Journal of The Korea Society of Computer and Information
Vol. 22 No. 10, pp. 73-81, October 2017

www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.10.073

# The big data analysis framework of information security policy based on security incidents

Seong Hoon Jeong*, Huy Kang Kim**, Jiyoung Woo***

## Abstract

In this paper, we propose an analysis framework to capture the trends of information security incidents and evaluate the security policy based on the incident analysis. We build a big data from news media collecting security incidents news and policy news, identify key trends in information security from this, and present an analytical method for evaluating policies from the point of view of incidents. In more specific, we propose a network-based analysis model that allows us to easily identify the trends of information security incidents and policy at a glance, and a cosine similarity measure to find important events from incidents and policy announcements.

▸Keyword: big data, information security policy, information security incident, text analysis, policy evaluation

## I. Introduction

정보보호 산업이 꾸준히 성장함에 따라 관련 사건사고와 정책과 관련된 소식들이 많이 보도되고 있다. 정부는 정보보호 산업 진흥을 위해 인력 양성, 산업체 지원, 제도 정비 등의 다양한 노력을 하고 있지만, 이러한 노력이 시장에 제대로 반영되고 있는지에 대한 연구는 부족한 실정이다.

이러한 상황에서 빅데이터는 정보보호 분야의 핵심 이슈를 신속하게 발견하고 이를 정책 결정에 활용할 수 있는 대안이 될 수 있다[1]. 미래창조과학부에서 발표한 '2016년도 연구개발사업 종합시행계획'에서 빅데이터를 10대 전략산업 중 하나로 선정하고 빅데이터 분석을 평가위원 선정에 활용하거나 기초연구 정책 지원 및 연구사업 추진 효율성 강화 등에 활용하려는 계획을 가지고 있다[2]. 유럽 연합의 여러 회원국과 호주, 싱가포르, 일본 등도 양질의 공공 데이터를 공개하고 데이터 기반의 정부 운영 플랫폼을 갖추는데 적극적으로 나서고 있다[3].

빅데이터 분석을 위해 뉴스나 블로그, 웹 포럼처럼 온라인 상의 데이터를 이용하면 별도의 데이터 수집 예산을 집행할 필요가 없고 현재의 사회적 이슈를 신속하게 파악할 수 있다. 이러한 방식은 전통적인 방식의 데이터 분석을 보완할 수 있다. 따라서 본 논문에서는 정보보호 동향을 가장 신속하게 파악할 수 있는 정보보호 침해사고 및 정책 뉴스를 수집하고 키워드 네트워크 분석 및 유사도 측정과 같은 본 연구에서 제안하는 방법론에 따라 정보보호 분야의 핵심 트렌드를 도출한다. 그리고 분석한 결과를 통해 정보보호 정책에 대한 제언을 덧붙인다.

## II. Related Works

최근 연구에서 빅데이터 분석을 정책분야에 활용하기 위한 다양한 시도를 하고 있다. 빅데이터에 축적되는 정보들은 현상 및 사람들의 의견, 감정 등을 포함하고 있기 때문에 이를 잘 분석하면 국민맞춤형 정책이 가능해지기 때문이다. 예를 들어, 지

• First Author: Seong Hoon Jeong, Corresponding Author: Jiyoung Woo
*Seong Hoon Jeong (seonghoon@korea.ac.kr), Graduate School of Information Security, Korea University
**Huy Kang Kim (cenda@korea.ac.kr), Graduate School of Information Security, Korea University
***Jiyoung Woo (jywoo@sch.ac.kr), Dept. of Big Data Engineering, Soonchunhyang University
• Received: 2017. 08. 01, Revised: 2017. 08. 21, Accepted: 2017. 09. 18.
• This work is supported by Electronics and Telecommunications Research Institute Research Fund(Title: A Study on Information Security Industrial Policy Direction) and also supported by Soonchunhyang University Research Fund(No: 20170270)

역과 위치에 대한 공간자료와 빅데이터를 결합한 공간빅데이터(Geospatial Big Data)를 활용하면 도시의 어느 부분으로 시장이 발달하는지 파악하여 도시성장을 관리할 수 있다[4]. 또한 보건복지 분야에서도 끊임없이 쌓이고 있는 공공 데이터를 활성화하여 복지사각지대를 발굴하고 정책을 보완할 수 있다[5]. 그리고 국방 분야도 병사관련 사고정보 분석을 통한 사고의 사전 예측이 가능하다. 또 다른 예는, 나라사랑카드 이용 현황 분석을 통해 징병 복지 개선 등의 사업화하고, 급여정보 분석을 통한 직업군인/노후설계 맞춤형 비즈니스 예측에 빅데이터를 활용하고자 하고 있다[6].

경기도에서는 환경분야 정책에 빅데이터를 활용하기 위한 시스템을 구축하여 활용중이다[7].

빅데이터를 정책 분야에 활용하려는 기존의 연구는 다음과 같다. 이영주[8]의 연구에서는 창업관련 정책을 빅데이터를 이용해 평가하였다. 해당 논문은 키워드를 추출하고, 키워드의 시계열 분석을 통해 주요 논의 방향을 파악하고, 키워드 사이의 연결망을 구축하여 네트워크 상에서 유사한 주제 그룹을 추출하였다. 본 연구에서는 도출된 키워드를 창업의 여러 영역으로 구분한 후 주요 단어가 시간에 따라 어떻게 바뀌는지 분석하였다.

