The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.2.301
32(2), 301–317

Analysis of the National Police Agency business trends using text mining

Hyunseok Sun[a] · Changwon Lim[a,1]

[a]Department of Applied Statistics, Chung-Ang University

(Received October 15, 2018; Revised November 21, 2018; Accepted January 31, 2019)

Abstract

There has been significant research conducted on how to discover various insights through text data using statistical techniques. In this study we analyzed text data produced by the Korean National Police Agency to identify trends in the work by year and compare work characteristics among local authorities by identifying distinctive keywords in documents produced by each local authority. A preprocessing according to the characteristics of each data was conducted and the frequency of words for each document was calculated in order to draw a meaningful conclusion. The simple term frequency shown in the document is difficult to describe the characteristics of the keywords; therefore, the frequency for each term was newly calculated using the term frequency-inverse document frequency weights. The L2 norm normalization technique was used to compare the frequency of words. The analysis can be used as basic data that can be newly for future police work improvement policies and as a method to improve the efficiency of the police service that also help identify a demand for improvements in indoor work.

Keywords: text-mining, unstructured format, the Korean National Police Agency, keyword extraction

1. 서론

디지털 데이터 수집 기술의 빠른 진보로 인해 웹 상에서는 텍스트, 이미지, 영상 데이터 등 다양한 비정형 데이터가 생산되고 있다. 다양한 정보의 80%가 정형화되지 않은 형태이며 이는 주로 텍스트에서 비롯된다는 표현을 할 정도로 텍스트 데이터의 생산은 끊임없이 이루어지고 있다 (Grimes, 2008). 이에 따라 방대한 양의 텍스트 자료를 처리하고 분석할 수 있는 적절한 패턴과 추세를 발견하는 것이 큰 이슈이다 (Talib 등, 2016). 이러한 비정형 텍스트 데이터의 대량 생산과 이를 통해 인사이트를 발굴하고 의사결정에 도움이 되는 기법으로 각광 받고 있는 분야가 텍스트 마이닝이다 (Sulova 등, 2017). 텍스트 마이닝은 정형화되지 않은 텍스트 형태에서 이를 구조화된 데이터로 변환하고 이를 통해 정보를 추출하는 방법이다 (Nahm과 Mooney, 2002). 정보를 추출하기 위한 방법으로는 문서의 군집화 및 분류, 정보 추출과 정보 검색, 트렌드 탐지 등이 있고 이와 같은 기법들은 다양한 분야에서 활용되고 있다 (Berry, 2004).

This research was supported by the Chung-Ang University Research Scholarship Grants in 2017.
[1]Corresponding author: Department of Applied Statistics, Chung-Ang University, Heukseok-ro 84, Dongjak-gu, Seoul 06974, Korea. E-mail: clim@cau.ac.kr

정보 처리에 대한 수요가 늘고 텍스트 마이닝을 통해 대용량 문서 자료에 대한 활용이 가능해지면서 다양한 분야에서 텍스트 마이닝을 활용한 연구가 진행되고 있다. 국내에서는 의학분야, 사회분야, 경제분야, 공학분야 등에서 텍스트 마이닝을 활용한 연구들과 새로운 알고리즘들이 개발되고 있다. 대표적으로 Bae 등 (2013)은 텍스트 마이닝을 이용하여 2012년 한국 대선 관련 트위터 자료를 분석함으로써 트위터 사용자들의 멘션 기반 네트워크를 구축하고 사회적 이슈를 포착하기 위한 실시간 트위터 트렌드 마이닝 시스템을 개발하였다. Cho와 Kim (2011)은 텍스트 마이닝을 활용한 산업공학 학술지의 논문 주제어 간 연관관계를 파악하기 위해 텍스트 자료 시각화, 군집분석, 연관성 분석 방안을 연구하였다. Song 등 (2013)은 경제 분야에서의 텍스트 자료를 통해 한국 경제학에 대한 시각과 연구 동향을 파악하였다. 이처럼 분야를 막론하고 다양한 곳에서 텍스트 자료를 활용한 연구가 나타나고 있으며, 본 논문에서는 경찰청 업무 향상 및 개선에 활용하기 위한 인사이트 발굴을 목표로 경찰청에서 생산한 문서들에 텍스트 마이닝 기법을 적용하는 방법론과 분석 결과를 설명한다.

본 논문에서 사용한 자료는 경찰청이 발행한 경찰백서와 우리나라 16개의 지방경찰청에서 약 1년 동안 생산된 업무 보고 문서이다. 우리는 이 문서자료에 텍스트 마이닝 기법을 적용하여 각 자료에 나타난 키워드를 파악하고 키워드의 연도별 변화를 파악하는 분석을 수행하였다. 또한 지방청마다 고유한 업무 특징을 비교 분석하여 최종적으로는 경찰청의 업무 향상 및 청내 업무 수요 파악에 활용될 수 있도록 하였다. 또한 본 연구와 유사한 연구를 바탕으로 앞으로 경찰청 내에서도 텍스트 관련 연구가 활발하게 이루어질 수 있도록 하고자 하는 것이 본 논문의 목표이다.

본 논문의 나머지 부분은 다음과 같이 구성되어 있다. 2장에서는 본 연구에서 사용한 자료에 대한 설명과 분석 절차와 방법론에 대해 설명하였다. 3장에서는 텍스트 마이닝 기법과 다양한 방법론을 토대로 각 자료를 요약하고 트렌드를 탐지한 결과에 대해 서술하였다. 마지막 4장에서는 본 연구의 의의와 앞으로의 활용가능성에 대해 논하였다.

2. 자료 설명 및 분석 방법

2.1. 자료 설명

경찰청에서 생산한 문서들 중에서 본 연구에서 사용된 자료는 경찰백서와 지방청 업무보고 자료이다. 우선 경찰백서 자료는 2010년부터 2016년까지 총 7개년의 경찰청 업무 활동을 기록한 텍스트 자료이다. 경찰백서는 경찰청에서 역점을 두고 추진해 온 업무와 활동 상황을 충실히 기술하여 각 분야별로 변화되어 가는 경찰의 활동을 기록해 놓은 연간 문서이다. 연도마다 목차의 이름은 조금씩 다르지만 매년 총 7가지의 반복되는 주제를 기반으로 작성되어있다. 이를 통해 각 주제별로 나타나는 단어들을 파악하고 연도별로 각 주제에서 나타나는 단어 트렌드를 파악하였다.

두 번째로는 각 지방청에서 생산하는 일일업무보고, 주간업무보고 등의 업무 보고 문서를 이용하여 분석에 진행하였다. 16개의 지방청(인천지방청 제외)에서 생산하는 문서로 자료 기간은 2016년 5월부터 2017년 5월까지 약 1년간의 업무 활동을 기록해놓았다. 기록된 업무 내용은 지방청 중요일정, 당면추진업무, 근무현황 및 부서별 활동 내용 등이 있다. 해당 자료를 통해 지방청별로 생산한 문서에서 나타나는 특징 키워드를 추출하였다.

분석에서 사용한 데이터의 구성은 Table 2.1과 같다. 경찰백서 텍스트 자료는 7개의 주제로 구분하여 분석을 진행하였고, 지방청 업무보고 텍스트 자료는 각 데이터를 생산한 16개의 지방청별로 구분하여 분석을 진행하였다.

