number of infected files
P	Adware.Generic.32911	A (Adware)	5
P	Adware.Generic.47501	A (Adware)	4
P	A.ADV.FunWebProducts	A (Adware)	3
P	A.ETC.Jyle	A (Adware)	2
P	A.TBR.BaroGo	A (Adware)	1
P	E.KSE.TWinSoft	B (Extended)	1
P	E.KSE.PointUp	B (Extended)	2
P	E.KSE.PointUp	B (Extended)	3
Y	A.ETC.waddon	A (Adware)	5
Y	E.RHT.Hosts	B (Extended)	5
Y	H.SHK.CWS	C (Hijacker)	5
Y	Trojan.Agent.AGZI	D (Trojan)	2
K	Adware.Generic.42292	A (Adware)	1
J	A.ETC.Jyle	A (Adware)	5

Table 2.2. Rate of infection for malware type with the Table 2.1

Malware type	A	B	C	D
Number of infected users (M_t)	4	2	1	1
Frequency of malware type	8	4	1	1
Sum of infected files	26	11	5	2
Infection rate (p_t)	$4/10 = 0.4$	$2/10 = 0.2$	$1/10 = 0.1$	$1/10 = 0.1$
Infection frequency per 1 infected user	$8/4 = 2.0$	$4/2 = 2.0$	$1/1 = 1.0$	$1/1 = 1.0$
Infected files per 1 infected user	$26/4 = 6.5$	$11/2 = 5.5$	$5/1 = 5.0$	$2/1 = 2.0$

Table 2.3. Daily mean infected rate for malware types

Ranking	Malware type	Daily mean infected rate (%)
1	Trojan	2.17
2	Extended	0.86
3	Adware	0.74
4	Virus	0.49
5	Worm	0.39
6	Spyware	0.28
7	Hijacker	0.25
8	Etc	0.18
9	HackingTool	0.09

3. 위험도 지수

본 절에서는 9가지 악성코드 종류에 대한 위험도의 가중치를 계산하여 악성코드 위험도 지수를 제안하기로 한다. 이를 위해 관심 대상인 9가지 악성코드 종류의 위험 정도를 계층적의사결정(analytic hierarchy process; AHP) 과정을 사용하여 측정하였다.

3.1. AHP 분석

AHP 분석을 하기 위해서는 일반적으로 다음과 같은 4가지 단계의 작업을 수행한다.

Table 3.1. The example questionnaire of AHP pairwise comparison sclae

Alternative	Importance ←									→ Importance								Alternative
	Extreme		Very Strong		Strong		Moderate		Equal	Moderate		Strong		Very Strong		Extreme		
	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	
A					O													B
A										O								C
B															O			C

Table 3.2. AHP weight of malware type

Malware type	Weight (9)	Weight (8)
Virus	0.117	0.118
Worm	0.211	0.218
Trojan	0.207	0.220
Spyware	0.196	0.206
Adware	0.053	0.054
Hijacker	0.066	0.063
Extended	0.037	-
Hacking Tool	0.089	0.096
Etc	0.023	0.025
Consistency index (CI)	0.033	0.032
Consistency ratio (CR)	0.023	0.023

단계 1: 의사결정문제를 상호 관련된 계층으로 분류하여 계층적 결정 구조를 형성한다.

단계 2: 의사결정 요소들 간의 쌍대비교로 판단자료를 수집한다.

단계 3: 역수행렬의 고유벡터를 사용하여 의사결정 요소들의 상대적 가중치를 추정한다.

단계 4: 평가대상이 되는 여러 대안들에 대한 종합순위를 얻기 위하여 의사결정 요소들의 상대적인 가중치를 종합화한다.