송태민과 송주영의 연구[9]에서는 소셜미디어에서 키워드를 추출하고, 문서대비 발생 빈도 수를 구하고, 시간에 따른 이 값의 변화를 구하여 미래를 예측하는 모델을 제시하였다. 본 논문에서는 새로운 신호를 찾아내는 것의 중요성을 강조하고 있으며, weak 신호를 찾는 방법으로 문서대비 발생 빈도수의 증가율을 제시하였다. 또한 정책의 수용도에 영향을 미치는 요소를 분석하고, 연관분석을 이용하여 정책의 수용과 관련이 높은 정책요인을 도출하였다. 이 연구는 새로운 동향을 찾기 위해 weak 신호의 중요성을 기존의 문헌을 들어 설명하고, 측정치를 제안하였다는 점에서 매우 중요한 연구이다. 하지만 다른 정책 연구와 같이 본 연구도 해당 방법론을 검증하지는 못했다. 또한 제시한 방법은 TF-IDF (term frequency-inverse document frequency)방식을 이용했는데, 이는 TF의 값 즉, 키워드가 어느정도 발생건수가 가져야만 값이 크게 나온다. 따라서 사건이 어느정도 진행된 후에 탐지가 된다는 한계를 가진다.

권신혁 외[10]은 학술정보 데이터를 수집하여 의미망 네트워크를 구축하여 전체 연구를 몇가지 전체 그룹으로 구분한다. 해당 논문은 연구가 수행된 지역을 교차분석하여 어느 지역에서 어떤 주제의 주된 연구가 발생하는지를 파악하였다.

유예림의 학위논문[11]에서도 주제어 빈도 분석과 토픽 모델링을 통해 주제어를 클러스터링(clustering)하는 방법을 제시하였다.

이정훈의 연구[12]에서는 국가 R&D 정책 수립을 위해 지식지도를 이용하였다. 발생 빈도를 반영하여 키워드 네트워크로 표현하였다. 이 논문에서는 성과보고서, 연구자 입력 키워드, 과학기술표준 상 주제 분류간 맵, 분야별 전문가 맵등을 구성하여 비교하였다. 해당 연구는 여러 네트워크를 구성하여 비교한다는 측면에서 기존 연구보다는 좀 더 발전된 연구이다.

기존의 연구는 텍스트 분석을 통해 키워드를 추출하고, 이의 트랜드 분석과 의미망 연결분석을 실시하였다. 대부분의 기존 연구는 데이터를 기반으로 현상을 보는 것에 머물러 있다. 일부 연구에서 weak 신호로 새로운 현상을 찾아내는 것의 중요성에 대해 언급하고 있지만, 아직은 다양한 방법이 제시된 것은 아니며, 기존에 제시한 방법에도 한계가 있다. 본 연구는 단순 발생건수로 트랜드를 파악하는 것이 아니라 발생 유사도를 기반으로 이벤트를 탐지하는 방법을 제시한다. 제안하는 방법은 유사도를 기반으로 유사도가 높은 영역에서 새로 등장한 단어 리스트를 탐지하여, 이벤트를 탐지한다. 이러한 방법은 기존 연구보다 빠르게 이벤트를 탐지할 수 있다. 또한 현황을 분석하고 이벤트를 탐지하는데에서 더 나아가 정책과 교차분석하는 방법론을 제시한다. 정책과 현황의 비교 분석을 정량화할 수 있는 방법도 제시하였다. 본 연구는 단순히 키워드 발생 건수나 연결망을 분석하는 빅데이터 기반 정책 연구를 한단계 발전시키는 역할을 할 것이다.

# III. The Proposed Methodology

## 1.1 Big-data analysis framework

분석 과정은 키워드 네트워크 시각화를 통한 단어 네트워크 파악, 중심성을 이용한 트렌드 분석, 코사인 유사도를 통한 월별 유사도 파악으로 구성된다. 전체적인 분석 과정에 대한 구성도는 [그림 1]에 나타내었다.

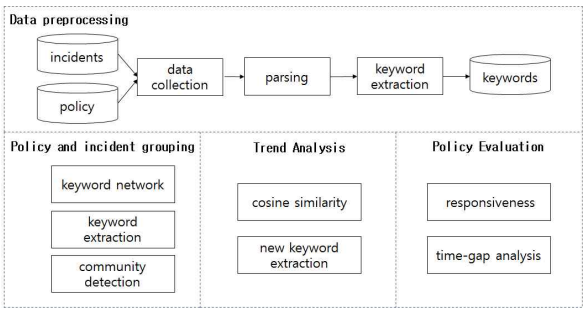



Fig. 1. Big data-based information security incident and policy analysis model

## 1.2 Policy and incident grouping

### 1.2.1 Keyword network

본 논문에서는 한글 텍스트를 분석하고 키워드를 도출하기 위해 텍스트 마이닝을 사용하였고 키워드의 중요도를 도출하기 위해 네트워크 분석(Network analysis) 기법을 이용한다.키워드 간의 상호연관관계를 파악하기 위해 그래프를 이용한 네트워크를 만들고, 네트워크에서 어떤 노드(node)가 중심 역할을 하는지, 중계자의 역할을 하는지 분석한다.

키워드 네트워크는 동일 문서에서 나타난 단어의 링크를 생

성하고, 두 단어를 포함한 문서의 개수를 가중치(weight)로 하여 네트워크를 구성한다. [그림 2]는 단어간 문서를 매개체로 2-mode 네트워크를 구성한 후, 이를 단어간의 네트워크로 구성한 모습이다.

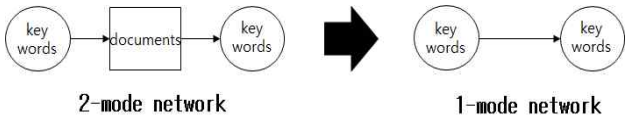



Fig. 2. Keyword network

#### 1.2.2 Keyword importance

정근하의 연구[13]에서는 네트워크 분석을 이용한 미래예측 방법 연구에서 트랜드를 나타내는 단어의 네트워크를 구성하고, 연결 중심성(degree centrality), 매개 중심성(betweenness centrality), 근접 중심성(closeness centrality) 등을 이용하여 핵심 트랜드를 도출하였다.