Table 2.1. Configuration of text data used for National Police Agency business analysis

자료 종류	자료 구성	자료 기간
경찰백서 (연도별 자료)	경찰행정 - 경찰 조직, 인력 등 경찰활동 추진배경 및 성과 교통안전과 경찰활동 국민생활과 경찰활동 범죄 추세 및 주요 범죄 검거 사회 안보와 경찰활동 세계화 시대 속의 경찰활동	2010.01.01–2016.12.31
지방청 업무보고 (일별 자료)	강원지방청	2016.06.15–2017.05.16
	경기남부지방청	2016.05.15–2017.05.16
	경기북부지방청	2016.05.17–2017.05.16
	경남지방청	2016.05.15–2017.05.16
	경북지방청	2016.05.18–2017.05.15
	광주지방청	2016.05.17–2017.05.16
	대구지방청	2017.03.24–2017.05.15
	대전지방청	2016.05.17–2017.05.15
	부산지방청	2016.05.17–2017.05.15
	서울지방청	2016.05.17–2017.05.16
	울산지방청	2016.05.17–2017.05.16
	전남지방청	2016.05.17–2017.05.16
	전북지방청	2016.05.17–2017.05.16
	제주지방청	2017.01.09–2017.05.16
	충남지방청	2016.05.17–2017.05.16
	충북지방청	2016.05.18–2017.05.16

2.2. 분석 수행 절차

데이터베이스화된 자료를 파이썬 프로그램 (Python Softwoare Foundation, 2017)의 내장 함수들을 바탕으로 한글 형태소 분석에 사용되는 라이브러리인 KoNLPy (Park과 Cho, 2014), 기계 학습 오픈 소스 라이브러리인 Scikit-learn (Pedregosa 등, 2011) 등을 사용하여 분석을 수행하였다. 텍스트 마이닝에서 의미 있는 결론을 도출하고자 각각의 데이터에 대해 전처리 과정을 수행하였고, KoNLPy 라이브러리에 내장된 트위터 형태소 분석기를 사용하여 문서에 나타난 단어들을 추출하였다. 트위터 형태소 분석기는 다른 오픈 소스로 제공되는 한글 형태소 분석기에 비해 처리 속도가 빠르고 분석된 결과를 살펴보았을 때 비교적 정확하고 일관된 형태로 단어가 추출되었기 때문에 분석에 사용되었다. 형태소 분석 과정 후, Scikit-learn 라이브러리에 내장된 함수들을 통해 Bag-of-words 표현 방식을 기반으로 문서-용어 행렬을 구축하고 단어빈도-역문서빈도(term frequency-inverse document frequency; TF-IDF) 가중치를 계산하여 데이터를 변환하였다. 변환된 데이터를 통해 해당 문서를 특징짓는 키워드를 추출하였다. 또한 연도별 또는 월별 자료의 경우 단순선형회귀분석을 통해 증가추세나 감소추세가 나타난 단어들을 추출하여 경찰청 업무 트렌드를 분석하였다. 이때, 데이터마다 문서의 길이, 즉 전체 단어의 수가 다르기 때문에 문서별 단어 출현 정도를 비교하기 위하여 L2 정규화(L2 normalization)를 적용하였다.

이외에도 텍스트 자료에 나타난 상위 단어들을 한눈에 파악할 수 있는 방법인 워드클라우드를 그려 전체적인 문서의 특징을 살펴보고, 회귀분석을 통해 파악한 증가추세와 감소추세 단어의 연도별 그래프를 통해 시각적인 분석을 수행하였다. Figure 2.1은 분석 수행 절차를 도식화한 형태이다.

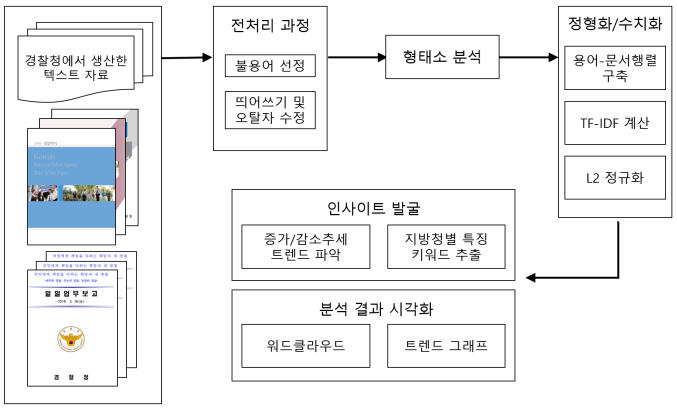


Figure 2.1. Text mining procedure for National Police Agency business analysis.

2.3. 전처리 과정

비정형 자료인 텍스트를 통해 의미 있는 결론을 도출하기 위해서는 텍스트에 반복적으로 나타난 의미 없는 단어를 파악하고 이를 분석에서 제외하고 이후의 과정을 진행하는 것이 필요하다. 원자료의 경우 페이지마다 반복적으로 나타난 문구, 예를 들어 경찰백서 데이터의 경우 목차 제목이 매 페이지마다 기술되어 전처리 과정 없이 분석을 진행할 경우 목차에 나타난 단어들이 상위 빈도를 차지하게 된다. 이는 텍스트 자료를 통해 경찰청의 업무 내용을 파악하는데 방해가 되는 요소로 판단되기 때문에 각 페이지마다 업무 내용과 별개로 반복적으로 나타난 문구들을 제거하는 작업이 필요하다. 지방청 업무 보고 자료의 경우 직책에 대한 언급이 많아 직책에 포함된 단어들이 상위 빈도를 차지하는 문제가 발생하여 직책을 뜻하는 단어들을 제거하는 전처리 과정을 진행하였다. 예를 들어 해당업무 담당자가 “폭력계장”인 경우 형태소 분석을 거치면 “폭력”과 “계장”으로 나오기 때문에 텍스트에 대한 특징 단어로 “폭력”이라는 잘못된 결과를 도출하게 된다. 따라서 직책을 나타나는 단어를 분석에서 제외하기 위해 “∼계장”, “∼팀장”, “∼관리관” 등의 단어를 뽑아내고 원자료에서 이 단어들을 삭제한 후 업무 내용에서 나온 단어만 분석에 반영될 수 있도록 전처리를수행하였다.

반복적으로 나타나거나 텍스트에 대한 분석 내용을 방해할 수 있는 단어들의 패턴을 파악하고 원자료에서 제외하는 과정을 거친 뒤 형태소 분석된 결과를 통해 불용어(stopword)를 선정하였다. 트위터 형태소 분석기를 이용하여 조사를 제외하고 명사, 동사, 형용사로 태깅된 단어를 추출하였고 이때 명사로 잘못 추출된 조사들을 불용어 처리하였다. 예를 들어, “도록”, “하기에” 등 잘못 추출된 단어들을 파악하고 이와 같은 단어들은 텍스트의 의미를 파악하는데 도움이 되지 않기 때문에 분석에서 제외시켰다.

2.4. 데이터 정형화 및 수치화

2.4.1. Bag-of-words 표현 본 연구에서는 분석에 사용되는 텍스트 데이터를 수치화하는 방법으로 Bag-of-words 방식을 이용하여 분석을 진행하였다. Bag-of-words 개념은 문서나 문장에 나타난 단어들의 출현빈도를 기반으로 단어에 대한 고유한 인덱스를 통해 해당 문서를 전체 단어 크기의 벡터 표현하는 방식이다 (Turney와 Pantel, 2010). Figure 2.2는 Bag-of-words 표현 방식의 예로, 형태소 분석 후 각 단어의 고유 인덱스를 기반으로 문장에서 단어가 나타나면 해당 인덱스에 +1을 하여 각각의 문장을 전체 단어 수 길이만큼의 벡터로 표현하는 과정이다. 이 표현 방법을 기반으로 용어-문서 행렬(term-

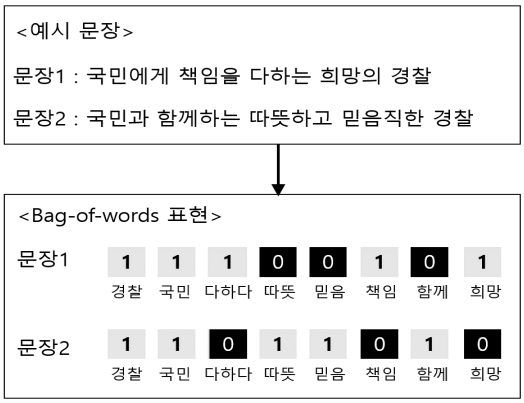


Figure 2.2. Example of Bag-of-words vector representation on text.

document matrix)를 구축하여 각 문서의 벡터 값이 그 문서에 출현한 단어의 빈도로 구성되게 데이터를 정형화 하였다.