AHP의 적용에서 가장 중요한 단계라 할 수 있는 첫 번째 단계는 의사결정의 문제를 설정하는 것이다. 여기서 의사결정자는 상호 관련되어 있는 여러 의사결정 사항들을 계층화하여야 한다. 두 번째 단계에서는 상위계층에 있는 요소들의 목표를 달성하는데 공헌하는 직계 하위계층에 있는 요소들을 쌍대비교하는 설문을 작성한 후 그로부터 비교행렬을 생성한다. 예를 들어 A, B, C 3개의 대안에 대한 중요도를 묻는 설문은 Table 3.1과 같이 작성할 수 있다. 세 번째 단계로 상대적 가중치를 추정할 때, 한 계층 내에서 비교 대상이 되는 n개 요소의 상대적인 중요도를 w_i, $i = 1, \ldots, n$ (단, $\sum w_i = 1$)라 놓고, 쌍대비교 행렬 $A = (a_{ij})$에서 a_{ij}는 w_i/w_j $(i, j = 1, \ldots, n)$로 추정할 수 있다. 이는 행렬 A를

$$A \cdot w = n \cdot w$$

로 표현하는 것과 동일하다 (Saaty, 1980, 2003). 여기서, $w = (w_1, w_2, \ldots, w_n)$이다.

위의 식에서 보면, 일관성 있는 정확한 가중치 w_i에 대응하여 정확한 쌍대행렬 A가 주어지지만, 실제 AHP 설문에 의해서는 평가자의 일관성 있는 정확한 w가 계산되지 않는다. 그러므로 AHP 설문에 의해 얻어진 쌍대행렬을 A^*이라 하면, 평가자의 가중치 w는 A^*의 고유치-고유근 분해를 통해 다음 식의 w^*로 추정한다.

$$A^* \cdot w^* = \lambda_{max} \cdot w^*$$

여기서, $\lambda_{\max}$는 행렬 A^*의 최대고유치로 항상 n보다 크거나 같기 때문에 $\lambda_{\max}$가 n에 근접하는 값 일수록 쌍대비교행렬 A^*의 수치들이 일관성을 가진다고 말할 수 있다. 그러므로 측정된 행렬 A^*의 일관성을 측정하기 위해 일관성 지수(consistency index; CI) $(\lambda_{\max}-n)/(n-1)$와 일관성 비율(consistency ratio; CR) (CI/RI) $\times$ 100%가 사용된다. 여기서 RI는 난수지수이다. 마지막 단계에서는 계층의 최상위에 있는 의사결정의 목적을 달성하기 위하여 최하위에 있는 대안들의 우선순위를 결정하는 종합 중요도 벡터를 산출한다 (Jeong, 2010; Jeong 등, 2012; Lee 등, 2014).

3.2. 악성코드의 위험도 가중치

$$A^* = \begin{pmatrix} 1.000 & 0.678 & 0.387 & 0.640 & 2.569 & 1.683 & 3.898 & 1.230 & 5.194 \\ 1.476 & 1.000 & 1.423 & 1.476 & 5.502 & 2.825 & 5.966 & 1.445 & 7.056 \\ 2.584 & 0.703 & 1.000 & 1.084 & 4.743 & 4.384 & 4.082 & 2.408 & 6.000 \\ 1.563 & 0.678 & 0.922 & 1.000 & 4.644 & 4.384 & 4.690 & 2.825 & 6.188 \\ 0.389 & 0.182 & 0.211 & 0.215 & 1.000 & 1.062 & 1.783 & 0.725 & 3.000 \\ 0.594 & 0.354 & 0.228 & 0.228 & 0.942 & 1.000 & 2.825 & 0.699 & 3.817 \\ 0.257 & 0.168 & 0.245 & 0.213 & 0.561 & 0.354 & 1.000 & 0.570 & 1.783 \\ 0.813 & 0.692 & 0.415 & 0.354 & 1.380 & 1.431 & 1.755 & 1.000 & 4.644 \\ 0.193 & 0.142 & 0.167 & 0.162 & 0.333 & 0.262 & 0.561 & 0.215 & 1.000 \end{pmatrix}, \qquad w^* = \begin{pmatrix} 0.297 \\ 0.534 \\ 0.522 \\ 0.495 \\ 0.134 \\ 0.166 \\ 0.093 \\ 0.225 \\ 0.059 \end{pmatrix}.$$