연결 중심성은 한 노드와 이웃한 노드들의 엣지(edge) 수를 측정하여 중심 역할을 하는 노드를 찾는 방법이다. 매개 중심성은 두 노드 사이를 경유하는 노드의 비율을 계산하여 중계자 역할을 하는 노드를 찾는 방법이다. 마지막으로 근접 중심성은 각 노드 간의 거리를 계산하여 다른 노드와 거리가 가까운 노드를 중심노드로 선정하는 방법이다.

여러 중요도 측정 지표중에서 키워드 중요도는 연결 중심성과 근접 중심성의 관점에서 측정하였다.

#### 1.2.3 Keyword grouping

네트워크를 구성하게 되면, 네트워크상의 커뮤니티(community)를 도출할 수 있다. 커뮤니티는 네트워크 연결상에서 연결강도가 높은 집단으로 다른 노드와는 연결강도가 낮은 특성을 가진다. 커뮤니티 탐지 방법을 통해 침해사고나 정책에 사용된 기술 요소 간 그룹을 알 수 있다.

### 1.3 Trend Analysis

#### 1.3.1 Event detection

본 연구에서는 이벤트 탐지 방법으로 코사인 유사도 측정 방법을 제안한다. 코사인 유사도(cosine similarity)는 두 벡터간의 방향의 유사성을 측정하는 방법이다[14]. 시간에 따른 연속된 벡터를 측정하면 시간에 따른 변화가 존재하는지를 측정할 수 있다.

코사인 유사도는 일반적으로 사용되는 방법으로 두 벡터사이의 유사도를 구하는 측정치이다. 이를 텍스트마이닝에 적용하기도 하는데, 두 문서 사이의 유사도를 측정하는데 사용한다[15]. 코사인 유사도 값은 [0,1]의 구간에 존재하게 되며, 일반적으로 0.5이상이면 두 벡터사이의 반응성이 높다고 판단한다.

권혁민 외 연구[16]에서는 코사인 유사도 기법을 응용하여 자기유사도(self-similarity)라는 지표를 제안하고, 로그의 반복성을 측정하는데 활용하였다. 본 연구에서는 단어의 발생 여부를 dummy variable로 나타내고, 시간에 따른 반복성 측정을 위해 자기유사도 지표를 텍스트 분석에 맞게 적용하였다.

코사인 유사도 측정을 위해 주요 단어 N개를 키워드로 선정하여 t 시점의 발생 여부를 나타내는 값을 벡터(vector)화 한다. 식(1)에서 K가 이에 해당한다. 키워드 벡터 K는 k를 원소로 갖는데, ki는 i번째 주요 단어의 발생 여부를 나타내는 값으로 1 또는 0의 값을 가진다.

$$\cos(\theta) = \frac{K_t \cdot K_{t-1}}{\| K_t \| \| K_{t-1} \|} = \frac{\sum_{i=1}^{N}(k_{t,i} \times k_{t-1,i})}{\sqrt{\sum_{i=1}^{N}(k_{t,i})^2} \times \sqrt{\sum_{i=1}^{N}(k_{t-1,i})^2}} \quad (1)$$

특정 침해사고가 연속적으로 발생한다면 이는 특정 침해사고의 중요성을 나타낸다고 할 수 있다. 본 연구에서는 이벤트 탐지를 위해 위에서 제안한 코사인 유사도를 이용한다. 마찬가지로 주요 정책 발제의 이벤트 또한 코사인 유사도로 측정한다.

#### 1.3.2 New trend detection

과거와는 다른 동향을 파악하기 위해서는 새로 도출되는 키워드를 모니터링하는 것이 필요하다. 해당 방법은 코사인 유사도가 높은 시점을 대상으로 단순히 비교대상 시점의 단어를 뽑아내고, 과거 단어 리스트에 없는 단어를 추출한다.

### 1.4 Policy Evaluation

정보보호 정책이 침해사고의 주요 키워드를 얼마나 다루고 있는지를 분석하여 정책을 평가하는 방법을 제안한다. 제안하는 방법의 첫 번째는 정책과 침해사고 커뮤니티 그룹 간 대응 정도를 수치화한다. 두 번째는 정책이 침해사고에 빠르게 대응하는지를 분석하기 위한 방법으로 시간차 분석을 제안한다.

#### 1.4.1 Policy responsiveness for incidents

본 연구에서는 대응력 분석을 위해 두 가지 지표를 제안한다. 첫 번째는 포함정도(coverage)로 침해사고의 키워드를 정책에서 어느 정도 포함하고 있는지를 나타내는 값이다. 두 번째는 특화성으로 특정 침해사고에 대응하고 있는지를 측정하는 값이다. 특화성은 일반적인 정책인지 특정 침해사고에 대응하는 정책인지를 판별하게 해준다.

Table 1. The cross analysis of policy vs. incidents.

| | policy group1 (group1) | … | policy groupn (groupn) |
|---|---|---|---|
| Incident group1 | x11 | | x1n |
| … | | xij | |
| Incident groupn | xn1 | | |

- 정책의 침해사고 대응력

$$Coverage_{group_i} = \frac{\sum_{j=1}^{n} x_{ij}}{w_i \sum_{i} \sum_{j=1}^{n} x_{ij}} \quad (2)$$

대응력(coverage)은 정책 그룹이 침해사고 그룹을 포함하고 있는 정도를 나타내는 값으로, 정책에 나타난 단어가 특정 침해사고 그룹 i에 있는 단어를 얼마나 많이 포함하고 있는지를 0~1사이의 값을 갖도록 설계되었다. 분자는 정책 그룹의 단어가 침해사고 그룹i에서 나타난 빈도를 더한 값이다. 분모는 정책과 침해사고에서 발생하는 모든 단어를 나타낸다. 지수값을 0~1로 나오게하고, 전체 사고 그룹 중 i그룹의 정도를 표현하기 위해 모든 사고 그룹의 합으로 나누어준다. 여기서 $w_i$는 침해사고 그룹i가 전체 침해사고 그룹에서 차지하는 비율로, 크기가 큰 그룹은 그 크기로 인해 공통 정책 단어의 수가 많아질수 있으므로 침해사고 그룹의 크기로 보정해준다.

