https://doi.org/10.33778/kcsa.2019.19.2.029

# 텍스트 마이닝을 활용한 개인정보 위협기반의 트렌드 분석 연구

김 영 희*, 이 택 현**, 김 종 명**, 박 원 형***, 국 광 호****


## 요 약

과학기술 분야를 비롯한 산업영역 전반에 걸쳐 기술의 방향성과 흐름을 확인하기 위한 연구가 중요하게 대두되고, 이를 위해 대량의 데이터와 정보에서 주요 토픽을 찾아내고 분석하기 위한 트렌드 연구가 활발히 이뤄지고 있다. 이와 함께 개인정보보호 영역 또한 전망과 흐름을 사전에 파악하고, 선제적 대응을 위한 활동이 요구되고 있지만, 광의적인 형태의 정보보안 영역과 개인정보보호 관련 솔루션 기반의 트렌드 연구등 기술위주의 연구만 이뤄지고 있다. 이에 본 연구에서는 텍스트 마이닝을 활용해 개인정보보호 영역에 국한된 위협기반의 트렌드 분석을 통해 주요 이슈 토픽을 도출하고 현재와 미래의 트렌드를 미리 확인하여, 개인정보처리 기업에서 개인정보보호에 필요한 정책수립과 효과적 대응을 위한 전략수립 방향성 탐색에 활용 될 것으로 기대된다.


# A Study on the Trend Analysis Based on Personal Information Threats Using Text Mining

Young-Hee Kim*, Taek-Hyun Lee**, Jong-Myoung Kim**, Won-Hyung Park***

Kwang-Ho Kook****


## ABSTRACT

For that reason, trend research has been actively conducted to identify and analyze the key topics in large amounts of data and information. Also personal information protection field is increasing activities in order to identify prospects and trends in advance for preemptive response. However, only research based on technology such as trends in information security field and personal information protection solution is broadly taking place. In this study, threat-based trends in personal information protection field is analyzed through text mining method. This will be the key to deduct undiscovered issues and provide visibility of current and future trends. Policy formulation is possible for companies handling personal information and for that reason, it is expected to be used for searching direction of strategy establishment for effective response.


**Keywords : Data Mining, Personal Information Protection, Text Mining, Information Security**

접수일(2018년 11월 30일), 수정일(1차: 2018년 12월 17일, 2차: 2019년 6월 24일), 게재확정일(2019년 6월 30일)

* 한화시스템 ICT (주저자)
** 서울과학기술대학교 IT정책전문대학 산업정보시스템
*** 극동대학교 산업보안학과
**** 서울과학기술대학교 기술경영융합대학 글로벌융합산업공학과 (교신저자)

# 1. 서 론

정보화 및 지식사회의 진입에 따라 정보의 취득·가공을 통해 생산성 증대와 효율성 향상을 가져왔다. 하지만 그 역기능으로 사이버위협 및 해킹, 정보유출과 오·남용 등으로 인해 유·무형적 손실이 증가하고 있으며, 이와 함께 정보가 활용되어지는 IT전반에 대한 보안이 중요하게 대두되고 있고, 위협과 방어측면의 정보보안 분야의 트렌드 또한 끊임없이 변화하고 있다.

이에 국·내외 기업 및 주요 기관에서 정보보안 분야의 트렌드를 분석하여 미래 방향성 및 전략 마련을 위해 노력을 기울이고 있으며, 이와 더불어 산업전반에 걸쳐 정보와 IT가 다양한 형태로 활용되어지는 현 상황에서 정보의 안전한 보호를 위한 방향성 모색과 문제점을 파악하고 선제적 대응하기 위한 활동이 중요하게 대두되고, 그 방법 중 해당 영역의 트렌드를 찾고 분석하는 연구가 대두되고 있다[1].

트렌드 분석은 해당 분야의 방향성 탐색과 예측을 위한 정량적 분석방법으로 정형·비정형 형태의 데이터를 정량화 하고 분석하여 예측을 위한 정보를 제공하는 방법으로 이를 통해 합리적인 의사결정을 위한 지표자료로 유용하게 활용되고 있으며, 정보보호 분야 또한 기관에서 매년 발간하고 있는 이슈 및 트렌드 전망 자료와 민간 및 학계에서 다양한 트렌드 연구를 통해 급변하는 환경을 진단하고, 미래 핵심 가치 발굴에 기여하고 있다[2]. 이와 더불어 최근 대량의 비정형 데이터를 활용한 분석 기술이 크게 발전함에 따라 방대한 양의 자료 분석이 용이해지고, 가공과 분석을 위한 적합한 환경이 마련되었다.

특히 인터넷 사용자 증가에 따른 대량의 데이터 수집과 효율적 저장기술의 발전에 따라 유의미한 데이터를 도출하는 데이터마이닝에 대한 중요성이 증대되었으며, 특히 디지털화 된 다양한 형태의 비정형 텍스트로부터 새로운 의미를 찾는 텍스트마이닝 방법론이 부각되고 있다. 이를 통해 다양한 분야의 트렌드 분석, 장바구니 분석 등 IT 전 분야에 걸쳐 활용되어 지고 있다[3]. 이와 함께 개인의 다양한 식별 정보를 활용한 새로운 서비스의 등장으로 개인정보의 수집과 활용이 증가됨에 따라 개인 프라이버시 보호를 위한 기술 연구와 미래의 방향성 탐색을 위한 트렌드 분석에 대한 관심이 증대되고 있지만 정보보안 영역이라는 광의적인 범위의 트렌트 분석이 주류를 이루며, 개인정보보호에 국한된 트렌드 분석에 대한 연구는 미흡한 실정이다.