2.4.2. 단어빈도-역문서빈도 단어빈도-역문서빈도인 TF-IDF는 여러 문서에서 특정 단어가 해당 문서 내에서 얼마나 중요한 것인지를 나타낼 수 있도록 고안된 방법이다 (Ramos, 2003). 단어 빈도는 해당 단어가 특정 문서 내에서 얼마나 자주 등장하는지에 대한 값을 의미하고 역문서 빈도는 전체 문서에서 단어가 나타난 문서의 비율에 역수를 취한 값이다. 이 두 가지 값을 곱하여 나타낸 값으로 이 과정의 목적은 문서 내에서 상대적으로 더 중요한 단어를 파악하기 위함이다. 문서 내에서 단어의 중요도가 높다는 것은 해당 단어가 모든 문서에서 빈번하게 나타나는 것이 아니라 특정 문서에서만 등장하여 그 문서의 키워드로 판별될 수 있다는 것을 의미한다. 일반적인 TF-IDF의 수식은 식 (2.1)과 같다

$$\begin{aligned} \mathrm{tfidf}(t,d,D) &= \mathrm{tf}(t,d) \times \mathrm{idf}(t,D), \\ \mathrm{idf}(t,D) &= \log \frac{|D|}{|\{d \in D : t \in d\}|}, \end{aligned} \tag{2.1}$$

여기서 $\mathrm{tf}(t,d)$는 특정 문서 내에서 단어 t의 총 출현 빈도를 나타내는 값이고 $\mathrm{idf}(t,D)$는 전체 문서 집합 D내에서 단어 t를 포함한 문서의 수의 비율의 역수이다. 여기에 문서 수가 많아지게되면 되면 전체 문서 집합 D의 크기가 커짐에 따라 $\mathrm{idf}(t,D)$의 값이 기하급수적으로 커지는 것을 방지하기 위해 로그를 취한 형태가 일반적이다. 본 연구에서는 데이터 형태에 맞는 TF-IDF 가중치를 구하기 위에 일반적인 형태에서 변형된 식을 이용하였다. 지방청별 업무 보고 문서에서 문서 단위는 지방청이고 전체 문서 집합 크기는 16에 불과하기 때문에 역문서빈도인 $\mathrm{idf}(t,D)$에 로그를 취하지 않았다. 문서별 특징 단어를 추출하기 위해 상위 단어들의 출현빈도의 분포를 살펴보았다. Figure 2.3은 경찰청 업무 보고 자료에서 나타난 상위 300개 단어의 출현빈도를 나타낸 그래프이다. 이 그림에서 볼 수 있듯이 각 지방청에 나타나는 단어들의 빈도를 비교해보면 상위 30개 단어의 출현 빈도는 하위 단어들의 출현 빈도에 월등하게 높게 나타났다. 상위 단어들의 출현빈도가 나머지 단어들에 비해 상당히 크기 때문에 일반적인 TF-IDF 가중치를 사용하는 경우, 가중치를 적용하더라도 문서의 특징 단어가 상위에 분포하지 않고 가중치를 적용하기 전의 결과와 동일한 상위 단어가 나타났다. 따라서 단어의 출현 빈도인 에 로그를 취하여 가중치를 계산하였다. 위와 같은 문제를 해결하고자 본 논문에서는 해당 자료에서 문서별 특징 키워드 추출을

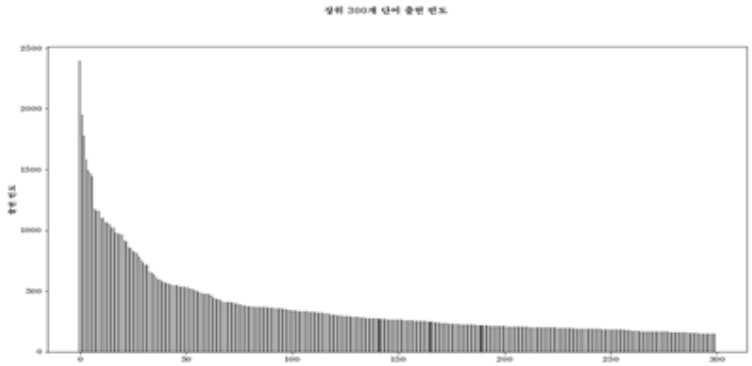


Figure 2.3. Histogram of the top 300 words in the National Police Agency's business report texts.

위한 기존 형태의 식을 변형하여 다음과 같은 새로운 식 (2.2)를 제안하였다:

$$\text{tfidf}(t,d,D) = \log(\text{tf}(t,d)+1) \times \frac{|D|}{|\{d \in D : t \in d\}|}. \tag{2.2}$$

2.4.3. L2 정규화 문서의 의미를 파악하는데 있어 문서의 길이가 서로 다른 경우 단순 출현빈도를 통해서는 한 단어에 대한 문서별 비교가 불가능하다. 따라서 단어에 대한 문서별 빈도 비교를 위해 TF-IDF로 계산된 값에 L2 정규화를 적용하여 분석을 수행하였다. L2 정규화는 벡터의 유클리디안 크기(Euclidean norm)가 1이 되도록 스케일을 조정하는 방법이다 (Kothe, 1983).

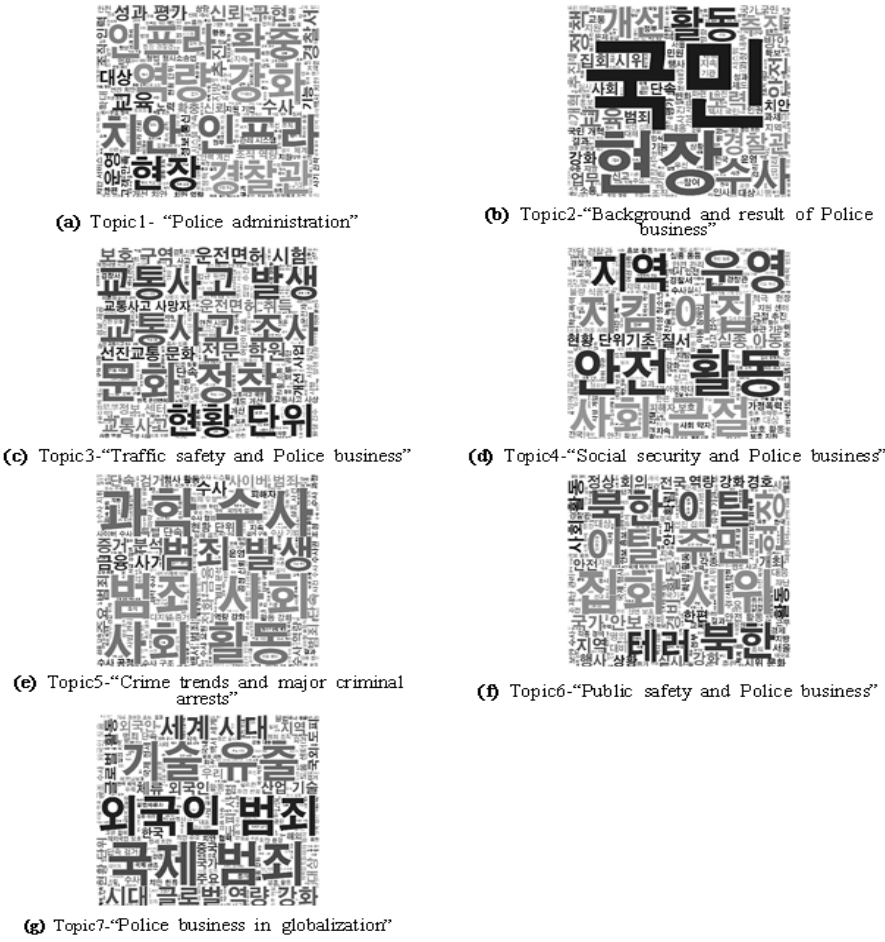
이렇게 스케일이 바뀐 문서 벡터는 문서의 길이, 즉 문서마다 나타난 단어의 수에 영향을 받지 않기 때문에 특정 단어에 대한 문서별 출현 정도 비교가 가능하다. 또한 L2 정규화를 사용하는 대표적인 이유로는 문서 벡터의 코사인 유사도 비교 시, 정규화된 벡터는 이미 크기가 1로 표준화되어있기 때문에 두 벡터의 내적만으로도 유사도를 측정할 수 있다는 장점이 있기 때문이다 (Leopold와 Kindermann, 2002). 본 논문에서는 문서 간 유사도 분석을 수행하지 않았지만 추후 연구에 사용될 수 있는 측면을 고려하여 L2 정규화를 수행하였다. 문서 벡터에 대한 정규화 과정을 자세히 살펴보면, 정규화 전의 문서 벡터는 전체 단어 개수에 의해 정의되고 각 원소의 값은 Bag-of-words 방식에 따라 단어의 TF-IDF 값으로 표현된다. 이렇게 표현된 벡터의 크기를 구하고 각 원소의 값을 벡터의 크기로 나누어주면, 각각의 값은 0과 1사이의 값을 갖게 되고 이 값의 의미는 해당 벡터에서 각 단어의 절대적인 수치가 아닌 상대적인 중요도를 의미하게 된다. 정규화 과정을 거친 뒤에는 각 문서 벡터의 크기가 모두 1로 동일하기 때문에 한 단어가 갖는 문서 내의 중요도를 비교할 수 있게 된다. TF-IDF 가중치가 반영된 문서 벡터 $\vec{\boldsymbol{v}}$가 L2 정규화 벡터 $\vec{\boldsymbol{u}}$가 되는 과정을 수식으로 표현하면 식 (2.3)과 같다.