위의 A^*는 5명의 정보보호 전문가로부터 9가지 악성코드에 대한 36회의 쌍대비교를 실시한 후 얻은 쌍대비교행렬을 기하평균 한 행렬을, w^*는 A^*의 첫 번째 고유벡터를 의미한다. 이로부터 9가지 악성코드의 위험도 가중치는 Table 3.2처럼 계산된다. 여기서, 전문가의 의견에 대한 일관성지수는 0.034 ($\lambda_{\max} = 9.268$)로 신뢰수준이 매우 높은 것으로 나타났다 (Saaty, 1980; Lee 등, 2014). Table 3.2에서 9가지 악성코드 중 worm의 위험도가 0.212로 가장 높고 다음으로 trojan 0.207, 그리고 spyware 0.196임을 볼 수 있다. 그런데, extended는 기타 확장 악성코드로 해당 기간에 결측치가 많아 실제 분석에서는 extended를 제외한 가중치를 사용하는 것이 바람직하다고 판단되었다. 따라서, extended 악성코드를 제외하면, trojan의 위험도는 0.220, worm의 위험도는 0.218, spyware의 위험도는 0.206으로, 9개 악성코드에서는 worm의 위험도가 가장 높았던 것에 비해, 8개 악성코드 내에서는 trojan의 위험도가 가장 높아짐을 알 수 있다. 하지만, 그들의 차이는 유의하지는 않다고 판단된다.

3.3. 악성코드의 위험도 지수

Extended를 제외한 8개 악성코드로 이루어진 위험도 지수를 제안하기로 한다. 이제, 하루(t일) 중 k번째($k = 1, \ldots, 8$) 악성코드 종류에 1번 이상 감염된 확률을 p_{tk}, 사용자 n_t명 중 해당 악성코드 종류에 1번 이상 감염될 사용자 수를 M_{tk}, 그리고 해당 악성코드 종류의 AHP 가중치를 w_k(단, $\sum_k w_k = 1$)), 해당 악성코드 종류의 2007년 12월 한 달간 평균 표본감염률을 p_k로 나타내기로 하자. 2007년 12월을 기준으로 삼은 이유는 감염 추세가 향후 전반적으로 감소하는 추세를 따르고 있으며, 네트워크를 통한 악성코드 관리운영이 초기에 비해 다소 안정화된 비교적 이른 시기를 선택하였기 때문이다. 이 제, k번째 악성코드 p_k 대비 t일 감염률 p_{tk}의 증감률(%)은 $(p_{tk}/p_k - 1) \times 100$이 되며, 이들을 악성코드 위험도를 나타내는 AHP 가중치를 사용하여 가중평균하면

$$\xi_t = \sum_k w_k \left(\frac{p_{tk}}{p_k} - 1 \right) \times 100 = \left(\sum_k w_k \frac{p_{tk}}{p_k} - 1 \right) \times 100$$

Table 3.3. Information security risk index value on July 2009

Week day	Date	Index	Date	Index	Date	Index
Sun	28-Jun-09	4.894	5-Jul-09	−14.941	12-Jul-09	69.823
Mon	29-Jun-09	17.963	6-Jul-09	6.537	13-Jul-09	77.415
Tue	30-Jun-09	10.254	7-Jul-09*	8.803	14-Jul-09	40.992
Wed	1-Jul-09	13.161	8-Jul-09	19.909	15-Jul-09	14.595
Thu	2-Jul-09	8.074	9-Jul-09	83.934	16-Jul-09	6.744
Fri	3-Jul-09	3.346	10-Jul-09	141.658	17-Jul-09	5.425
Sat	4-Jul-09	−8.850	11-Jul-09	113.198	18-Jul-09	−13.145

∗: DDos attack occurrence day.

이 얻어진다. 여기서, p_k는 p_{tk}의 상대적인 변화를 나타내기 위한 기준값에 불과하므로 확률변수가 아닌 고정 상수로 간주할 수 있는데, 이 경우 ξ_t의 표본값은

$$\hat{\xi}_t = \left(\sum_k \left(\frac{w_k}{p_k} \right) \hat{p}_{tk} - 1 \right) \times 100 = \left(\frac{1}{n_t} \cdot r' \cdot M_t - 1 \right) \times 100$$