이에 본 연구에서는 국내 개인정보 영역의 뉴스기사를 분석하여 개인정보 위협 기반의 트렌드를 분석하고자 한다. 기존 정보보안 기반의 기술 예측을 위한 트렌드 분석 위주의 연구에서 탈피해 개인정보보호 분야에 국한된 주제를 대상으로 2007년에서 2016년까지의 9년간의 국내 주요 일간지 및 IT전문 기사들 중심으로 텍스트 마이닝 기법을 통해 주요 기술 동향과 트렌드 분석, 시계열 분석을 통해 개인정보보호 분야의 의미 있는 정보를 확인하고자 한다. 이를 통해 개인정보보호와 관련된 현재의 트렌드를 미리 점검하고 개인정보보호를 위한 주요 이슈 및 관련 문제점을 짚어보고, 그 해결방안을 찾기 위한 주요 단초를 제공하여 개인정보처리기관에서 선제적으로 대응하기 위한 정보로 활용 될 것이다.

# 2. 선행 연구

## 2.1 정보보안 영역의 트렌드 분석

전 산업영역 및 기술영역 전반에 걸쳐 시간 변화에 따른 추세 및 경향을 분석하기 위해 과거와 현재의 데이터와 현상을 바탕으로 미래 예측 위한 가설을 도출하는 트렌드 분석 연구가 다양한 분야에서 이뤄지고 있다[4]. 이처럼 기술의 추세와 흐름을 분석하기 위한 연구가 정보통신 분야 전반에 걸쳐 중요하게 대두되고, 이를 위해 대량의 데이터를 분석하고 감춰진 의미를 발굴하는 트렌드 분석 연구가 활발히 이뤄지고 있다. 이를 활용해 관련 기업 및 정부기관 등에서는 현재의 현상을 바

탕으로 미래의 방향성을 탐색하고 대응하기 위한 의사결정 자료로 활용되어지고 있다[2].

정보보안 영역 또한 정보의 오남용과 유출을 위한 해킹 기법 및 위협이 나날이 증대되고, 이에 대응하기 위한 방법 또한 급변하고 있어 트렌드 분석을 활용한 선제적 대응이 무엇보다 필요한 실정이다. 이에 국·내외 정보보안 기관 및 학계에서도 정보보안 분야의 트랜드 연구와 보고서를 매년 발표하고 있으며, 주요 연구로는 정보보안 영역의 현황과 기술을 파악하고 이를 통해 새롭게 이슈화되거나 점진적으로 요구되는 기술 탐색을 위해 연도별 정보보안 기술을 분석을 하고 이에 대한 연관성 분석을 통해 각각의 트렌드 간 비교를 통해 의미 도출하였다[1][17][18]. 이와 함께 빠르게 변화하는 정보보안 환경에 적용하기 위한 보안솔루션에 대한 미래 트렌드를 분석하고 변화된 환경에 맞게 최적화된 투자를 위한 보안솔루션 로드맵을 도출하는 연구와 인터넷 및 이용자 환경 등의 변화로 인한 정보보안 패러다임 변화에 대한 위협분석 연구 또한 진행되었다[5]. 그리고 한국인터넷진흥원에서 2010년부터 매년 1회 정보보안 10대 산업이슈 전망 보고서를 발간하고, 기업에서는 이를 활용해 이슈의 빠른 분석과 정보보호 관련 미래 정책 수립에 활용하고 있으며, 금융보안연구원 또한 금융 IT 보안 10대 이슈 전망 보고서 발간 등 국내 주요 기관에서 정보보안 영역 전반의 트렌드 보고서를 발간하고 기업에서 이를 활용하고 있다.

### 2.2 개인정보보호 분야 트렌드 분석 필요성

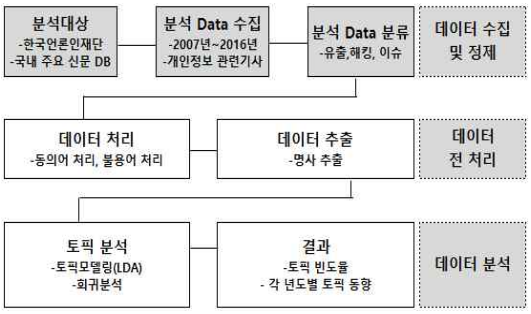
개인정보유출 및 침해등에 따른 상담 및 피해구제건수가 꾸준히 증가하고 있으며, 국내의 IT와 관련된 인프라 성장 추세와 함께 e-비즈니스 생태계에서 개인정보의 가치와 자산으로써의 중요성이 부각되어, 개인정보유출 및 침해 사례가 크게 증가 하고 있다. 이처럼개인정보 활용과 보유가 기업의 비즈니스 연속성에 큰 영향을 주는 중요 관리 대상으로 인식되고, 이를 보호하고 안전하게 관리하기 위한 활동이 무엇보다 중요한 실정이다[16]. 이에 개인정보보호 분야 또한 트렌트 연구를 통해 향후 대응책 마련을 위한 다양한 형태의 연구가 진행되고 있으며, 그 예로 광의적인 형태의 정보보안에 특화된 개인정보 기술분야 연구와 개인정보유출을 방어하기 위한 보안솔루션 도입을 위한 트렌드 분석 연구가 다수 진행되었다[6][7]. 하지만 해당 연구의 경우 개인정보보호를 위한 정보보안솔루션과 특정 기술위주의 국한된 연구만 이뤄지고 있으며, 기존에 활발하게 진행되는 있는 정보보호영역에 대한 트렌드 분석의 경우도 개인정보에 국한되지 않고 광의의 보안영역의 분석이 주류를 이루고 있어 개인정보보호에 관한 실직적인 트렌드 연구가 필요한 실정이다. 이처럼 선행연구에서 밝힌 것과 같이 기업의 해킹 및 침해사고가 개인정보와 상당수 연관되어 있고, 개인정보에 대한 중요성이 크게 증가함에 따라 기업에서 실제 발생하고 있는 개인정보보호 관련 위협에 대한 트렌드를 분석하고 이를 통해 중점 관리되어야 할 주제와 현황을 진단하고 대응할 수 있는 개인정보보호 측면의 트렌드 연구가 필요한 실정이다. 이에 본 연구에서는 개인정보보호에 대한 정량적 데이터를 텍스트 마이닝 분석을 통해 개인정보취급 현황 및 보호를 위한 지표를 제시 한다.