$$\begin{aligned} &\vec{\boldsymbol{v}} = (v_1, v_2, \ldots, v_V), \quad V : \text{전체 단어 개수}, \\ &\vec{\boldsymbol{u}} = \frac{\vec{\boldsymbol{v}}}{\|\vec{\boldsymbol{v}}\|_2} = \frac{\vec{\boldsymbol{v}}}{\sqrt{v_1^2 + v_2^2 + \cdots + v_V^2}}. \end{aligned} \tag{2.3}$$

3. 분석 결과

3.1. 경찰백서 분석 결과

우선 경찰백서에 나타난 7가지의 주제에 대해 연도를 구분하지 않고 워드클라우드를 통해 탐색한 결과, 각 주제를 대표하는 키워드들이 뚜렷하게 구분되어 나타났다. Figure 3.1은 각 주제를 탐색하기 위한 방



(a) Topic1- "Police administration"

(b) Topic2-"Background and result of Police business"

(c) Topic3-"Traffic safety and Police business"

(d) Topic4-"Social security and Police business"

(e) Topic5-"Crime trends and major criminal arrests"

(f) Topic6-"Public safety and Police business"

(g) Topic7-"Police business in globalization"

Figure 3.1. Word clouds about each topic from National Police Agency White Paper.

법으로 사용된 주제별 워드클라우드이다. Table 3.1은 각 주제별로 나타난 상위 단어들을 나타낸 표이다. 이러한 데이터 탐색을 바탕으로 각 주제가 어떤 키워드를 중심으로 업무가 기록되었는지 확인할 수 있다.

7가지 주제에 대해 워드클라우드를 살펴본 결과, 각 주제마다 추출된 상위 단어들이 뚜렷하게 구분됨을 알 수 있다. 경찰행정 부분에서는 "치안", "인프라", "확충", "역량", "강화" 등이 나타났다. 이와 같은 키워드를 통해 경찰청의 조직 및 인력 개선 등 치안 인프라를 확충하기 위한 경찰 행정 업무에서의 활동을 알 수 있다.

두 번째 주제인 경찰활동 추진배경 및 성과에 해당하는 워드클라우드를 살펴보면, 눈에 띄는 단어는 "국민", "개선", "수사", "추진", "현장"이다. 해당 주제에서는 국민을 중심으로 경찰청에서 추진 및 개선한 업무 내용을 기술한 것을 알 수 있다. 경찰청의 비전과 추진방향과 관련된 키워드들을 살펴보면, 치안, 안전, 교육, 단속 등이 빈번하게 나타난 것을 파악하였다. 또한 집회 시위에서의 인권과 안전을 강조한 것을 알 수 있다.

Table 3.1. 7 topics from National Police Agency White Paper and top keywords in each topic

Topic	Top keywords
1. 경찰행정 - 경찰 조직, 인력 등	강화, 역량, 인프라, 치안, 확충
2. 경찰활동 추진배경 및 성과	국민, 개선, 수사, 추진, 현장
3. 교통안전과 경찰활동	교통사고, 문화정착, 운전면허, 현황
4. 국민생활과 경찰활동	사회근절, 안전활동, 운영, 지킴이집
5. 범죄추세 및 주요범죄검거	검거, 과학수사, 금융, 사이버, 범죄
6. 사회안보와 경찰활동	북한, 시위, 이탈, 집회, 테러
7. 세계화 시대 속의 경찰활동	국가명, 국제, 기술유출, 범죄, 외국인

세 번째 주제인 교통안전과 경찰활동 주제에서는 “교통사고”, “문화정착”, “선진교통”, “운전면허” 등이 나타났다. 이와 같은 단어의 출현을 통해 경찰청의 교통안전 업무가 어떻게 이루어지는지 알 수 있는 자료로 활용될 수 있다. “교통문화”, “선진”, “운전면허”, “제도” 등을 통해 선진교통문화 정착을 위한 경찰활동을 살펴볼 수 있는 주제이다.

네 번째 주제인 국민생활과 경찰활동의 워드클라우드를 살펴보면, “사회근절”, “아동”, “안전활동”, “운영”, “지킴이집” 등이 크게 나타났다. 이를 통해 국민생활에 밀접한 경찰활동으로 범죄 근절, 아동의 안전 등이 주로 이루어진 것을 추측할 수 있다. 또한 해당 주제에서 국민의 안전을 위한 경찰활동으로 범죄 근절 추진과 건전한 사회환경 조성 및 아동 등 사회적 약자 보호활동이 어떤 키워드를 중심으로 나타나는지 파악할 수 있다. “가정폭력”, “성폭력” 등의 단어를 통해 국민의 안전을 위협하는 일상생활 범죄 예방에 힘쓰고 있으며, “보호”, “실종” 등을 통해 실종 아동 찾기 캠페인과 아동 보호 활동에 노력을 기울이고 있는 것을 확인 할 수 있다.

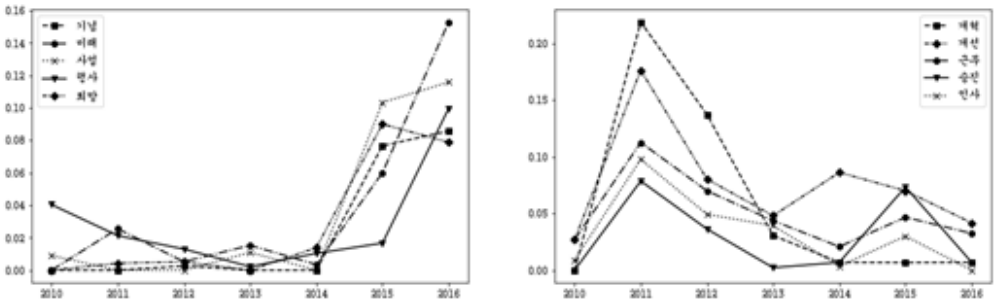
다섯 번째 주제인 범죄 추세 및 주요 범죄 검거의 주요 키워드로는 “검거”, “금융”, “과학수사”, “범죄”, “사이버” 등이 나타났다. 따라서 해당 주제에서는 범죄 발생 현황과 범죄의 형태들이 기술된 것을 알 수 있다. “금융사기”, “사이버범죄”, “전화사기” 등이 나타난 것을 통해 대두되고 있는 범죄 형태는 무엇인지 알 수 있다. 또한 이와 같은 범죄에서 피해자 발생을 막기 위한 경찰청의 과학수사 및 단속과 검거 활동이 중점으로 이루어지고 있는 것을 확인할 수 있다.

여섯 번째 주제인 사회안보와 경찰활동 부분에서는 “집회”, “시위”, “테러”, “북한”, “이탈” 등의 단어가 나타났다. 해당 주제에서 나타난 키워드를 통해 사회안보를 위협하는 요소가 무엇이고 이를 막기 위한 경찰청의 주요 업무 내용이 무엇인지 알 수 있는 주제이다. “테러”, “북한”, “집회”, “시위” 등이 상위에 나타났고 안정된 사회를 위한 경찰활동이 어떻게 이루어졌는지를 알 수 있다.

마지막으로 세계화 시대 속의 경찰활동에서는 “국제”, “기술유출”, “범죄”, “외국인” 등이 나타났다. 상위 키워드를 살펴보면 해당 주제에서는 세계적인 범죄 형태와 이를 해결하기 위한 경찰청의 글로벌 활동이 어떻게 이루어졌는지 알 수 있다. 국제 정세에 따른 경찰청의 활동으로, 외사범죄 단속과 검거활동 등이 중점적으로 기술되어있는 것을 알 수 있다. 이처럼 워드클라우드를 통해 각 주제에서 나타나는 경찰청의 주요 업무 내용이 어떤 키워드를 중점으로 전개되는지를 살펴 볼 수 있었다.