- 정책의 특화정도

$$Skewness_{group_i} = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^3}{(\frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2)^{3/2}} \quad (3)$$

정책의 특화정도는 정책이 특정 사고그룹에 치우친 정도를 나타내는 것으로, 정책의 특정 기술에 특화되어있는지 여러 내용을 포함하고 있는지를 나타낸다. 본 연구에서는 일반적으로 사용되는 Skewness 공식[17]을 사용하였다.

#### 1.4.2 Time-gap analysis

정책이 침해사고의 주요 이슈를 얼마나 다루고 있는지를 분석하기 위해 침해사고의 주요 키워드와 정책의 주요 키워드의 일치도를 코사인 유사도를 이용하여 산출하는 방법을 제안한다. 또한 정책과 사건사고의 유사도를 시간차를 변경하면서 측정한다. 즉, 동일 월 비교, 침해사고를 분석하는 시점을 기준으로 +1, ..., +n 시점의 정책 그룹과 비교한다.

$$Gap(n) = \frac{KP_t \cdot \ KE_{t-n}}{\| KP_t \| \ \| KE_{t-n} \|} \quad (4)$$

식 (4)에서 KP(keyword for policy)는 정책그룹에서의 키워드 벡터를 나타내고, 벡터의 요소는 식 (1)에서 설명한 대로 특정 키워드의 발생 여부를 나타내는 더미변수이다. 마찬가지로 KE(keyword for event)는 침해사고 그룹에서의 키워드 벡터를 나타낸다. 식 (4)의 경우 침해사고 발생 시점으로부터 n시점 이후의 정책 그룹과의 유사도를 측정하는 것을 나타내었다.

## Ⅳ. Experiment Results

이 장에서는 정보보호 침해사고 및 정책 관련 뉴스를 수집·분석하여 정보보호 정책의 동향을 파악하고 평가한다. 분석 데이터는 주기적으로 데이터를 수집하는 웹 크롤러(Web crawler)를 개발하여 온라인 뉴스 사이트, 정책기관, 해커포럼 등으로부터 정보보호 관련 뉴스를 수집하여 사용하였다. 수집 데이터 통계는 [표 2]에 나와있다.

관련 데이터는 정보보호 침해사고의 전체 데이터와 정책 관련 전체 뉴스 데이터를 이용하였기 때문에 빅데이터라고 할 수 있다. 빅데이터의 의미는 크기의 의미도 있지만 크기가 작더라도 일부 샘플링 된 데이터가 아닌 전체 데이터를 이용한다는 관점이기도 하다[18]. 본 연구에서는 해당기간동안 정보보호 관련 뉴스 중 중복을 제거하고, 전체 데이터를 사용하였다.

Table 2. Collected data statistics

| | Period | # of documents |
|---|---|---|
| Incidents | 2014-01-01~ 2015-08-30 | 689 |
| Policy | 2014-01-01~ 2015-08-30 | 1,617 |

뉴스 데이터 분석을 수행하기 전에 데이터를 파싱하고 단어를 추출하는 과정이 필요하다. 먼저 뉴스 제목, 시간 등을 제거하는 파싱 과정을 거친 후 파이썬의 자연어처리 패키지인 koNLPy를 이용하여 조사를 제거하고 단어를 추출하였다[19].

### 1.1. Policy and incident grouping

#### 1.1.1 Keyword network and group detection

키워드 네트워크 분석을 통해 추출한 단어 간의 연결 관계를 분석하기 위해 그래프 시각화 도구인 Gephi를 이용하였다[20]. Gephi를 이용하여 뉴스 데이터를 분석할 때 분석이 용이하도록 사건·사고 뉴스는 연결 횟수(weight)가 3 이상인 엣지(edge), 정책 뉴스의 경우 연결 횟수가 4 이상인 엣지를 사용해서 그래프를 계산하였다. 키워드 분석 등의 계산이 필요한 분석은 파이썬을 이용하여 모듈을 구성하였다.

[그림 3]의 사건·사고 키워드 네트워크는 5개의 그룹으로 나뉘었고, 각 그룹의 주요 단어로 그룹별 특징을 파악해 금융사기, 악성코드(웹 해킹), 정보통신망법, 시스템취약점, 스마트폰으로 이름을 붙였다. 가장 많은 단어가 쓰인 그룹은 시스템해킹으로 주요 단어는 다운로드, 이메일, 악성코드였다. 사건·사고 네트워크의 추이를 알아보기 위해 시간별 발생 빈도를 살펴보면 악성코드 그룹의 발생건수가 가장 높으며 꾸준한 증가 추세에 있는 것을 확인할 수 있다.

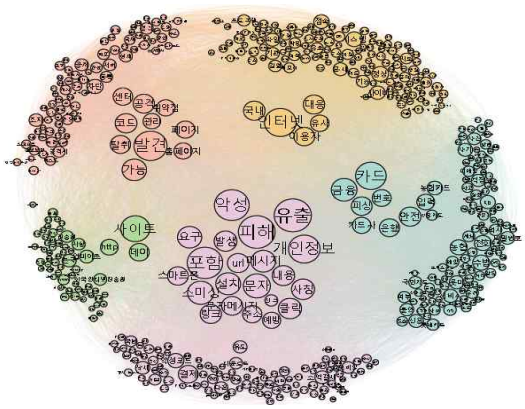



Fig. 3. Incident-related keyword network

### 1.1.2 Keyword extraction

커뮤니티 탐지 결과 각 커뮤니티를 색을 다르게 하여 표시하였다. 5개의 그룹으로 나뉘어졌다. [그림 3]에서 가운데 위치해있는 단어가 그룹간을 연결해주는 역할을 하는 매개 중심성이 높은 단어이다. 원의 크기는 연결 중심성가 높은 단어이다. 이 둘의 값을 비교하여 주요 단어를 추출하여 [표 3]에 나열하였다.