## 3. 연구 방법

### 3.1 연구 설계

본 연구는 개인정보보호에 관한 뉴스 정보를 토픽모델링 기법을 활용하여 위협기반의 트렌드를 분석하고자 한다. 먼저 개인정보보호와 관련된 흐름 및 이슈를 파악하기 위해 실시간성을 바탕으로 정보를 빠르고 직관적으로 전달하는 매체인 뉴스 정보 활용을 위하 주요 주제 및 이슈를 직접적으로 파악 할 수 신문 데이터를 활용한다. 수집된 기사데이터를 기반으로 데이터 전 처리 후 토픽모델링 기법을 이용해 개인정보 관련 주요 토픽 추출과 비중 분석 및 선형회기분석을 통한 시계열 분석 단계로 이뤄진다. 이를 위해 먼저 국내 개인정보보호 관련 기사 데이터를 바탕으로, 개인정보유출, 오·남용 등 위협과 관련된 주제를 포함한 20

08년에서 2016년 동안의 9년간의 기사 데이터 수집하고 동의어 처리, 불용어 처리 등의 전처리 과정의 통해 정제된 데이터를 추출하였다. 해당 데이터를 바탕으로 토픽 모델링 중 LDA (Latent Dirichlet allocation) 기법을 활용하여 개인정보 위협 관련 주요 토픽을 추출하고 분류 한다. 이와 함께 각각의 토픽에 대한 연도별 추이 분석을 통해 개인정보보호 관점에서의 기간별 상승(Hot) 주제와 점차 빈도가 줄어드는 하향(Cold) 주제를 밝혀 시계열 분석을 한다. 분석 설계 모델은 (그림 1) 과 같다.




(그림 1) 분석 설계 모델

### 3.2 데이터 수집 및 전처리

본 연구에서는 신문기사 데이터를 수집하고 활용해 개인정보보호 위협 관련 트렌드를 분석 한다. 트렌드 분석을 위해서는 관련 분야의 데이터 수집이 가장 먼저 선행 되어야 하며, 이에 신문기사 중 개인정보 위협과 관련된 주요 키워드 검색을 통해 데이터를 수집하였다. 먼저 한국언론진흥재단에서 제공하는 뉴스 빅데이터 웹사이트(www.bigkinds.or.kr)로 부터 2007년 1월 1일부터 2016년 12월 31일간의 9년간의 기사 데이터에서 “개인정보” 키워드로 검색한 8만여 건의 기사데이터를 1차 수집하고, 해당 자료 중 개인정보와 무관한 주제의 기사를 제외하기 위해 개인정보보호와 관련된 이슈 키워드로 재 검색하여 약 2천 건의 기사 제목, 본문으로 구성되는 데이터를 추출하고 데이터베이스화 하였다. 데이터 수집은 현황은 다음 <표 1> 과 같다.

<표 1> 데이터 수집 현황

| 언론사 | 년도 | 문서개수 |
|---|---|---|
| 국민일보,한계레,경향신문,서울신문,내일신문,세계일보,문화일보,한국일보,디지털타임스,전자신문 외 지역종합지, 방송사 | 2016년 | 153개 |
| | 2015년 | 225개 |
| | 2014년 | 811개 |
| | 2013년 | 206개 |
| | 2012년 | 223개 |
| | 2011년 | 185개 |
| | 2010년 | 158개 |
| | 2009년 | 82개 |
| | 2008년 | 262개 |
| | 2007년 | 76개 |
| | 합 계 | 2,302개 |

수집되고 데이터베이스화된 기사 텍스트의 경우 비정형 데이트 형태로 정보의 추출과 분석에 용의하게 하기 위해 전 처리 단계를 거친다. 먼저, 기사 내 문장 분석을 위한 단위인 단어로 분리하는 Tokenization 과정과, 분석 과정에서 불필요한 단어를 미리 제거하는 불용어 처리 과정을 거친 후 마지막으로 동일 의미를 내포하지만 다양한 형태로 표현된 단어를 공통 의미를 가진 단어로 변환하는 동의어 처리과정을 거쳐 최종 분석데이터를 도출한다. 데이터 전처리 현황은 다음 <표 2>와 같다.