워드클라우드를 통해 상위 단어들과 전체적인 출현 단어를 시각적으로 확인하여 각 주제에 대한 전반적인 탐색을 거친 후, 각 주제별로 나타난 단어들의 변화를 파악하는 분석을 수행하였다. 각 주제별로 분석 결과에 기술된 내용은 해당 자료에서 발굴한 인사이트를 위주로 제시하였다.

첫 번째 주제로 경찰행정 - 경찰 조직, 인력 주제에서는 특정한 추세나 패턴이 도출되지 않았다. 연도별로 경찰행정 부분에서 비슷한 방식으로 서술되어 주요 단어의 연도별 추세가 나타나지 않았고 전체적으로는 “치안”, “평가”, “국민”, “강화”, “개선”, “역량” 등의 단어가 상위 단어로 나타났다. 해당 주제 특



(a) Upward trend words in Topic2 (b) Downward trend words in Topic2

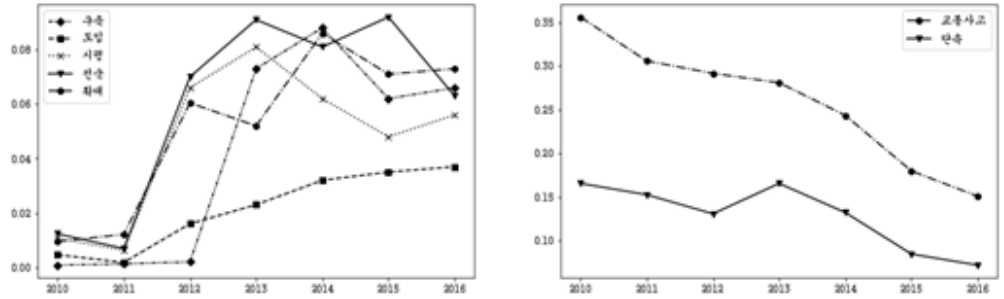
Figure 3.2. Time series plot of words with (a) upward and (b) downward trend in Topic2-"Background and result of Police business".

성상 경찰의 행정업무에 대한 전반적인 내용이 서술되어있고, 경찰 인력현황이나 경찰 인프라 구축현황 등의 수치화된 형태로 기술되어 텍스트를 통한 경찰 행정의 전반적인 내용을 설명하는데 한계가 있다. 하지만 경찰 행정의 방향을 파악할 수 있는 기초 자료로 활용될 수 있는 가능성을 보이고 있다. 행정 부분에서 중요한 키워드인 치안 인프라 확충과 역량 강화를 통한 신뢰 구현을 비전으로 경찰청에서의 행정업무가 이루어지고 있는 것으로 보인다. 경찰청의 목표와 비전을 세우고 이를 이루기 위해 노력하는 경찰의 모습을 반영하기 위한 부분으로 경찰청의 정책 홍보 등에 활용하여 더 발전적인 경찰 행정이 이루어질 수 있을 것이다.

두 번째 주제로 경찰활동 추진배경 및 성과에 대한 분석 결과를 살펴보면 2010년에 나타난 상위 단어들은 2011–2016년에는 나타나지 않고 반대로 2011–2016년에 나타난 상위 단어들은 2010년에는 하위권에서 나타났다 (해당 연도별 키워드 빈도표는 본문에 싣지 않았다). 이를 통해 2010년의 경찰활동 추진배경 및 성과에 대한 기술 방식이 2011년 이후 변화하였음을 알 수 있다. 해당 주제에서 나타나는 전체적인 단어의 연도별 흐름을 파악하기 위하여 경찰활동 추진배경 및 성과 부분에서 2010년 단어는 제외하고 분석을 수행하였다. Figure 3.2는 해당 주제에서 나타난 증가 추세 단어와 감소 추세 단어의 시계열 그래프이다. "기념", "미래", "사업", "행사", "희망"이라는 단어는 2014년 전까지는 변화가 미미하다가 2014년 이후 급격히 증가하는 추세를 보였다. 이는 경찰청 70주년 기념사업이 실시되며 2015년부터 비약적으로 증가한 것으로 파악되었다. 반면 "개혁", "근무", "승진", "인사"는 경찰활동 추진배경 및 성과에서 언급이 감소되는 경향을 보이는 것으로 나타났다.

세 번째 주제로 교통안전과 경찰활동에 관한 분석 결과로 Figure 3.3에서 나타난 것과 같이 증가 추세를 가진 단어와 감소 추세를 띄는 단어를 파악하였다. "교통사고"와 "단속"에 대한 언급은 감소추세를 보이는 반면, 2011년을 기점으로 "확대", "전국", "도입", "시행", "구축"은 증가하는 경향이 있는 것으로 나타났다. 이는 2012년을 기점으로 도시교통정보시스템(Urban Traffic Information System; UTIS)의 구축과 2014년에 정식 도입 이후, 2015년부터 확대 시행중인 것을 통해 관련 단어들의 언급이 증가한 것으로 추정된다. 이러한 추세를 갖는 단어를 통해 경찰청의 사고를 줄이고 원활한 도시 교통 체계를 구축하기 위한 교통 관리 업무에 활용될 수 있을 것이다. 키워드들의 연도별 상대적인 언급량과 교통사고 수, 부상자 수 및 출퇴근 도로 혼잡도 등의 수치 데이터와의 연관성을 비교해보고 어떤 교통 관련 정책이 효과적으로 이루어 졌는지에 대한 평가지표로써 활용될 수 있을 것이다.

네 번째 주제인 국민생활과 경찰활동에 대한 분석에서는 해당 주제에서 연도별 변화가 뚜렷하게 나타난 단어들을 파악하였다. Figure 3.4는 해당 주제에서 뚜렷한 추세를 가진 단어를 그린 그래프이다.



(a) Upward trend words in Topic3 (b) Downward trend words in Topic3

Figure 3.3. Time series plot of words with (a) upward and (b) downward trend in Topic3-"Traffic safety and Police business".

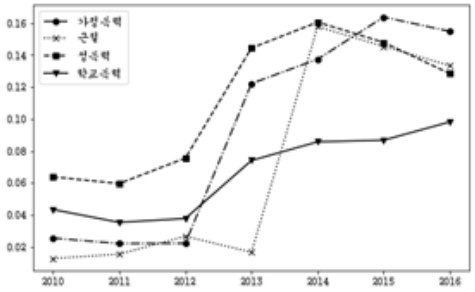


Figure 3.4. Time series plot of words with upward trend in Topic4-"Public safety and Police business".

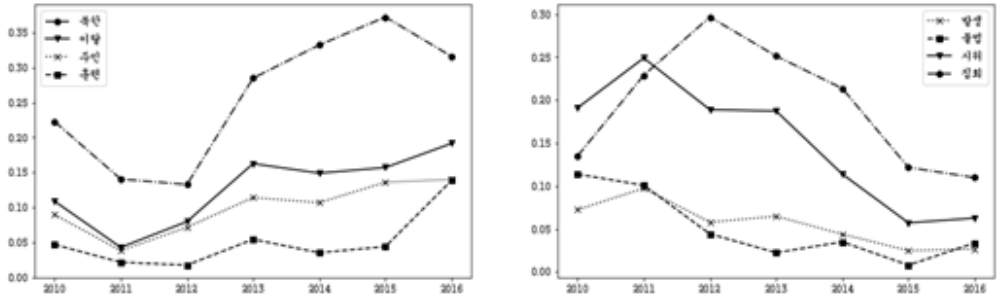
2013년을 기준으로 "가정폭력", "학교폭력", "성폭력", "근절"이 급증하였다. 이는 2014년을 기점으로 경찰청에서 '4대 사회악 근절 전략'을 추진하여 관련 업무에 대한 언급이 증가한 것으로 보인다. 그리고 해당 주제에서는 뚜렷한 감소 추세로 나타난 단어는 나타나지 않았다.

다섯 번째 주제인 범죄 추세 및 주요 범죄 검거에서는 추세를 가진 단어를 파악할 수 없었다. 각 연도별로 "범죄", "수사", "검거", "사이버" 등의 단어가 상위 단어로 나타났고 각 연도별로 기술된 내용 전개가 비슷하게 이루어진 것을 확인하였다. 주제 특성상 범죄 종류별 발생 건수, 검거 현황 등의 수치화된 자료로 기술되어 해당 주제에서는 추세를 가진 단어를 확인할 수 없었다.