Table 3. Key words for incident groups

| Group | Key words, # | Keywords |
|---|---|---|
| #1 - Smart phone | 127 | 악성, 피해, 유출, 포함, 메시지, 스미싱 |
| #2 - Internet banking | 120 | 카드, 금융, 피싱, 번호, 카드사, 은행, 농협카드 |
| #3 - System hacking | 138 | 인터넷, 시스템, 접속, 파일, 정상, 메모리 |
| #4 - Malware and web hacking | 89 | 발견, 탈취, 코드, 공격, 취약점, 홈페이지, 차단, 스크립트 |
| #5 - Treat responsiveness | 60 | 사이트, 제로데이, 업데이트, 한국인터넷진흥원, MS, 안내 |

정책도 마찬가지로 키워드 네트워크 구축하고 커뮤니티 탐지를 수행하였더니 크게 4개의 그룹으로 나뉘어졌다. [그림 4]는 그 결과이다. 각 그룹의 키워드는 [표 4]에 정리되어 있다.

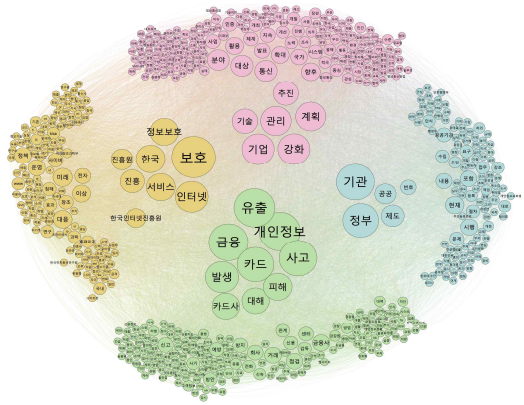



Fig. 4. Policy-related keyword network

Table 4. Key words of policy group

| Group | Key words, # | Keywords |
|---|---|---|
| #1 - Corporate management | 191 | 기업, 강화, 관리, 계획, 추진, 체계, 발표 |
| #2 - Public Agency | 140 | 정부, 기관, 제도, 시행, 포함, 도입, 요구 |
| #3 - Emerging technology | 120 | 보호, 한국인터넷진흥원, 서비스, 침해, 보안사, 대응, ICT |
| #4 - Financial institute | 189 | 유출, 개인정보, 금융, 사고, 카드, 발생, 피해 |

## 1.2. Trend Analysis

### 1.2.1 Event detection

마지막으로 월별 유사도를 알아보기 위해 월별로 자주 사용된 상위 30개의 단어를 구하여 월단위 코사인 유사도를 구하고, 그 결과를 [그림 5]에 시각적으로 표현하였다.

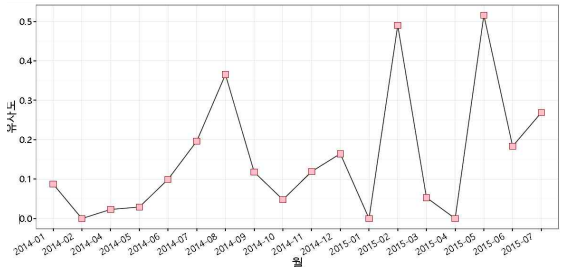



Fig. 5. Monthly cosine similarity of incidents

침해사고 기사의 월별 코사인 유사도는 대체로 0.1 이상의 값을 보였고 최대값은 0.5였다. 높은 유사도를 보인 기간은 2014년 8월~9월, 2015년 2월~3월, 5월~6월이었다. 2014년 8월~9월 사이에 공통적으로 등장하는 단어는 사칭, 택배, 스미싱 등이었다. 관련 기사를 확인해본 결과 민방위, 카카오톡 계정사칭, 초대장 등의 스미싱이 기승을 부린 것으로 확인되었다. 2015년 2월~3월에는 공유기, security, Microsoft 등의 단어가 사용되었고, Microsoft가 2월과 3월에 긴급 보안 업데이트를 발표하고 공유기 취약점을 악용하는 사례가 발생하였다. 2015년 5월~6월 역시 스미싱 단어가 자주 언급되었고 랜섬웨어, TeslaCrypt가 새로 등장하였다.

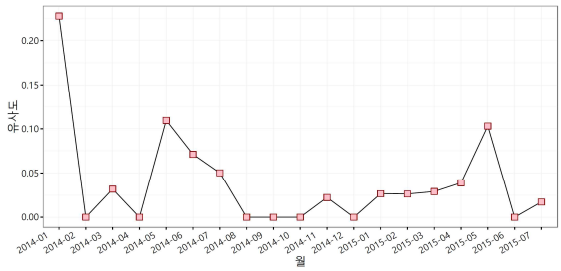



Fig. 6. Monthly cosine similarity of policy

[그림 6]에 정책 기사의 월별 코사인 유사도를 표현하였다. 정책의 월별 유사도는 대체로 0.1 미만의 값을 보였고 최대값은 0.22였다. 정책의 코사인 유사도는 침해사고의 코사인 유사도에

비해 전체적으로 낮은 값을 보였는데, 이는 정책의 특정 이슈가 여러 달에 걸쳐 다루어지는 경우가 낮음을 의미한다. 그러나 2014년 1월~2월, 5월~6월, 2015년 5월~6월의 코사인 유사도는 상대적으로 높은 값을 보였다. 1월~2월에 공통적으로 등장하는 단어는 '주민등록번호'였다. 2014년 1월 국민·롯데·농협 3개 카드사에서 고객정보 유출 사건이 발생하면서 개인정보 유출에 대한 대응의 필요성이 기사화되었기 때문이다. 2014년 5월~6월에는 미래창조과학부와 KISA가 함께 스타트업 관련 정책을 새로 발표하였고 해당 기간에 또 다른 특정 이슈가 발생하지 않았기 때문에 코사인 유사도가 높은 값으로 나타난 것으로 보인다.