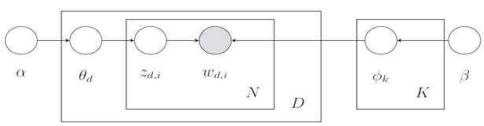
<표 2> 데이터 전처리

| 전처리 | 분리 예시 |
|---|---|
| 단어 분리 | 개인정보보호/ 를/ 위한/ 안전성확보 |
| 불용어 처리 | 개인정보/ 안전성확보조치/ 기준<br>- 동사, 조사, 형용사 어미 형태 제외<br>- 두 단어 이상 명사는 필수 명사 1단어로 변환하여 활용 |
| 동의어 처리 | “주민번호”, “주민등록”, “주민등록증 번호” -> “주민등록번호로 동의어 처리 |

### 3.3 연구 방법론

본 연구는 개인정보 위협기반 트렌드 분석을 위해 토픽모델링 기법을 활용한다. 토픽모델링은 문서가 가지는 잠재적 의미와 주제들을 찾아 추론해 내는 방법으로, 이를 통해 대량의 문서들에서 출현할 확률이 높은 키워드를 도출하여 해당 토픽의 의미와 주제를 찾는 확률 모형이다[8]. 단어들의 집합들로 이뤄진 문서에는 토픽의 집합이 확률

적으로 내포되고, 단어들 간의 분포에 대한 분석을 통해 구조를 파악하여 문서에서의 주제를 파악한다. 이를 통해 문서에서 주요하게 다뤄지고 있는 토픽을 추정하고 확인 할 수 있다. 이처럼 문서의 토픽을 분석하는 주요 기법으로는 LSI (LatentSemantic Indexing), pLSA(Probabilistic Latent Semantic Analysis) 등이 주로 활용 된다 [9][10]. 최근에는 문서 내 주제별 단어의 확률을 기반으로 한 LDA(Latent Dirichlet Allocation) 기법을 활용해 기존 토픽모델링의 단점을 보완한 해당 방법론을 통한 연구가 활발히 이뤄지고 있으며, LDA 알고리즘 모델링 방법은 분석에 필요한 문서 내 다양한 형태의 주제들이 포함되어 있다는 가정 하에, 주제별 분포와 주제별로 생성된 단어의 확률을 모델링 하고 이를 통해 토픽의 확률을 추론하고 도출한다. 이는 관찰 되어지는 변수(observed variable)와 문서 내 감춰져 있는 주제 즉, 보이지 않는 변수(hidden variable)를 찾고 추론해준다[11][12]. 이를 통해 전체 문서들의 집합에서 주제 도출, 문서 내 주제의 비율, 주제에 포함되어있는 단어의 확률을 알아낸다. LDA 알고리즘은 아래 (그림 2) 과 같다[8].




(그림 2) LDA 알고리즘의 문서 생성 과정

α : 문서별 토픽 k의 Dirichlet prior weight, Θ값을 결정하는 파라미터

Θd : 문서별로 포함된 토픽의 비율

Zd,n: 문서 d의 n번째 단어의 토픽(index)

Wd,n: 문서 d의 n번째 단어, 문서에 관측되는 변수, index

η : 토픽별 단어 w의 Dirichlet prior weight, β값을 결정하는 파라미터

βk : 토픽별 단어 w의 생성확률

K : 전체 토픽의 개수

LDA 알고리즘을 통해 관찰되는 변수는 단어(Wd,n)이며, hyper parameter에 해당하는 α,η와 문서 내 단어 추출에 사용되어지는 hidden parameter인 β가 있다. 최종적으로 문서에서 관찰되어지지 않는 z인 hidden variable로 이뤄진다. 해당 모델에서 Z는 각 문서별 주제에 해당되는 토픽의 비율인 Θ로부터 만들어지고 Θ는 α값에 의해 형태가 정해지는 값으로 Dirichlet 분포를 따른다. 이와 함께 문서를 대표 토픽별 단어의 생성확률인 β는 Dirichlet 분포를 따르는 값이며 η 값에 의해 결정되어진다. 이를 통해 최종 단어인 W가 단어의 토픽 값인 Z와 토픽별 단어들 간의 비율인 β값에 의해 결정된다.

## 4. 개인정보 위협 트렌드 분석

### 4.1 개인정보위협 토픽 추출 및 정의

개인정보 위협 기반의 토픽을 추출하기 위해 전 처리과정을 마친 해당 데이터베이스에서 LDA 알고리즘을 활용하여 토픽을 추출 하였으며, 사용된 Tool 로는 R 패키지의 “topicmodels”을 활용하였다. 또한 토픽 모델링 시 최적의 토픽 추출 개수 및 샘플링 반복 횟수는 효율적 해석이 가능한 수준에서 연구자가 결정할 수 있다[13][14]. 이에 본 연구에서는 토픽의 수를 10, 15, 20, 25의 점진적 증감과 샘플링 반복 획수를 500, 1000, 2000, 5000으로 설정을 변경하며 토픽모델링을 수행하였으며, 이를 바탕으로 효율적으로 해석 가능하다고 판단되는 15개 토픽 및 샘플링 회수 50,000회로 설정하여 분석 하였다. 이와 함께 세부 설정값 중 α파라미터와 β값은 기본값 형태의 알고리즘을 적용 하였다. LDA 알고리즘을 통해 추출된 15개의 토픽의 경우 해당 토픽 내 확률 값이 높게 나타난 단어로 이뤄져 있다. 세부적으로 도출된 토픽 내용의 경우 개인정보취급 기관, 개인정보보호 취급자 위반, 개인정보 유출 및 오·남용에 따른 주요 위협 등 개인정보보호에 있어 위협이 되는 주요 토픽을 포함하고 있다. LDA 알고리즘을 통해 추출된 15개의 토픽과 해당 토픽 내 확률 값이 높게 포함된 단어로 이뤄져 있다.