여섯 번째 주제인 사회안보와 경찰활동 주제에서 나타난 추세로는 2011년을 기점으로 "집회", "시위", "발생", "불법"은 감소하는 경향을 보였고, "북한", "안보"는 증가하는 추세로 나타났다. Figure 3.5는 사회안보 관련 주제에서 나타난 증가 및 감소 추세를 가진 단어에 대한 그래프이다. 특히 2013년을 기점으로 "이탈", "주민", "훈련"이 증가한 것을 보아 2013년 이전에는 불법시위, 집회에 관한 치안 업무를 중심으로 기술되었고 이후에는 북한 이탈 주민에 대한 정책과 테러 대응에 관한 업무를 중심으로 기술된 것을 확인할 수 있다.

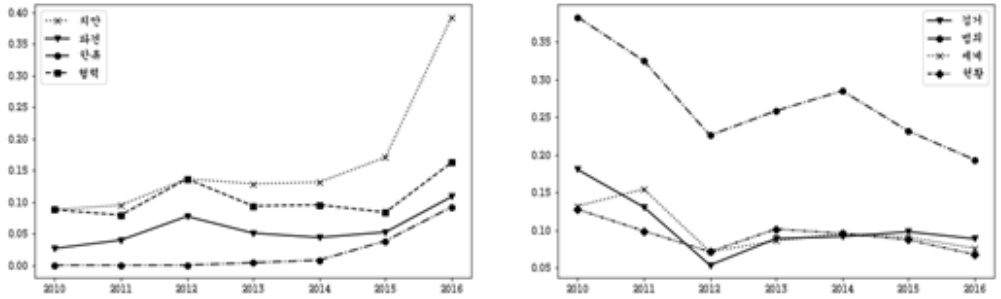
마지막으로 세계화 시대 속의 경찰활동 분석 결과를 살펴보면 "범죄", "검거", "증가", "세계", "현황"은 감소하는 경향을 나타낸 반면, "치안", "한류", "협력", "파견"은 증가하는 형태로 나타났다. 이는 2015년 한국형 선진 치안 시스템의 파급으로 인한 결과로 추정되며 Figure 3.6을 통해 증가 및 감소 추세 단어의 경향성을 확인할 수 있다.

분석결과를 종합해보면, 각각의 주제에서 증가하는 경향이 있는 단어와 감소하는 경향이 있는 대표적인



(a) Upward trend words in Topic6 (b) Downward trend words in Topic6

Figure 3.5. Time series plot of words with (a) upward and (b) downward trend in Topic6-"Social security and Police business".



(a) Upward trend words in Topic7 (b) Downward trend words in Topic7

Figure 3.6. Time series plot of words with (a) upward and (b) downward trend in Topic7-"Police business in globalization".

단어들을 살펴보고, 이러한 추세를 통해 어떠한 정책이 반영된 결과인지 살펴보았다. 경찰백서의 자료 특성을 생각해보면, 업무를 담당한 부서에서 한 해 동안의 성과를 정리하고, 새로운 목표를 이루기 위한 정책을 제시하고 있다. 이러한 특성 상 경찰백서 분석 결과는 과거의 각 주제마다 업무를 요약적으로 제시하고 있다. 추후 경찰백서의 발간이나 정책을 제안하는데 있어 본 분석 결과를 활용하여 과거의 트렌드를 살펴봄으로써, 더 효율적인 업무가 진행될 수 있을 것으로 보인다.

3.2. 지방청 업무 보고 분석 결과

16개의 지방청 업무 보고 문서를 분석한 결과, 단순히 단어의 단순 출현 빈도를 통해서 각 지방청별 주요 업무는 파악할 수 있었지만, 각 지방청마다의 업무적 특징을 나타내는 키워드를 찾기에는 부족하였다. 텍스트 데이터 탐색을 목적으로 각 지방청별 워드클라우드를 그린 결과는 Figure 3.7과 같다.

대부분의 지방청에서 "수사", "단속", "검거", "홍보", "교통" 등의 단어가 상위 빈도를 차지하였다. 또한 대구지방청에서의 "수성", 대전지방청에서의 "중부", "동부", 서울지방청에서 "광화문", "청사", 부산지방청에서 "해운대", 충남지방청에서 "아산", "서산" 등 각 지방청별로 지역을 뜻하는 단어들이 상위 키워드를 차지하였다. Table 3.2는 각 지방청별로 출현 빈도 상위 10개의 단어를 정리한 표이다. Table 3.2와 Figure 3.7에서 나타난 결과를 통해 대략적인 지방청별 업무 내용을 알 수 있다. 데이터 탐색을 통해 전체적으로 대부분의 지방청에서 수사, 검거, 치안, 교육, 홍보, 교통 업무 등이 중점적으로 이루어

Figure 3.7. Word clouds for each Metropolitan Police Agency.

지는 것을 알 수 있다. 전반적인 업무 내용 파악과 더불어 지방청별로 문서를 통해 업무의 차이를 파악하기 위해서 해당 지방청 문서를 특징짓는 키워드를 추출하는 과정이 필요하여 식 (2.2)의 TF-IDF 식을 사용하였다. TF-IDF를 사용하기 전 상위 단어를 통해서는 해당 지방청의 업무 보고 내용을 비교하기 어렵기 때문에 Table 3.3에서와 같이 TF-IDF 가중치를 이용해 단어 빈도를 변환하여 상위 단어를 비교하였다.

우선 TF-IDF 가중치를 적용한 결과인 Table 3.3을 통해 각 지방청별로 특징적인 상위 키워드를 살펴보면 각 지방청마다 상위를 차지하고 있는 키워드는 해당 지방청에서 관리하는 행정지역 명으로 나타났다. 이를 통해 각 지방청마다 어느 지역에 관심이 집중되어 있는지를 알 수 있다. 강원지방청에서 나타난 지역 순위는 “삼척”, “인제”, “철원”, “춘천” 순으로 나타났고 이외에도 하위 단어들에서 지역 키워드들이 다양하게 나타났다. 경기남부지방청의 경우, “중원”, “오산”, “과천”, “시흥”순으로 지역 키워드들이 나타났고, 경기북부지방청의 경우, “동두천”, “포천”, “킨텍스” 순으로 나타났다. “군단”이라는 단어도 상위 키워드인 것을 보아 육군과의 업무 협업이 이루어진 것으로 추측된다. 경남지방청의 경우, “거창”,

Table 3.2. Top 10 words by freqeuncy for each Metropolitan Police Agency

지방청	상위 10개 단어 (출현 빈도 기준)
강원지방청	교육, 교통, 단속, 대상, 수사, 안전, 장소, 점검, 홍보, 현장
경기남부지방청	경정, 교육, 단속, 대상, 실시, 안전, 예방, 장소, 특별, 홍보
경기북부지방청	경정, 교육, 단속, 대상, 실시, 안전, 예방, 장소, 특별, 홍보
경남지방청	경남, 교육, 교통, 단속, 대상, 안전, 점검, 현장, 홍보, 수사
경북지방청	검거, 교육, 대비, 대상, 대상, 안전, 점검, 치안, 현장, 수사
광주지방청	계획, 광주, 교육, 교통, 단속, 대상, 수사, 장소, 현장, 점검
대구지방청	교육, 교통, 대상, 수사, 수성, 예방, 장소, 점검, 현장, 대구
대전지방청	검거, 교통, 단속, 대상, 범죄, 수사, 신고, 장소, 점검, 대전
부산지방청	교통, 단속, 대상, 수사, 신고, 안전, 장소, 점검, 현장, 부산
서울지방청	교육, 근무, 단속, 서울, 수사, 실시, 안전, 점검, 현장, 교통
울산지방청	교통, 남부, 단속, 대비, 대상, 동부, 울산, 중부, 홍보, 안전
전남지방청	계획, 교육, 기능, 단속, 대상, 실시, 안전, 점검, 현장, 수사
전북지방청	교육, 군산, 대상, 수사, 신고, 전북, 점검, 참석, 현장, 장소
제주지방청	교육, 교통, 단속, 동부, 서부, 안전, 장소, 제주, 홍보, 예방
충남지방청	검거, 교육, 수사, 안전, 장소, 점검, 특별, 현장, 홍보, 단속
충북지방청	교육, 교통, 대상, 안전, 장소, 점검, 참석, 현장, 형사, 수사