월별 유사도를 측정하면 주요 이벤트를 탐지해낼 수 있고, 침해사고와 정책의 주요 이벤트를 비교할 수 있다. 비교 결과 침해사고의 주요 이벤트와 정책의 주요 이벤트는 일치하지 않는 것으로 나타났다.

#### 1.2.2 New trend detection

Table 5. List of new-coming incident topic words

| Month | Keywords |
|---|---|
| JAN | 워드프레스 |
| FEB | 팀뷰어, 팟플레이어, Sykipot |
| MAR | 공공아이핀, KIST |
| APR | 바이러스토탈, 크립토락커, 워드프레스, script, virobot |
| MAY | 크립토락커, 워드프레스, ARP, QEMU, script, 가상머신, 알리바바 |
| JUN | 크립토락커, 아키에이지, link, tech, porn, hoax |
| JUL | 크립토락커, VNC, 구글 해킹 |
| AUG | defkor, 호환, APT29, IE11, bind, micro |

○ 새로운 이슈 요약

✓ 새로 등장한 단어를 종합해보면, 대부분 웹사이트의 취약점을 이용한 해킹으로 인해 개인정보가 유출되는 사고나 가능성에 대한 것임

✓ 그 외에 원격제어 프로그램에 대한 해킹 사례나 취약점 보고가 등장함

✓ 사용자의 문서나 이미지 등을 암호화시키고 돈을 요구하는 악성코드인 랜섬웨어(크립토락커)가 스마트폰 이용 확대에 따라 모바일로 확대되고 있음

✓ SNS, 클라우드에 대한 해킹사고도 보고됨

Table 6. List of new-coming policy topic words

| Month | Keywords |
|---|---|
| JAN | 라인, 특보, 트레이딩, 관측소 |
| FEB | 국방부장, 포스트타워(개인정보보호토론회 개최장소), VoLTE, intelligence |
| MAR | egisec, 특보, ccfp, seconexpo, 코드게이트, cykor |
| APR | 비콘, WPA, 표적, ifsec. 여행객 |
| MAY | 망중립성, 무역센터, 엑스코 |
| JUN | 얼라이언스, 군사과학, 도상, NPAPI, COPS, W3C, 정보 처리 |
| JUL | 라인, CTB, 특보, bug, capture (CTF (capture the flag; 깃발뺏기 방식의 해킹방어대회)에서 파생되어 나온 일반 단어), NAS |
| AUG | 없음 |

표 6에서 보는 것과 같이 정책의 동향은 침해사고의 동향과 일치도가 낮은 것으로 나왔다.

### 1.3. Policy evaluation

#### 1.3.1 Responsiveness

Table 7. The cross analysis of incidents and policy

| Incident group | Policy group | | | | |
|---|---|---|---|---|---|
| | #1 - Corporate management | #2 - Public Agency | #3 - Emerging technology | #4 - Financial institute | total |
| #1 - Smart phone | 1,605 | 807 | 669 | 2,209 | 5,290 |
| #2 - Internet banking | 252 | 14 | 199 | 409 | 874 |
| #3 - System hacking | 983 | 37 | 282 | 219 | 1,521 |
| #4 - Malwares and web hacking | 4,419 | 337 | 5,453 | 2,049 | 12,258 |
| #5 - Treat responsiveness | 277 | 10 | 55 | 91 | 433 |
| total | 7,536 | 1,205 | 6,658 | 4,977 | 20,376 |

정책뉴스에서 사건사고의 키워드의 발생 빈도를 점수화하여 정책을 평가하면, 정책 그룹별, 사건사고 그룹별 상대평가가 가능하다.

Table 8. Policy responsiveness index

| | #1 - Corporate management | #2 - Public Agency | #3 - Emerging technology | #4 - Financial institute |
|---|---|---|---|---|
| keyword,# | 191 | 140 | 120 | 189 |
| Responsiveness | 70.6 | 8.3 | 39.2 | 46.2 |
| Specialization | 0.51 | 0.46 | 0.66 | 0.17 |

기업관리>미래기술>금융기관개인정보보호>공공기관개인정보보호 순으로 정책의 사건사고 언급도가 높은 것으로 나타났다. 또한 공공기관개인정보보호 정책군에서는 사건사고에서 나타나는 정보보호 관련어의 언급이 다른 군에 비해 현저히 낮은 것으로 보아 기술적 분석에 따른 정책 수립이 미흡하다고 판단된다.

### 1.3.2 Time-gap analysis

Table 9. Sensitivity analysis of policy response over incidents

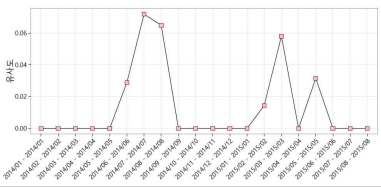
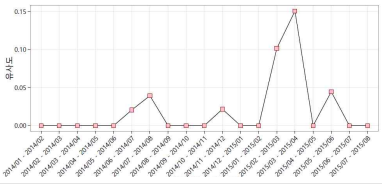
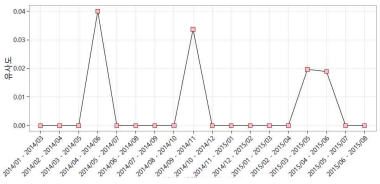
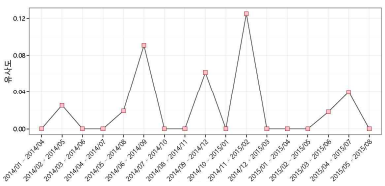
| Time-gap | Cosine similarity |
|---|---|
| 0 |  |
| +1month |  |
| +2month |  |
| +3month |  |