<표3> 개인정보 위협기반 토픽 및 세부내용

| 업종/위반 | [P1] 정보통신-마케팅 불법활용 | [P2] 정보통신-수집/활용미동의 | [P3] 정보통신-안전성 확보조치 | [P4] 정보통신-개인정보제공/활용동의 | [P5] 정보통신-위치정보 | [P6] 정보통신-처리위탁위반 | [P7] 공공기관-개인정보 무단열람/조회 |
|---|---|---|---|---|---|---|---|
| | 마케팅활용, 쇼핑몰,광고대행사 불법매매, 삼자제공동의및고지, 포탈사, 동의통보사항 목적외이용,텔레마케팅, 데이터베이스, 대행업체 수집이용고지, 고유식별정보 | 미동의수집활용, 통신사, 목적외이용, 수집이용고지, 동의통보사항, 신용정보, 취급방침위반, 처리위탁위반, 불법매매, 협력업체, 보유기간, 텔레마케팅,이메일 | 포탈사, 통신사, 외부노출방지조치, 암호화, 비밀번호, 홈페이지, 접속기록보관관리, 접근권한, 기술적보호조치, 관리기술조치미흡, 관리적보호조치, 파기위반, 접근통제 | 통신사, 유출사금전피해, 삼자제공동의및고지, 미동의수집활용, 텔레마케팅, 처리위탁위반, 불법도용, 마케팅활용, 취급방침위반, 목적외이용, 협력업체,파기위반, 활용동의 | 포탈사해외, 위치정보, 파기위반, 미동의수집활용, 정보통신기업, 민감정보, 접속기록보관관리, 취급방침위반,비밀번호, 주민등록번호,계획수립및시행,접근권한, 암호화 | 통신사,유출사금전피해, 처리위탁위반,무단조회열람,소규모기업,협력업체, 마동의수집활용, 텔레마케팅,삼자제공동의및고지, 취급방침위반,목적외이용, 불법매매, 개인정보처리시스템 | 무단조회열람, 공무원, 공공기관,목적외이용, 접속기록보관관리, 보안교육및인식, 접근권한, 무단반출, 개인정보처리시스템, 삼자제공동의및고지, 보유기간, 마동의수집활용, 불법매매 |
| [P8] 공공기관-안전성 성확보조치 | [P9] 공공기관개인정보 제공/활용동의 | [P10] 금융기관-안전성 확보조치 | [P11] 금융기관-마케팅 불법활용 | [P12] 금융기관-개인정보제공/활용동의 | [P13] 금융기관-개인정보무단열람/조회 | [P14] 금융기관-개인정보불법거래 | [P15] 유통서비스-마케팅불법활용 |
| 공무원,공공기관, 접속기록보관관리, 목적외이용, 서류저장매체보관,접근권한,보안교육및인식,cctv, 마동의수집활용, 삼자제공동의및고지,보유기간,개인정보처리시스템, 계획수립및시행 | 공무원,공공기관, 삼자제공동의및고지,접속기록보관관리,마케팅활용, 목적외이용, 서류저장매체보관, 미동의수집활용, 보안교육및인식, 파기위반, 협력업체 보유기간, 수집이용고지 | 금융사, 카드사, 유출사금전피해, 협력업체, 외주직원, 계열사공유, 계획수립및시행, 처리위탁위반, 삼자제공동의및고지, 물리적보호조치, 마케팅활용, 접근권한, 서류저장매체보관 | 카드사, 유출사금전피해, 금융사, 협력업체, 마케팅활용, 이차피해, 목적외이용, 마동의수집활용, 텔레마케팅, 불법매매 과태료, 광고대행사, 수집이용고지 | 금융사, 카드사, 유출사금전피해, 삼자제공동의및고지, 불법매매, 미동의수집활용, 파기위반,협력업체, 목적외이용, 계열사공유,신용카드정보, 접근권한, 마케팅활용 | 금융사,무단조회열람, 보험사, 미동의수집활용, 삼자제공동의및고지,,카드사,접근권한,목적외이용, 불법매매, 외부노출방지조치, 동의통보사항,접속기록보관관리,개인정보처리시스템 | 금융사, 카드사, 유출사금전피해, 불법매매, 삼자제공동의및고지, 미동의수집활용, 마케팅활용, 접근권한, 목적외이용, 파기위반, 처리위탁위반, 스팸메일 광고대행사 | 마케팅활용, 유통, 삼자제공동의및고지, 광고대행사,쇼핑몰, 내부직원, 마동의수집활용, 목적외이용, 불법매매, 대행업체, 수집이용고지, 동의통보사항, 협력업체 |

세부적으로 도출된 토픽 내용의 경우 개인정보 취급 기관, 개인정보보호 취급자 위반, 개인정보 유출 및 오·남용에 따른 주요 위협 등 개인정보보호에 있어 위협이 되는 주요 토픽을 포함하고 있다. 주요 토픽은 <표 3>과 같다.

## 4.2 개인정보위협 토픽 분석 결과

개인정보관련 주요 위협에 관한 토픽을 도출하기 위해 데이터 전체 기간 동안의 토픽에 대한 비중을 분석하였다. 전체기간 동안의 토픽 점유율의 경우 [P12] 업종: 금융기관, 위반사항: 개인정보 제공·활용 동의, [P10] 업종: 금융기관, 위반사항: 안전성확보조치 토픽이 높은 비중을 나타냈으며 [P1] 업종: 정보통신 기업, 위반사항: 마케팅 불법활용, 등의 순으로 낮음 비중을 보였다. 해당 결과는 아래 <표 4>와 같다.