이다. 여기서, $\hat{p}_{tk} = M_{tk}/n_t$, $r' = (r_1, \ldots, r_8) = (w_1/p_1, \ldots, w_8/p_8)$, $M_t{}' = (M_{t1}, \ldots, M_{t8})$이다. 또한, 네트워크 이용자 n_t의 값이 충분히 크므로, $\hat{\xi}_t \sim N(\xi_t, \sigma_t^2)$를 가정할 수 있다. 여기서, $\sigma_t^2 = \mathrm{Var}(\hat{\xi}_t) = (100/n_t)^2 r' \mathrm{Cov}(M_t) r$이 되며, 편의상 $\mathrm{Cov}(M_t)$가 t에 상관없이 일정하다고 가정하면, M_t, $t = 1, \ldots, N$을 이용하여 σ_t^2을 $s_t^2 = (100/n_t)^2 r' \Sigma r$로 추정할 수 있다. 이에 따라, N일 간의 우도값은 다음과 같다.

$$\log L = -\frac{1}{2} \sum_{t=1}^{N} \left[\log\left(2\pi\sigma_t^2\right) + \left(\frac{\hat{\xi}_t - \xi_t}{\sigma_t} \right)^2 \right] \simeq -\frac{1}{2} \sum_{t=1}^{N} \left[\log\left(2\pi s_t^2\right) + \left(\frac{\hat{\xi}_t - \xi_t}{s_t} \right)^2 \right].$$

$\hat{\xi}_t$의 값은 결국 k번째 악성코드 종류의 p_k 대비 t일 표본감염률 $\hat{p}_{tk}$의 증감률(%)들의 악성코드 위험도별 가중평균값이 되는 것이므로, 이 값을 관찰하였을 때 급격히 상승하는 현상이 일시적이라도 나타난다면 네트워크가 악성코드들의 급격한 감염률 증가로 위험에 빠질 수도 있음을 경고하는 일종의 메시지로 해석 될 수 있을 것이다. 따라서 본 연구에서는 $\hat{\xi}_t$을 t일의 종합위험도지수로 부르기로 한다. Figure 3.1은 2007년 8월부터 2010년 4월까지 위험도지수(일별 자료)에 이동평균을 적합한 그림이다. 이동평균은 지수의 저항선을 잘 나타내 준다. 그림에서 빨간색 점은 법적 공휴일을 의미하며, MA5는 5일 이동평균(파랑색 점선), MA20은 20일 이동평균, MA60은 60일 이동평균을 각각 나타낸다. 또한, 그림에서 가로축의 2007.5는 2007년의 한 가운데 날로 2007년 6월 30일을, 2008.0은 2008년 1월 1일을 의미한다. 본 논문에서는 2007년 12월을 기준으로 매일 8개 악성코드 감염률의 증감을 비교하였기 때문에, 2007년 12월(가로축 값이 2008.0 근처)이 지수값 0을 중심으로 하고 있음을 볼 수 있다. 2007년 8월의 위험도지수는 2007년 12월에 비해 평균 50%가량 높았으나, 점차적으로 감소하였으며, 2008년 4월 6월에 다시 평균 30%가량으로 위험도지수가 높아졌고, 이어서 2008년 9월에는 2007년 12월 수준으로 다시 감소하였다가, 2008년 후반에는 급격히 위험도지수가 재차 증가하였음을 볼 수 있다. 그리고 2009년 7월 이후의 위험도지수는 2007년 12월에 비해 평균 20% 정도 낮은 수준임을 볼 수 있다. Table 3.1의 초록색 수직선은 2009년 7월 7일로 DDoS 발생일이다. 이 날을 기준으로 2일 후 지수가 크게 상승하기 시작했으며, 3일과 4일 후인 2009년 7월 10일과 7월 11일에 해당 값이 141.658, 113.198로 상당히 높아졌음을 볼 수 있다 (Table 3.3 참고).