Table 10. Key topics of policy and incident at the point with high similarity

| Time | Common keywords | Incidents | Policy |
|---|---|---|---|
| 2014-JUL | 스미싱 | 택배, 등기소포, 민방위, 카카오톡 계정사칭, 초대장 등 스미싱이 기승을 부린 것으로 확인됨 | 미래부, 추석 겨냥 선물 배송 택배 사칭 스미싱에 사용자 주의 당부 |
| 2014_AUG | 스미싱 | | |
| 2015-MAR | 공유기 | 공유기 취약점을 이용해 공유기에 접속된 스마트폰 악성 앱 배포 | IOT 특허 소개 및 투자 필요에 대한 뉴스 |
| 2015-MAY | 북한 | 워터링홀 공격을 통한 악성코드 유포 사실 | 박근혜 대통령이 북한의 도발에 대해서는 한미 양국이 연합방위 능력을 통해 확고하게 대응 |

당월 비교 결과 스미싱, 공유기, 북한 등의 키워드와 관련해서 많은 국민이 피해를 입을 수 있는 공격이나 국가 안보와 관련된 사건에 대해서는 즉각적인 대응책을 발표하는 것으로 나타났다. +1개월과 +2개월에서는 당월에 발생했던 스미싱, 북한 이슈가 이어진 기간에 약간 높은 키워드 유사도를 보였다. +3개월에서 평균적으로 가장 높은 코사인 유사도를 보였다. 이 결과는 사건·사고가 사후에 정책으로 반영되는 시점은 사건 발생 후 3개월임을 나타낸다.

## IV. Discussions

본 연구의 한계는 제안한 방법론을 정량적으로 평가하기 어렵다는 것이다. 하지만 제안한 모델을 적용했던 시점 이후의 정보보호 침해사고 동향을 보고 간접적으로 평가할 수는 있다. 본 연구의 분석 당시 새로운 이벤트로 탐지 되었던 랜섬웨어와 공유기 해킹의 예를 들어보면, 랜섬웨어는 정책에서 다루어지지 못한 반면, 공유기 취약점에 대해서는 [표 10]에서 보듯이 IoT 보안으로 다루어졌다. 정책의 사각지대였던 랜섬웨를 살펴보면, 랜섬웨어는 2016년 하반기부터 이슈가 되고 있다. 본 연구에서 데이터 분석을 실시할 당시에 새로운 주요 이벤트로 탐지가 되었는데, 정책에서는 중요하게 다루어지지 않았다. KISA 레포트에 따르면, 국내 램섬웨어는 2015대비 2016년에 급격히 늘어난 것으로 조사되었다. 다음 표는 국내 랜섬웨어 피해 민원접수 현황이다. [표 11]에서 나타난 것과 같이 2016년 4분기에서 전 분기 대비 증가폭이 높은 것으로 나타났다[21].

Table 11. Randsomeware damage reports[21]

| | 1st quarter | 2nd quarter | 3rd quarter | 4th quarter | total |
|---|---|---|---|---|---|
| 2015 | 0 | 127 | 58 | 585 | 770 |
| 2016 | 176 | 353 | 197 | 712 | 1,438 |

## V. Conclusions

기존에 빅데이터를 정책 연구에 활용하고자 하는 연구는 많이 이루어졌지만, 상대적으로 구체적인 분석 방법과 결과를 제시한 연구는 부족하였다. 본 연구는 정보보호 분야의 침해사고 트랜드를 파악하고, 정책을 분석하기 위한 모델을 제시한다. 이를 통해 뉴스미디어의 데이터를 빅데이터로 구축하고, 이를 효율적으로 분석할 수 있다.

모델에 대한 평가를 위해 뉴스데이터를 수집하고, 텍스트마이닝을 이용해 정보보호 침해사고 동향과 정책의 대응력을 분석하였다. 이 방법은 새로운 데이터를 추출하고 활용할 필요 없이 기존의 온라인 데이터를 이용하여 사람이 일일이 분석하기 어려운 대용량의 자료를 분석할 수 있다는 장점이 있다. 또한 시각화를 통해 결과를 쉽게 파악할 수 있고, 구체적인 수치 데이터를 통해 결과에 신빙성을 더해주었다. 이러한 아이디어를 활용하여 빅데이터 분석을 다양한 분야에 적용함으로써 이전에는 보기 어려웠던 새로운 결과를 발굴해낼 수 있을 것으로 기대한다.

본 연구가 기존 연구 대비 차별점은 다음과 같다. 본 연구는 의미망 연결분석에서 그친 것이 아니라 커뮤니티 탐지로 확장시켜서 이벤트(사건사고)와 정책을 그룹화하고, 이 둘의 일치도를 계산하고, 시간차를 두어 일치도를 계산하는 방법이 기존 연구에서 사용되지 않았던 방법이라는 점이다. 이를 통해 정책을

평가하는 방법은 기존의 연구에서 시도된바 없으며, 시간차 분석을 통해 정책의 반응성을 측정하는 연구도 시도된바 없었다. 두 번째는 서론에서도 언급하였듯이 대부분의 기존 연구는 데이터를 기반으로 현상을 보는 것에 머물러 있는데, 본 연구는 단순 발생건수로 트랜드를 파악하는 것이 아니라 발생 유사도를 기반으로 이벤트를 탐지하는 방법을 제시하였다.

본 연구는 정보보호분야를 대상으로 하였으나, 이벤트와 정책을 비교하고자하는 다른 영역에도 적용가능하다. 예를 들면, 보건행정분야에서도 살충제 달걀 사건[22]과 같은 이벤트와 식재료 안정성 관련 정책을 비교 분석할 수 있다. 정책이 얼마나 현실을 반영하는지를 평가하고자할 경우에는 본 연구에서 제안한 방법론을 적용할 수 있다.

다음은 본 연구의 방법론이 적용할 때 발생할 수 있는 이슈에 대해 언급하였다. 본 연구에서 제안한 방법은 키워드의 수준과 커뮤니티 크기에 따라 분석 수준이 결정된다.