이와 함께 기간별 토픽의 시계열 분석을 위해 전체 데이터 기간을 3년간의 단위로 분리하고 각 기간별 비율(점유율)을 기간별 평균값으로 환산해 순위를 도출 하였다. 첫 번째로 1기간 (2016년 ~ 2013년)의 경우 금융기관인 카드3사의 개인정보 유출 사고 등에 따른 개인정보의 거래 관련 이슈

<표 4> 전체기간 개인정보보호 위협 토픽 비중

| 구분 | 업종 | 위반사항 | 비율 | 순위 |
|---|---|---|---|---|
| P1 | 정보통신 | 마케팅 불법활용 | 2.73% | 15 |
| P2 | 정보통신 | 수집/활용 미동의 | 4.56% | 11 |
| P3 | 정보통신 | 안전성확보조치 | 8.16% | 4 |
| P4 | 정보통신 | 개인정보 제공/활용 동의 | 5.63% | 8 |
| P5 | 정보통신 | 위치정보 | 3.68% | 14 |
| P6 | 정보통신 | 처리위탁위반 | 3.71% | 13 |
| P7 | 공공기관 | 개인정보 무단 열람/조회 | 5.47% | 10 |
| P8 | 공공기관 | 안전성확보조치 | 4.21% | 12 |
| P9 | 공공기관 | 개인정보 제공/활용 동의 | 5.94% | 7 |
| P10 | 금융기관 | 안전성확보조치 | 11.20% | 2 |
| P11 | 금융기관 | 마케팅 불법 활용 | 5.60% | 9 |
| P12 | 금융기관 | 개인정보 제공/활용 동의 | 15.16% | 1 |
| P13 | 금융기관 | 개인정보 무단 열람/조회 | 7.12% | 5 |
| P14 | 금융기관 | 개인정보 불법 거래 | 10.07% | 3 |
| P15 | 유통서비스 | 마케팅 불법 활용 | 6.76% | 6 |
| 총합 | | | 100% | |

토픽 및 개인정보의 안전성 확보를 위한 기술적·관리적 안전성 확보 조치 기준 등에 관한 이슈가 많이 대두되었으며, 관련해 [P14]업종:금융기간, 위반사항: 개인정보 불법 거래, [P10] 업종: 금융기관, 위반사항: 안전성확보 조치 순으로 높게 나타났다. 두 번째로 2기간(2014년 ~ 2011년)에서는 기업에서 준수해야할 개인정보 안전성 확보를 위한 기술적·관리적 보호조치 강화와 개인정보처리자의 보안의식 미성숙에 따른 개인정보 불법열람 등에 관한 이슈로 [P3] 업종: 정보통신 기업, 위반사항: 안전성확보조치, [P3] 업종: 정보통신 기업, 위반사항: 수집/활용 미동의 순으로 높게 나타났다. 마지막으로 3기간(2010년 ~ 2008년)의 경우 공공기관 및 유통서비스 기업등 정보보안 관련 수준이 상대적으로 낮은 업종에 대한 위협이 증가하였으며, 이에 따라 [P6] 업종: 정보통신, 위반사항: 처리위탁위반, [P7] 업종: 공공기관, 위반사항: 개인정보 무단/열람/조회, 등에 관한 토픽의 비중이 상대적으로 높게 나타났음을 확인할 수 있다. 해당 결과는 아래 <표5>와 같다.

<표5> 기간별 비중

| Rank | 1기간(2016년 ~2014년) | | 2기간(2013년 ~2011년) | | 3기간(2010년 ~2008년) | |
|---|---|---|---|---|---|---|
| | 구분 | 비율 | 구분 | 비율 | 구분 | 비율 |
| 1 | P14 | 0.2022 | P3 | 0.3213 | P6 | 0.1419 |
| 2 | P10 | 0.1379 | P2 | 0.0934 | P7 | 0.1402 |
| 3 | P3 | 0.1379 | P7 | 0.0819 | P1 | 0.1274 |
| 4 | P12 | 0.0943 | P10 | 0.0729 | P3 | 0.1224 |
| 5 | P4 | 0.0816 | P13 | 0.0630 | P15 | 0.1205 |
| 6 | P13 | 0.0490 | P5 | 0.0627 | P4 | 0.0963 |
| 7 | P11 | 0.0488 | P6 | 0.0557 | P2 | 0.0771 |
| 8 | P8 | 0.0436 | P4 | 0.0449 | P8 | 0.0499 |
| 9 | P2 | 0.0421 | P14 | 0.0426 | P9 | 0.0284 |
| 10 | P5 | 0.0377 | P1 | 0.0423 | P12 | 0.0254 |
| 11 | P15 | 0.0368 | P9 | 0.0332 | P10 | 0.0221 |
| 12 | P1 | 0.0280 | P12 | 0.0274 | P14 | 0.0145 |
| 13 | P7 | 0.0267 | P8 | 0.0228 | P5 | 0.0138 |
| 14 | P6 | 0.0226 | P11 | 0.0189 | P13 | 0.0102 |
| 15 | P9 | 0.0109 | P15 | 0.0171 | P11 | 0.0099 |