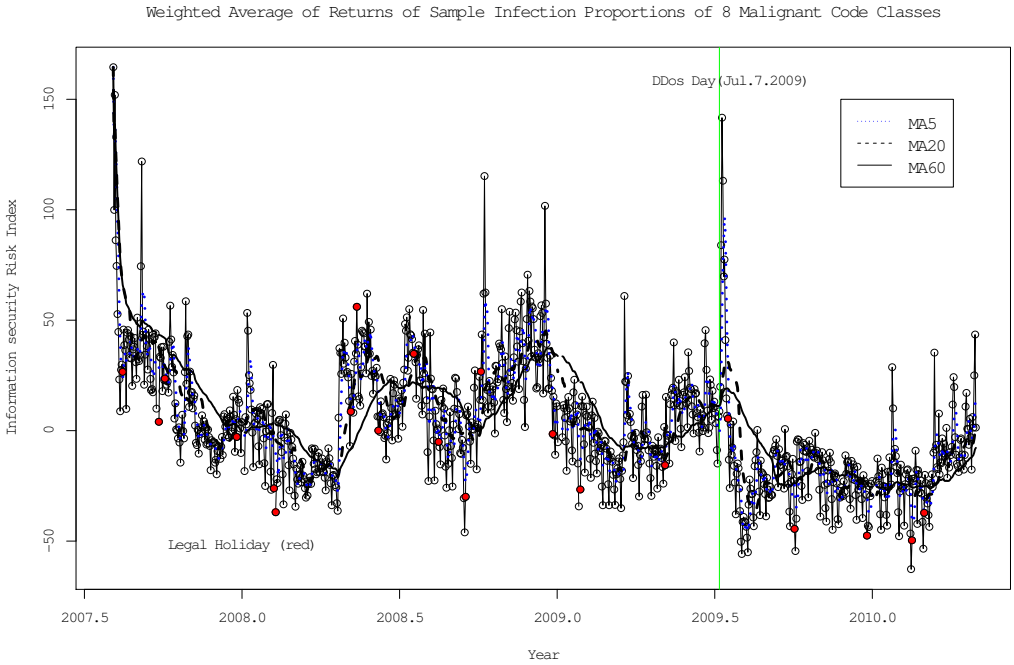

Figure 3.1. Fitting the moving average curves for information security risk index.

4. 관리도를 이용한 정보보호 위험도 예보

매일의 위험도지수를 계산하여 그래프로 나타내면, 위험도지수가 특이한 움직임을 보이는 순간을 포착할 수 있다. 이와 같은 아이디어는 실제 산업계의 품질관리 현장에서 널리 사용되는 관리도(control chart)기법과 동일하다. 관리도에서는 매 시점의 품질 특성을 나타내는 통계량을 계산하여 그래프에 그리는데, 사전에 지정된 관리 한계선(control limit)을 벗어나는 점이 나타나는 경우 품질에 이상이 있는 것으로 판단하게 된다. 일반적인 관리도에서 관리 한계선은 관심있는 통계량의 표준오차에 3을 곱한 값이다. 이를 일컬어 이른바 '3σ선'이라고 한다. 본 연구에서 다룬 위험도지수 경우, 관리 한계선을 어느 정도로 해야 하는가에 대해서는 추가적인 고려가 필요하겠지만, 일반적인 관리도처럼 '3σ'선을 적용하여, 관리 한계선을 벗어나는 경우에 위험을 예보하는 방안을 생각하기로 한다.

주어진 기간의 전체 자료로부터 3.3절에서 언급한 표본공분산행렬 근사를 사용하고,

$$\log\left(\frac{\xi_t}{100}+1\right)=\beta_0+\beta_1 t+\beta_2 t^2+\epsilon_t,\quad t=(\text{year}-2007)+\frac{\text{day}}{365.25}$$

의 선형모형으로 요일별 추세를 추정하기로 한다.

Figure 4.1은 일요일의 평균위험도지수 추세 관리도를, Figure 4.2는 월요일부터 토요일의 추세 관리도를 보여준다. Figure 4.1의 일요일 그림의 수직선은 Figure 3.1과 같이 2009년 7월 7일의 DDoS 발생일을 나타내고, 실선은 적합된 평균위험도지수의 추세선을 의미한다. 이 추세선의 위 아래에 대칭적으로 그려져 있는 두 개의 곡선은 관리 한계선이다. 그림에 평균위험도지수의 단기 추세 곡선으로 해석할 수 있는 lowess 곡선(자홍색 Lowess 곡선)과 5주 이동평균(MA5), 20주 이동평균(MA20)을 추가하여 나타내었다. 해당 그림은 R 프로그램을 사용하여 구현하였다 (Venables 등, 2010). Figure 4.1의 일요