Table 3.3. Top 10 keywords by TF-IDF for each Metropolitan Police Agency

지방청	상위 10개 단어 (TF-IDF값 기준)
강원지방청	남면, 동해, 삼척, 성취, 속초, 알파, 인제, 철원, 춘천, 품격
경기남부지방청	과천, 광명, 산지, 소사, 시흥, 안산, 안양, 오산, 의왕, 중원
경기북부지방청	군단, 남양주시, 동두천, 레드, 발전소, 석탄, 스프링, 클라우드, 킨텍스, 포천
경남지방청	거제, 거창, 구름, 김해, 남해, 마산, 밀양, 사천, 진주, 하동
경북지방청	구미, 급습, 무지개, 문경, 영덕, 영주, 의성, 칠곡, 포항시, 플랜트
광주지방청	광산, 금남, 도산, 마네킹, 무등, 발전소, 양동, 오명, 월곡, 조선대
대구지방청	계명대, 관문, 네거리, 무학, 부처님오신날, 서방, 성서, 수성, 영남, 일기
대전지방청	관인, 내동, 네거리, 대덕, 둔산, 순마, 인동, 중리, 즉결, 현충원
부산지방청	기장, 동래, 동백, 벡스코, 사하, 송도, 연제, 영도, 위안부, 태화
서울지방청	관악, 광진, 구름, 금천, 남대문, 마포, 성동, 성북, 양천, 은평
울산지방청	미리내, 북구, 빅워크, 산지, 삼산, 수공, 언양, 일계, 젊음, 태화
전남지방청	광양, 나주, 목포, 순천, 영광, 영암, 완도, 유투브, 편암함, 해남
전북지방청	고창, 군산, 김제, 덕진, 송전탑, 완산, 완주, 익산, 정읍, 철근
제주지방청	관광대, 비단, 서귀포, 올레, 일게, 제주, 카지노, 한라, 한림, 해안
충남지방청	공주, 금산, 농림부, 당진, 밀실, 보훈처, 산자부, 유투브, 해수부, 홍성
충북지방청	괴산, 단양, 보은, 옥천, 육거리, 제천, 진천, 창간, 청주, 흥덕

“밀양”, “남해” 등의 지역이 나타났고, 경북지방청의 경우, “칠곡”, “문경”, “영덕” 등이 나타났다. 광주지방청에서는 “무등”, “광산”, “조선대”가 상위를 차지하였고, 대구지방청에서는 “성서”, “무학”, “수성” 순으로 나타났다. 이처럼 각 지방청 별로 지역에 대한 언급량을 통해 어느 관할지역에 업무가 중점적으로 이루어지고 있는지 분석 결과를 통해 알 수 있다.

이후 각 지방청의 업무적 특성을 띄는 단어를 찾기 위해 지역을 제외하고 상위 키워드를 살펴보았다. Table 3.4는 TF-IDF 값을 적용하고 각 지방청마다 지역 이름을 제외하고 업무적인 특성을 띄는 단어를 나열한 결과이다. 지방청별 상위 10개 업무 관련 특징 키워드를 살펴보면 대략적인 지방청의 특색을 파악할 수 있다. 전반적인 주요 업무로는 교통에 관한 업무와 단속, 검거, 수사 업무 및 홍보와 교육에

Table 3.4. Top 10 business-related keywords by TF-IDF for each Metropolitan Police Agency

지방청	상위 10개 업무 관련 키워드(TF-IDF값 기준)
강원지방청	맑음, 성취, 수사, 스크린도어, 안전, 얼음, 평창동계올림픽, 품격, 플래시몹, 현장
경기남부지방청	단속, 담벼락, 등교시간, 마사회, 민사법, 안전, 유치권, 전조등, 트랙터, 홍보
경기북부지방청	골프장, 군단, 레드, 발전소, 발전소, 변전소, 석탄, 스프링, 클라우드, 킨텍스
경남지방청	교육사, 국화, 뉴스레터, 무슬림, 북카페, 영상편지, 육군훈련소, 자동차번호판, 중공업, 포위
경북지방청	급습, 당번, 몸싸움, 무지개, 쌀값, 제철, 진보, 포스코, 플랜트
광주지방청	건국, 고인, 단속, 마네킹, 민주화, 발전소, 수사, 오명, 점검, 현장
대구지방청	개제, 교육, 교통, 부처님오신날, 수사, 스타디움, 일기, 장소, 특약, 현장
대전지방청	갑반, 교통, 북카페, 수사, 연막, 월드컵경기, 즉결, 철도공사, 현충원, 효도
부산지방청	교섭, 부두, 생존권, 실험실, 위안부, 육군훈련소, 재벌, 진보, 최저임금, 환경미화원
서울지방청	구름, 대사관, 수신기, 인양, 잠수, 전자상거래, 제야, 증후군, 철야, 타종행사
울산지방청	단속, 대비, 맑음, 빅워크, 산지, 안전, 젊음, 중공업, 포위, 홍보
전남지방청	가람, 결선, 면세, 미팅, 어선, 유투브, 통근, 투광, 트로피, 편안함
전북지방청	봉쇄, 송전탑, 월드컵경기, 장수, 재개, 지리산, 철근, 프로축구, 하나은행, 황조
제주지방청	관광대, 교통, 기내, 안전, 예방, 월드컵경기, 장방형, 카지노, 해안, 홍보
충남지방청	공정위, 농림부, 밀실, 보훈처, 산자부, 식권, 유투브, 타자, 한일, 해수부
충북지방청	교통, 수사, 장소, 점검, 증평, 창간, 축전, 취하, 한가위, 휴양림

관한 업무가 중점적으로 다루어졌고, 특징 키워드를 살펴보았을 때, 어느 관할지역에 관심도가 높은지를 파악할 수 있다. 또한 지역명을 제외한 업무 관련 상위 특징 키워드를 살펴보았을 때, 지방청의 특색을 파악할 수 있는 단어가 나타났다. 예를 들어, 강원지방청에서의 "평창동계올림픽"이란 단어를 보았을 때, 동계올림픽 개최에 대비하기 업무가 진행되었을 것으로 추측된다. 경기남부지방청의 경우 "전조등", "등교시간", "트랙터" 등이 상위로 나타난 것으로 보아, 해당 지방청 교통 단속 업무에서의 특징을 대표할 수 있는 결과로 보인다. 이 밖에도 제주지방청에서 특징 단어로 나타난 "카지노", "관광대", "해안" 등을 통해 해당 지방청 업무에서의 대표적인 업무 동향을 파악할 수 있다. 충남지방청의 경우, "산자부", "해수부", "농림부"가 나타난 것으로 보아 다른 기관과의 업무 협력이 이루어진 것을 확인할 수 있었다.

이처럼 분석 결과를 통해 전체적인 지방청의 주요 업무를 살펴보고, 단어의 중요도를 고려하여 해당 지방청의 특징적인 키워드를 살펴보았다. 이러한 결과를 통해 지방청의 업무 방향성을 파악할 수 있는 기초 자료로 활용하고 정책 의사 결정시 전반적인 업무 요약과 지방청 특색을 파악할 수 있는 자료로 활용되어 경찰청 업무 향상에 기여할 수 있을 것으로 기대된다.

4. 결론

본 논문에서 사용한 경찰백서 자료를 통해 이용하여 7가지 주제에 대한 워드클라우드를 그려봄으로써 경찰청 업무에 대한 전반적인 구성을 살펴볼 수 있었다. 또한 단어들의 트렌드를 분석하여 이슈가 되었던 정책들을 파악할 수 있었다. 2013년을 기준으로 '4대 사회악 근절 전략' 관련 단어들이 증가한 것과 2012년을 기준으로 도시교통정보시스템이 도입되어 교통 정책에서 이슈가 되었음을 확인할 수 있었다. 또한 '한국형 선진 치안 시스템'이 2015년에 실시되어 관련 단어들이 증가하였음을 확인하였다. 범죄 추세 및 주요 범죄 검거 주제와 경찰행정 주제에서는 추세가 나타나진 않았지만, 추후 연구에서 형태별 범죄수 데이터 또는 부서별 인력 배치와 같은 정형화된 데이터와 비정형 텍스트 데이터에서 나온 결과와의 연관성을 파악하여 인사이트를 도출하는 연구로 이루어 질 수 있다. 그리고 지방경찰청 업무 보고 자료를 이용하여 지방청 특징 키워드 추출을 위한 방법을 살펴보았고, 이 결과를 토대로 지방청 간 업무

특성 차이를 갖는 단어들을 비교하였다. 이를 통해 전체적인 지방청의 업무 중점 사항과 각 지방청만의 특징을 갖는 업무들을 파악하여 전반적인 지방청 업무의 중요도를 확인할 수 있었다.