## 4.3 Hot / Cold 개인정보 위협 분석

본 연구에서는 개인정보보호 위협에 대한 주요 토픽 및 기간별 분석과 토픽 내 세부 사항의 변화 분석을 진행 하였다, 하지만 해당 분석으로 시간에 따른 위협 변화를 가시적으로 파악할 수 있었지만, 개별 위협 수준에서의 시간 경과에 따른 중요도 증가 및 감소 추세를 분석하기에는 한계점이 존재한다. 이에 토픽에 대한 전체분석 기간 동안의 비중에 대한 추세를 분석하여 비중의 증가 추세를 보이는 개인정보 위협 Hot 토픽 및 감소 추세를 보이는 개인정보 위협 Cold 토픽에 대한 트렌드를 분석하였다. 각 토픽의 연도별 추세를 증가와 감소로 나눠서 판단하는 기준의 경우 통계분석 기법 중 선형회기분석을 진행하고 해당 결과 중 회귀계수 값을 이용하였다[15]. 분석데이터는 9년간의 연도별 토픽의 비중 평균값을 활용하였으며, 연도는 독립변수로, 개인정보 위협 토픽에 대한 비중값을 종속변수로 활용하였다. 이를 통해 15개의 개인정보 위협과 관련된 세부 위협 토픽에 대한 회귀계수 값을 산출하였다. 다음으로 분석된 토픽 중 유의수준이 5% 이내의 유의한 회귀계수를 가지는 토픽대상으로 추이를 분석하였다. 분석된 개인정보 위협 토픽 중 회귀계수 값이 양수의 형태를 가지면 Hot, 음수 형태를 나타내면 Cold로 분류하여 토픽의 추이를 도출하였다. 이를 통해 Hot 토픽 4개와 Cold 토픽 5개를 개인정보 위협에 따른 추세로 분석 되었으며, Hot 토픽의 경우 신용정보 및 고유식별 정보등에 대한 위협이 증가되고 있는 금융기관 대상으로 한 위협 토픽이 상대적으로 크게 증가 하였으며, 개인정보에 대한 금전적 가치 증가에 따른 불법 거래 및 마케팅 활용과 개인정보보호를 위한 안전성확보 조치에 대한 중요도가 꾸준히 증가하는 것으로 분석된다. 이와 반대로 Cold 토픽의 경우 공공기관 및 정보통신기업 관련 토픽이 하향 추세를 보이며, 위협부분의 경우 과거 개인정보보호 관련 규제 및 개인정보보호에 대한 인식부족에 따른 개인정보취급자를 통한 개인정보 오·남용 등의 침해 행위가 개인정보보호에 관한 교육 및 지식 습득 등의 인식 제고 활동을 통해 관련 토픽에 대한 노출 빈도가 낮아지는 것으로 판단할 수 있다. 해당 내용은 <표6> 과 같다.

<표 6> 회귀 계수와 유의 수준

| 구분 | 비율 | 순위 | Hot/cold |
|---|---|---|---|
| P1 | -5.07544 | 0.0014 | Clod |
| P2 | -1.18006 | 0.2765 | - |
| P3 | -0.15461 | 0.8815 | - |
| P4 | -0.75294 | 0.4760 | - |
| P5 | 1.25697 | 0.2491 | - |
| P6 | -5.33846 | 0.0011 | Cold |
| P7 | -8.52651 | 0.0001 | Cold |
| P8 | -0.52951 | 0.6128 | - |
| P9 | -1.46810 | 0.1855 | Cold |
| P10 | 6.02575 | 0.0005 | Hot |
| P11 | 3.18530 | 0.0154 | Hot |
| P12 | 4.72799 | 0.0021 | Hot |
| P13 | 1.54377 | 0.1666 | - |
| P14 | 4.51416 | 0.0028 | HoT |
| P15 | -3.07647 | 0.0179 | Cold |

## 5. 결론

본 연구는 개인정보 침해사고 및 오·남용 등 위협을 기반으로 한 신문기사 데이터를 활용해 개인정보처리 기관별 개인정보 영역 관련 주요 위협 도출하고 시계열 분석을 통해 기간별 토픽 변화 추이를 분석 하였다. 그동안 광의의 정보보안 분야에 국한된 기술 예측 위주의 트렌드 분석 연구가 진행되고 있고, 최근 위협이 크게 증가하고 있는 개인정보보호 분야에 대한 연구의 필요성이 증가되고 있다. 이에 본 연구를 통해 개인정보 위협 동향과 트렌드 분석 등 개인정보보호 분야의 의미 있는 정보를 확인하고 이를 활용해 개인정보보호와 관련된 현재의 트렌드를 미리 점검하고 개인정보보호를 위한 주요 이슈 및 관련 문제점을 짚어보고 그 해결방안을 찾기 위한 주요 단초를 제공하여 관련 개인정보처리기관에서 활용 가능한 모델을 제시하였다.

세부적으로 살펴보면 정보통신 기업, 금융 기업, 공공기관, 유통서비스 기업에 업종별 개인정보 취급 기관에 대한 토픽을 도출하고 각 업종별 개인정보 무단/열람 조회, 개인정보보호 안전성 확보조치, 개인정보 불법 거래 등의 주요 토픽 및 세부 항목을 도출하였다.

이와 함께 기간별 Hot 및 Cold 토픽을 분석한 결과 전반적인 업종별 Hot 토픽은 개인정보를 포함한 주요 신용정보를 취급하는 금융기업, 유출정보에 대한 불법 거래, 마케팅 활동 등 금전적 거래에 대한 토픽, 개인정보를 안전하게 보관하고 관리하기 위한 안전성확보 조치에 대한 토픽들이 상승되고 있다. 이를 종합하면 금융기관을 대상으로 한 대외 및 내부 취급자에 의한 개인정보 오·남용에 대한 위협이 크게 증가하고 있으며 해당 영역의 보호조치 강화와 꾸준한 관리·감독이 필요할 것으로 보인다. 이와 반대로 Cold 형태를 토픽의 경우 업종별로는 공공기관 및 정보통신 기업 관련 토픽이 하향 추세를 보이며, 위협부분의 경우 과거 개인정보보호 관련 규제 및 개인정보보호에 대한 인식부족에 따른 개인정보 오·남용 등의 행위가 개인정보보호에 관한 인식 제고 활동을 통해 관련 토픽에 대한 노출 빈도가 낮아지는 것으로 판단 할 수 있다.