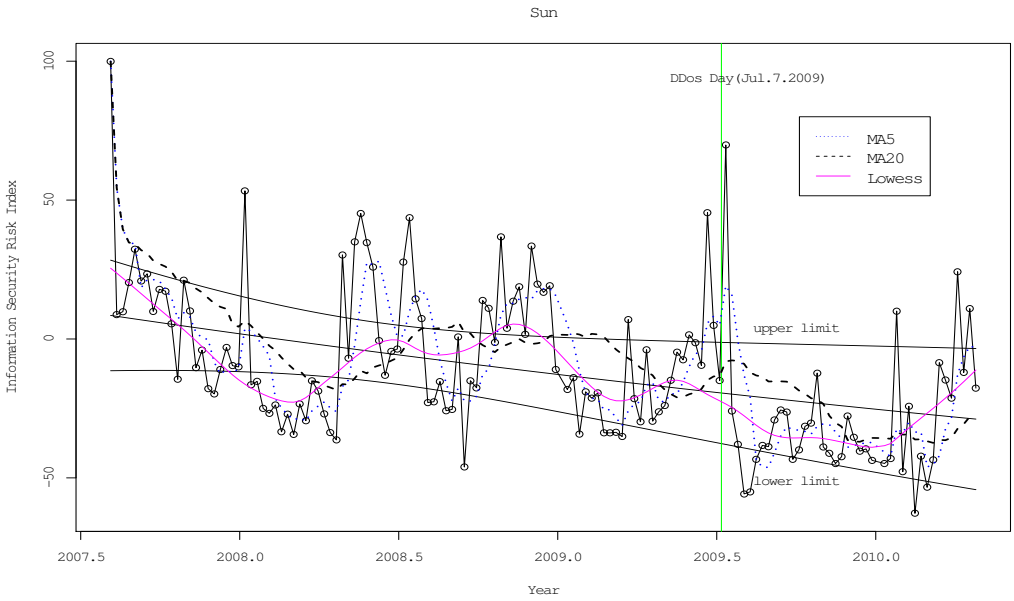

Figure 4.1. Control chart of information security risk index on Sunday.

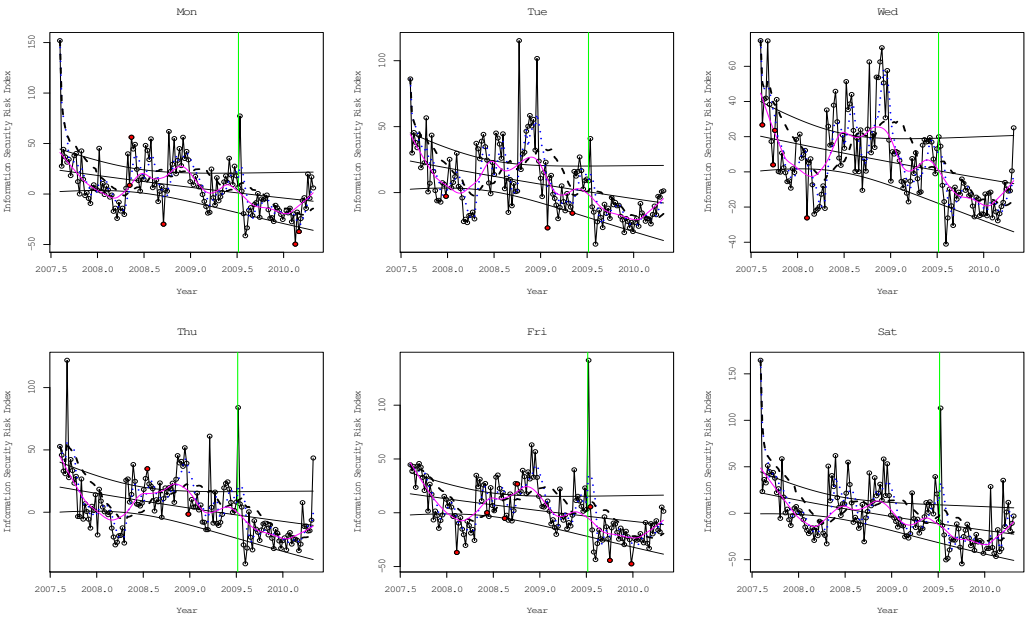

Figure 4.2. Control chart of information security risk index from Monday to Saturday.