경찰청 업무에서 새로운 정책을 제안하는 경우 또는 지방청 별로 다른 지방청과의 업무 비교를 통해 개선점을 제안하는 경우, 새로운 정책에 대한 근거로 이러한 분석 결과를 활용할 수 있을 것이다. 4대 사회악 근절 캠페인에 대한 성과를 검토하는 경우, 텍스트에 나온 언급량을 통해 해당 업무에 얼마나 집중했는지를 알 수 있는 자료로 활용될 수 있을 것이다. 물론 이에 대한 성과로 주민들의 체감안전도, 범죄재범률 등의 정량적인 지표가 기본적으로 활용되고 이에 추가적으로 텍스트 데이터 기반의 정량적인 지표를 활용하여 더 나은 의사결정의 도구로서의 기능을 고려할 수 있다. 또한 지방청별 업무 보고에 나타난 단어들에 대한 비교를 통해 전반적인 지방청별 업무가 어떻게 이루어지고 있는지 살펴보았다. 공통적으로 나타난 주요 업무는 교통, 교육, 수사, 현장점검 등의 업무가 주를 이루었다. 추가적으로 해당 지방청에서 특징적으로 나타나는 키워드를 비교해보았을 때, 해당 지방청의 지역적 특징이 뚜렷하게 구분되었다. 이를 활용하여 지역별 범죄건수 또는 사고 발생수 같은 정형화된 데이터와 지역 언급량과의 상관관계를 통해 지방청별로 관할지역에 대한 지역 집중도가 효율적으로 이루어지고 있는지 평가할 수 있는 자료로 활용할 수 있다. 지역적 특징을 제외하고 업무적인 특징을 비교하였을 때, 지방청별로 상호보완적인 정책 결정 구조를 확립할 수 있을 것이다. 예를 들어, 전남지방청의 경우 '유투브'를 통한 지방청 홍보 업무를 수행하여 해당 지방청에서 TF-IDF값이 높게 나타났고, 이에 대한 긍정적인 효과를 살펴봄으로써 다른 지방청에서도 시도해 볼 수 있을 것이고, 새로운 정책 아이디어를 수립하는데 활용될 수 있을 것이다.

이처럼 두 가지 자료에 대해 단어의 증가 및 감소 추세를 파악하거나 특징 단어의 중요도 판별을 수행함으로써 경찰청 업무에서 어떤 정책이 반영되었는지 확인할 수 있었고, 연구 결과를 통해 앞으로 어떤 흐름의 정책이 필요한지에 대한 자료로 활용할 수 있을 것이다. 더불어 지방청 업무 보고 자료에 대한 분석 결과를 통해 지방청의 업무가 해당 지역에 알맞게 이루어지고 있는지에 대한 평가의 기초 자료로 활용되어 각 지방청의 업무 개선에 도움이 될 수 있을 것으로 기대한다. 추후 연구에서는 경찰 업무 관련 신문기사와 같이 경찰백서나 업무보고 자료 외의 경찰 관련 텍스트 자료를 이용하여 문서에 나타난 핵심 단어들을 살펴보고 자료에 맞는 분석 기법을 적용하는 분석으로 수행함으로써 경찰청 자료 사용자들의 요구나 전반적인 흐름을 파악하고 향후 정책 마련 시 기초 자료로 활용될 수 있도록 할 계획이다.

본 논문에서는 키워드의 출현 빈도에 따른 현상 분석을 중점으로 다루었기 때문에 해석 가능성이 높은 Bag-of-words 방식으로 연구를 진행하였다. Bag-of-words 방식 외에 Word2Vec (Mikolov 등, 2013)나 Glove (Pennington 등, 2014) 등의 임베딩 모형을 통한 단어와 문서에 대한 축소된 벡터의 분산 표현 방식 등이 있고, 최근 단어의 임베딩을 이용하여 문서 분류와 감성 분석 등 다양한 분야에서 활용이 시도되고 있다. 추후에는 본 연구에서 나타난 분석 결과를 기초 자료로 활용하고 다양한 기계학습 알고리즘들과 함께 더 수준 높은 인사이트 발굴을 목표로 후속 연구를 진행할 계획이다.

References

Bae, J. H., Son, J. E., and Song, M. (2013). Analysis of Twitter for 2012 South Korea Presidential Election by text mining techniques, *Journal of Intelligence and Information Systems*, **19**, 141–156.

Berry, M. W. (2004). Survey of text mining, *Computing Reviews*, **45**, 548.

Cho, S. G. and Kim, S. B. (2011). Finding meaningful pattern of key words in IIE transactions using text mining. In *2011 Fall Conference Proceedings of Korean Institute of Industrial Engineers*, 443–452.

Grimes, S. (2008). Unstructured data and the 80 percent rule, *Carabridge Bridgepoints*, 10.

Kothe, G. (1983). Topological vector spaces. In Topological Vector Spaces I, *Springer, Berlin, Heidelberg*, 123-201

Leopold, E. and Kindermann, J. (2002). Text categorization with support vector machines. How to represent texts in input space?, *Machine Learning*, **46**, 423–444.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 3111–3119.

Nahm, U. Y. and Mooney, R. J. (2002). Text mining with information extraction. In *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 60–67.

Park, E. L. and Cho, S. (2014). KoNLPy: Korean natural language processing in Python. In *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, 133–136.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., and Vanderplas, J. (2011). Scikit-learn: machine learning in Python, *Journal of Machine Learning Research*, **12**, 2825–2830.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543.

Python Software Foundation (2017). Python Language Reference, version 3.6. Available from: http://www.python.org

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, **242**, 133-142.

Song, H. J., Park, K. S., Jung, H. E., and Song, M. (2013). Trend Analysis of Korean Economy in the Economic Literature by text mining techniques. In *Proceedings of the 20th Conference on Korea Society for Information Management*, 47–50.

Sulova, S., Todoranova, L., Penchev, B., and Nacheva, Radka. (2017). Using text mining to classify research papers. DOI:10.5593/SGEM2017/21/S07.083

Talib, R., Hanif, M. K., Ayesha, S., and Fatima, F. (2016). Text mining: techniques, applications and issues, *International Journal of Advanced Computer Science & Applications*, **1**, 414–418.

텍스트 마이닝 기법을 이용한 경찰청 업무 트렌드 분석

선현석[a] · 임창원[a,1]

[a]중앙대학교 응용통계학과

(2018년 10월 15일 접수, 2018년 11월 21일 수정, 2019년 1월 31일 채택)

요 약

최근 통계적인 기법을 이용하여 대량으로 생산되고 있는 텍스트 데이터를 통해 다양한 인사이트 발굴을 하기 위한 연구가 활발히 진행되고 있다. 본 연구는 경찰청에서 생산하는 텍스트 데이터를 통해 연도별 경찰청의 업무 트렌드를 파악하고, 각 지방청별로 생산되는 문서에서 주요 키워드를 파악하여 지방청 간의 업무 특성을 비교하고자 하였다. 의미 있는 결론을 도출하기 위해 각 자료 특성에 맞는 전처리 과정을 시행하고 문서별 단어 빈도수를 계산하였다. 문서에 나타난 키워드의 단순 출현 빈도로는 해당 키워드가 문서에서 갖는 중요도를 설명하기 힘들기 때문에 단어-역문서 가중치를 이용하여 각 단어에 대한 빈도수를 새롭게 계산하였고 단어의 문서별 및 연도별 빈도 비교를 위해 L2 정규화 기법을 이용하였다. 이러한 분석은 향후 경찰청 업무 개선 정책에 새롭게 활용될 수 있는 기초 자료로 사용될 수 있으며, 경찰청 업무 효율성 향상 및 청내 업무 개선 수요 파악을 위한 방법으로 활용될 수 있다.

주요용어: 텍스트 마이닝, 비정형 데이터, 경찰청, 단어-역문서 빈도, 키워드 추출

이 논문은 2017년도 중앙대학교 연구장학기금 지원에 의한 것임.

[1]교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: clim@cau.ac.kr