이처럼 개인정보 위협 트렌드에 대한 시계열 분석 자료를 통해 급변하는 개인정보 위협에 대한 현황을 진단하고 그 추세를 파악하여 향후 집중 관리해야 될 영역을 찾고 관리할 수 있는 트렌드 정보를 제공하는 것에 의의가 있다. 하지만 본 연구의 경우 분석에서 활용된 데이터가 신문기사를 기준으로 분석하여, 다양한 문헌, 논문, 대외 보고서등의 추가 데이터 확보와 분석이 필요할 것으로 보이며, 이와 함께 분석 자료에 대한 가시성 확보를 위한 시각화와 관련 세부 토픽들 간의 연계 구조를 분석하여 위협 토픽 간의 상호 관계성 분석 등의 추가 연구 필요할 것으로 보인다.

## 참고문헌

[1] 김원필. (2015). “정보보안에 대한 연구 트렌드 분석”, 한국정보통신학회논문지, 제19권, 제5호, pp. 1110-1116.

[2] 정철우, 김재준. (2012). "텍스트마이닝을 활용한 건설분야 트랜드 분석", 한국디지털건축 인테리어학회논문집, 제12권 2호, pp53-60.

[3] 조수곤, 김성범. (2012). “텍스트마이닝을 활용한

산업공학 학술지의 논문 주제어간 연관관계 연구", 대한산업공학회지, 제 38권, 제1호, pp. 67-73.

[4] Galit, S., Nitin, R. P., & Peter, C. B. (2011). Data mining for business intelligence. New Jersey: Wiley. Greiner, MA, & Franza, RM (2003). Barriers and bridges for successful environmental technology transfer. Journal of Technology Transfer, 28, pp. 167-177.

[5] 최은혁. (2011). "악성코드 동향으로 살펴본 스마트 기기의 보안 위협", 정보보호학회지, 제21권, 제3호, pp. 7-11.

[6] 안홍기. (2007). "산업기밀 정보유출방지와 개인정보보호의 현황과 전망", 정보과학회지, 제25권, 제8호, pp. 42-47.

[7] 남기효, 박상중, 강형석, 남기환, 김성인. (2008) "개인정보보호기술의 최신 동향과 향후 전망", 정보보호학회지, 제18호, 제6호, pp. 11-19.

[8] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), pp. 77-84.

[9] Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, & R. A.Harshman. (1990). "Indexing by Latent Semantic Analysis", Journal of the AmericanSociety for Information Science (JASIS), Vol. 41, No. 6, pp. 391-407.

[10] Hofmann, T. (1999). "Probabilistic latent semantic indexing", Proceedings of the 22ndannual international ACM SIGIR conference on Research and development ininformation retrieval, 50-57.

[11] 이원상, 손소영. (2015). "공간빅데이터 연구 동향 파악을 위한 토픽모형 분석", 대한산업공학회지, 제41권, 제1회, pp. 64-73.

[12] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003), Latent dirichlet allocation, Journal ofMachine Learning Research, 3, 993-1022.

[13] Hornik, K., Grün, B. (2011). topicmodels : An R package for fitting topic models. Journal of Statistical Software, Vol. 40, No. 13, 1-30.

[14] Song, M., Kim, S. Y. (2013). Detecting the knowledge structure of bioinformatics by mining-full-text collections. Scientometrics, 96(1), 183-201.

[15] Griffiths, T., & Steyvers, M. (2004). Finding Scientific Topics. Proceedings of the National Academy of Sciences, 101 (suppl. 1), 5228-5235.

[16] 김영희, 국광호 (2014) "개인정보의 안전성 확보조치 기준에서의 우선순위 정립에 관한 연구 ", 융합보안논문지, 제10권 제4호, pp. 9-17.

[17] 이택현, 국광호 (2018) "데이터 마이닝 기법을 이용한 소규모 악성코드 탐지에 관한 연구" , 융합보안논문지, 제19권 제1호, pp. 11-17.

[18] 김종민, 정병수 (2016) "데이터마이닝을 이용한 DDoS 예측 모델링" , 융합보안논문지, 제16권 제2호, pp. 63-70.

## 〔저자소개〕

김 영 희 (Young-Hee Kim)
2014년 서울과학기술대 산업정보시스템 석사
2018년 서울과학기술대 산업정보시스템 박사
2012년 ~ 현재 한화시스템 ICT
email : sorak75@naver.com

이 택 현 (Taek-Hyun Lee)
2013년 서울과학기술대 산업정보시스템 석사
2018년 ~ 현재 서울과학기술대 산업정보시스템 박사과정
email : futp@naver.com

김 종 명 (Jong-Myoung Kim)
2014년 서울과학기술대 산업정보시스템 석사
2014년 ~ 현재 서울과학기술대 산업정보시스템 박사과정
email : solojong@nate.com

박 원 형 (Won-Hyung Park)
2002년 서울과학기술대 산업정보시스템 학사
2005년 서울과학기술대 정보산업공학과 석사
2009년 경기대학교 정보보호학과 박사
2012 ~ 현재 극동대학교 산업보안학과 교수/학과장
email : whpark@kdu.ac.kr

국 광 호 (Kwang-Ho Kook)
1979년 서울대학교 산업공학과 학사
1981년 서울대학교 산업공학과 석사
1989년 조지아공대 산업공학과 박사
1993년 ~ 현재 서울과학기술대학교 글로벌융합산업공학과 교수
email : khkook@seoultech.ac.kr