일 관리도를 살펴보면 전체적으로 위험도지수가 3 상한선을 초과하여 군집을 이루고 있는 것이 중앙에 3개, 그리고 오른쪽 끝 부분에 1개 정도 발견되는데, 처음 것은 2008년 4월경에 시작하여 7월경까지였고, 두 번째 것은 2008년 10월경에 시작하여 12월경까지, 세 번째 것은 2009년 5월경부터 7월초까지,

그리고 마지막 것은 2010년 4월경이었음을 확인할 수 있다. 또한, 초단기 추세 곡선으로 해석할 수 있는 5주 이동평균 곡선이 단기 추세 곡선으로 해석할 수 있는 20주 이동평균 곡선을 강하게 상향 돌파한 것이 역시 4개 정도 발견되는데, 처음 것은 2008년 4월경에, 두 번째 것은 2008년 10월경에, 세 번째 것은 2009년 4월경에, 그리고 마지막 것은 2010년 3월경임을 볼 수 있다. 따라서, 일요일 관리도의 경우 위험도지수의 3 상한선 초과 시점과 5주 이동평균 곡선이 20주 이동평균 곡선을 상향 돌파한 시점이 약간의 차이는 있지만 매우 비슷하게 관찰되고 있음을 발견하였다. Figure 4.2의 월요일 관리도와 화요일 관리도에서도 일요일 관리도에서와 유사한 상황이 관찰되었는데, 일요일 관리도에서의 마지막 네 번째 군집은 나타나지 않았다. 이는 이 네 번째 군집이 자료 관찰의 마지막 부분에 걸쳐 있었기 때문인 것으로 판단된다. 또한, 수요일 관리도에서는 3σ 기준으로는 일요일 관리도의 세 번째 군집에 해당하는 DDoS 발생 근방의 군집이 사라졌으나, 이동평균 곡선 기준으로는 아직 미약하지만 세 번째 군집이 관찰되고 있음이 나타났다. 목요일, 금요일, 토요일 관리도에서는 수요일 관리도에서 사라졌던 DDoS 발생 근방의 세 번째 군집이 다시 발견되고 있으며, 특 별히 DDoS 발생 직후 목요일, 금요일, 토요일 위험도지수는 기준값인 2007년 12월 한 달간 평균 위험도지수보다 증가율이 각각 80%, 140%, 110% 정도로 폭등하였음을 확인할 수가 있었다. 이는 DDoS 발생일이 화요일이었으므로 DDoS 발생 이틀 후부터 위험도지수가 급반등했다는 사실을 말해 주며, 이는 관리도를 사용하여 제안한 지수를 활용할 수 있음을 나타내는 결과라 할 수 있다.

5. 결론

정보보호 수준 측정과 관련되어 많은 연구가 진행되어 왔는데, 이는 주로 월별 혹은 분기별 정보보호와 관련된 사회적 현상에 대한 진단 연구 혹은 설문조사가 주를 이룬다고 볼 수 있었다 (KISA, 2010, 2011, 2012). 그런데 본 연구에서는 KT 네트워크 서버에 축적되어 있는 방대한 악성코드 감염 자료를 기초하여, 통계분석의 시작이 되는 기초 자료 정리를 실시하고, 이에 대해 다양한 통계적 분석을 시도하는 것에 연구 의의를 두고자 한다. 결론적으로 본 연구에서는 결측으로 인해 사용자 구분이 불가한 악성코드 Extended를 제외한 나머지 8개 악성코드를 사용하여 새로운 위험도 지수를 제안하였다. 그리고, 제안한 위험도 지수를 정보보호 위험 수준 예보에 활용할 수 있는 가능성을 타진하였다. 위험도 지수에 따르면, 관찰기간 중 DDoS 발생 때까지 총 3번 정도의 네트워크 위험도 군집이 발견되었고, 관찰기간 마지막에는 4번째 위험도 군집이 다시 시작되는 것으로 파악되었다. 이제, 이러한 관리도를 실질적인 정보보호 위험도 측정 지수로 사용하기 위해서는 악성코드 감염정보 이외에 해당 시스템의 트래픽 양과 취약성 분석에 의한 사용자 접속 경로 등 부가적 정보가 결합되어야 하리라 생각된다. 또한, 효율적인 네트워크 위험 관리를 위해서는 악성코드 감염에 대한 평균 추세 연구 뿐 아니라 극단값 추세 연구도 병행되어야 할 것으로 생각된다. 즉, 정보보호 위험도 예보 및 예측을 위해서는 보다 다양한 기술적 데이터 수집이 필요하며 이들 자료의 결합을 연구할 필요가 있다 (Choi와 Jeong, 2015; Jeong, 2012, 2013). 끝으로, 본 연구 방법이 악성코드 감염 현황을 기계적으로 수집하는 기타 다수의 다양한 시스템 등에 활용될 수 있으리라 생각되며, 통계적 분석이 네트워크 상에서 수집되는 다양한 종류의 빅데이터 분석에 도입될 수 있기를 기대한다.