본 논문에서는 키워드 분석수준을 uni-gram으로 설정하고, 레벨 설정에 따른 영향을 최대한 줄이기 위해서 데이터 전처리를 수행하였다. 한글의 경우 띄어쓰기 문제가 발생하는데, 하나의 단어로 합칠 수 있는 부분은 합쳐서 uni-gram 수준에서도 키워드가 잘 도출될 수 있도록 하였다.

커뮤니티 크기 조절 문제와 관련해서는 네트워크 분석에서 커뮤니티 탐지 알고리즘의 설정에 따라 분석 수준이 달라진다는 것이다. 본 연구에서는 큰 커뮤니티를 도출하여 그 수를 제한하였는데, 이럴 경우 분석결과가 일반적이게 된다. 하지만 커뮤니티를 세분화하여 여러개 도출하게 되면, 좀 더 상세한 정보를 얻을 수 있다.

마지막으로 키워드를 추출할 때 정보보호 관련 단어로만 추려서 키워드 네트워크를 구축하면 정보보호 기술 측면에서의 평가가 가능할 것이나, 현재 정보보호 관련 사전이 구축되어 있지 않은 관계로 정보보호 관련 단어만으로 추출하여 정책을 평가하기 어렵다는 것이다.

## REFERENCES


[1] Korea Local Information Research &Development Institute, "A study on the big data utilization strategy by analysis of big data utilization case," 2014.

[2] Ministry of Science, ICT and Future Planning, "2016 Future creation science department R & D project comprehensive plan," 2016.

[3] Lee, J.H., "Utilization of big data for realization of Gov. 3.0," KIPA, 2013.

[4] Kim, D.J., "A smart and reliable government implementation plan with spatial data," KRIHS, 2013.

[5] Lee, Y.H., "Big data and its applications in the health and welfare sectors," Health And Welfare Policy Forum, 2015.

[6] Korea Research Institute for Strategy, “Defense data big data utilization plan,” 2013.

[7] Kim, D.Y., Lee, J.I., Song, M.Y., Kim, H.S., Choi, M.A., "Building and using big data for environmental management in Gyeonggi-Do," Policy Research, pp. 1-152, 2016.

[8] Lee, Y.J., "Methodological implications of policy analysis using big data," The Korean Association for Regional Information Society, pp. 95-109, 2015.

[9] Song, T.M., Song, J.Y., "Future signals of health and welfare policies and issues using social big data," J Health Info Stat, 41(4), pp. 417-427, 2016.

[10] Kwon, S.H., Moon, H.J., Kim, K.W., "Meaning of the world big data research network analysis based IT policy study," SAPA, pp. 327-343, 2016.

[11] Yu, Y.L., "Analysis of media coverage on 2015 revised curriculum policy using Big Data Analysis," Department of Education, The Graduate School of Seoul National University, 2015.

[12] Lee, J.H., "A Study of National R&D Big Data based Knowledge Map Application Method," Department of Information and Communication Engineering, Graduate School of Kongju National University, Gong Ju, Korea, 2016.

[13] Jeong, G.H., "A Study of foresight method based on textmining and complexity network analysis, KISTEP, 2011.

[14] Singhal, A., "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4), pp. 35-43, 2001.

[15] Tan,P.-N., Steinbach M., Kumar,V., "Introduction to Data Mining", Addison-Wesley, ISBN 0-321-32136-7, chapter 8; pp. 500, 2005.

[16] Kwon, H., Kim, T., Yu, S. J., Kim, H. K., "Self-similarity based lightweight intrusion detection method for cloud computing," In Proceedings of the Third international conference on Intelligent information and database systems-Volume Part II (pp. 353-362). Springer-Verla, 2011.

[17] Paul T., von Hippel, "Mean, Median, and Skew: Correcting a Textbook Rule". Journal of Statistics Education. 13 (2), 2005.

[18] Schonberg, V.M,, "The world that Big Data creates," 21st century books, 2013.

[19] Parko, E.J., Cho, S.Z., "KoNLPy: Korean natural language processing in Python," The 28th Annual Conference on Human & Cognitive Language Technology. 2014.

[20] Bastian M., Heymann S., Jacomy M., "Gephi: an open source software for exploring and manipulating networks," International AAAI Conference on Weblogs and Social Media, 2009.
[21] KISA, "'16 Randsomware Trend and '17 forecasting report", 2017
[22] https://namu.wiki/w/2017%EB%85%84%20%EC%82%B4%EC%B6%A9%EC%A0%9C%20%EA%B3%84%EB%9E%80%20%ED%8C%8C%EB%8F%99

## Acknowledgement

This research was supported by Electronics and Telecommunications Research Institute (Research Title: A Study on Information Security Industrial Policy Direction) and also supported by the Soonchunhyang University Research Fund(20170270).

## Authors

Seong Hoon Jeong obtained his M.S. degree in Information Security from Korea University, in Seoul, South Korea, in 2017. He is currently a Ph.D. student studying the major of Information Security at Korea University. His research interests are in the areas of intrusion detection system, software-defined network, and system security.

Huy Kang Kim received his B.S. degree in Industrial Management in 1998, M.S. and Ph.D degrees in industrial and systems engineering from KAIST in 2000 and 2009. He founded A3 Security Consulting, the first information security consulting company in Korea in 1999. Currently he is an associate professor in Graduate School of Information Security, Korea University. Before joining Korea University, he was a technical director (TD) and a head of information security department of NCSOFT (2004~2010). His research interests include solving security problems in online games based on the user behavior analysis and data mining.

Jiyoung Woo receivedthe B.S., M.S., Ph.D degree in industrial engineering from KAIST in 2000, 2002, and 2006. She is a currently professor at Big data engineering department, Soonchunhyang University. From 2008 to 2010, she was a researcher with AI lab of Arizona University, USA. She was a research professor at Graduate School of Information Security, Korea University from 2010 to 2016. Her research interest includes data mining and business intelligence.