References

Choi, H. Y. and Jeong, H. C. (2015). Multivariate time series modeling for information security data, *Journal of the Korean Data Analysis Society*, **17**, 1309–1318.

Jeong, H. C. (2010). Study on AHP and non-parametric verification on the importance of the diagnosis indicators of personal information security level, *Journal of the Korean Data Analysis Society*, **12**,

1499–1510.

Jeong, H. C. (2012). A study on Korea domain registration forecasting, *Journal of the Korean Data Analysis Society*, **14**, 1889–1898.

Jeong, H. C. (2013). A development of spam score card using the data mining method, *Journal of the Korean Data Analysis Society*, **15**, 697–707.

Jeong, H. C., Lee, J. C., and Jhun, M. (2012). A study for obtaining weights in pairwise comparison matrix in AHP, *The Korean Journal of Applied Statistics*, **25**, 531–541.

Lee, J. C., Jhun, M., and Jeong, H. C. (2014). A statistical testing of the consistency index in analytic hierarchy process, *The Korean Journal of Applied Statistics*, **27**, 103–114.

KISA (2010). *The Study on the Public Publication Promotion related to the Information Security*, Korea Internet & Security Agency.

KISA (2011). *Survey for Information Security Industry in Korea*, Korea Internet & Security Agency.

KISA (2012). *Internet Security Focus, Statistics*, Korea Internet & Security Agency, Available from: http://www.kisa.or.kr/public/library/

Saaty, T. L. (1980). *The Analytic Hierarchy Process*, McGraw-Hill, New York.

Saaty, T. L. (2003). Decision-making with the AHP: Why is the principle eigenvector necessary, *European Journal of Operational Research*, **145**, 85–91.

Venables, W. N., Smith, D. M., and The R Development Core Team (2010). *An Introduction to R*, Available from: http://cran.r-project.org/

네트워크 수집정보를 이용한 정보보호 위험도 예측지수 개발

박진우[a] · 윤석훈[a] · 김진흠[a] · 정형철[a,1]

[a]수원대학교 응용통계학과

(2016년 5월 25일 접수, 2016년 7월 6일 수정, 2016년 8월 10일 채택)

요 약

본 연구에서는 네트워크 가입자들로부터 수집된 악성코드 감염 정보에 기초하여 악성코드 감염에 대한 위험정도를 파악할 수 있는 지수 산출 문제를 다루었다. 계층적 의사결정 방법을 사용하여 여러 악성코드들의 상대적 위험 가중치를 제안하였으며, 이들 가중치를 결합하여 위험도 지수를 산출하였다. 개발된 위험도지수에 대한 시계열 분석 및 통계적 모형 적합을 시도하였으며, 관리도를 통해 정보보호 위험을 예보할 수 있는 지수의 활용성을 살펴보았다.

주요용어: 악성코드, 계층적 의사결정론, 정보보호 위험도 지수, 관리도

[1]교신저자: (18323) 경기도 화성시 봉담읍 와우리 수원대학교, 응용통계학과. E-mail: jhc@suwon.ac.kr