Journal of Digital Convergence
Vol. 19. No. 3, pp. 167-176, 2021

ISSN 2713-6434 / eISSN 2713-6442
https://doi.org/10.14400/JDC.2021.19.3.167

인과관계문형 기반 사회이슈 발생원인 도출 방법 연구

이남연[1*], 이재형[2]
[1]한신대학교 IT경영학과 조교수, [2]한신대학교 IT경영학과 학사과정

A study on the method of deriving the cause of social issues based on causal sentences

Namyeon Lee[1*], Jae Hyung Lee[2]
[1]Assistant Professor, Department of IT Management, Hanshin University
[2]Student, Department of IT Management, Hanshin University

요 약 최근 빅데이터 분석 기술이 발전하면서 사회 이슈를 분석하기 위해 그 동안 많은 텍스트 마이닝 기법을 활용한 연구들이 진행되어왔다. 사회이슈를 도출하기 위한 기존의 연구들을 살펴보면 다량의 텍스트 데이터를 뉴스, SNS 등으로부터 수집하여 토픽 모델링, 네트워크 분석 등의 기법을 이용하여 데이터로부터 이슈를 추출하고 분석하는 방식으로 연구들이 이루어져왔다. 사회 이슈는 다양한 사회현상들이 누적되어 나타나는 결과물이다. 하지만 기존 연구들이 가지는 한계점은 사회적으로 나타나는 이슈, 즉 결과에 대한 분석에 초점이 맞춰져 있어 해당 이슈의 발생 원인을 밝히는 것에는 한계를 가진다는 것이다. 사회이슈에 적절하게 대응하기 위해서는 어떠한 사회이슈가 존재하는지를 확인하는 것뿐만 아니라 사회이슈의 발생 원인을 파악하는 것이 필요하다. 이러한 한계점을 극복하기 위해서 본 연구에서는 사회이슈와 관련한 텍스트로부터 사회이슈의 원인이 되는 요인을 도출하는 방법을 국어학의 품사이론을 기반으로 제안하였다. 이를 위해서 2017년 1월부터 2019년 12월까지의 3년 동안의 사회이슈와 관련한 뉴스데이터를 수집하여 수집된 텍스트 내 단어들의 인과관계를 인과문형을 찾아 분석한 후 기존 텍스트마이닝 기법 접목하여 사회이슈의 원인 단어들을 찾는 방법론을 제안하였다.

주제어 : 사회이슈, 인과문형, 텍스트마이닝, 빅데이터, 연결어미

Abstract With development of big data analysis technology, many studies to find social issues using texts mining techniques have been conducted. In order to derive social issues, previous studies performed in a way that collects a large amount of text data from news or SNS, and then analyzes issues based on text mining techniques such as topic modeling and terms network analysis. Social issues are the results of various social phenomena and factors. However, since previous studies focused on deriving social issues that are results of various causes, there are limitations to revealing the cause of the issues. In order to effectively respond to social issues, it is necessary not only to derive social issues, but also to be able to identify the causes of social issues. In this study, in order to overcome these limitations, we proposed a method of deriving the factors that cause social issues from texts related to social issues based on the theory of part of Korean linguistics. To do this, we collected news data related to social issues for three years from 2017 to 2019 and proposed a methodology to find causes based causal sentences based on text mining techniques.

Key Words : Social issues, Causal sentence, Text-mining, Big data, Connection ending

*This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2018R1C1B5086335)
*Corresponding Author : Namyeon Lee(nylee@hs.ac.kr)
Received December 28, 2020 Revised February 19, 2021
Accepted March 20, 2021 Published March 28, 2021

1. 서론

현대 사회의 급속한 변화에 따라 많은 사회적인 문제들이 발생하고 있으며 이 중 일부는 사회적인 관심이 큰 사회적 이슈가 된다. 즉, 사회적 이슈는 현재 우리가 경험하는 현상 중 대중의 관심을 많이 받으며 사회적인 영향력이 큰 현상이라고 할 수 있다.

과거에는 해당 분야의 전문가로부터 의견을 수렴하는 top-down 방식으로 사회적 이슈를 선정하고 분석해왔다. 하지만 전문가들을 통한 사회이슈 분석 방법은 소수의 전문가가 사회 전반에 이르는 문제들을 다루기 때문에 전문가들의 배경, 경험, 지식의 범위 등에 따라 사회현상에 대한 분석 결과가 다르게 나타날 수 있는 오류를 내포하고 있다[1]. 이러한 문제들로 인해 최근에는 빅데이터와 텍스트 분석을 활용하여 사회이슈를 도출하고 분석하고자 하는 연구들이 활발하게 이루어지고 있다[2, 3, 4]. 하지만 기존의 텍스트 분석을 활용한 사회이슈 도출 및 분석 연구들은 현재 어떠한 이슈들이 나타나고 있는지 이슈를 발견하는 것에 초점이 맞춰져 있다. 따라서 현재 어떠한 이슈가 나타나고 있고, 해당 이슈에 대한 설명은 가능하나 왜 이러한 이슈가 발생하게 되었는가에 대해서는 설명하기 어려운 한계점이 있다.

사회적 이슈는 여러 원인들로 인해 발생하는 결과물이다. 예를 들어, 최근 대기오염은 사회적으로 이슈가 된 사회 문제 중 하나이다. 대기오염의 원인으로 미세먼지, 자동차 배출가스 등이 원인요인으로 고려될 수 있으며, 이러한 원인 요인들로 인해 대기오염이라는 현재의 상황이 나타나게 된 것이다. 사회적인 이슈들은 이슈의 원인이 되는 요인들이 누적되면서 새로운 결과 이슈들을 만들어내고, 이러한 원인 요인들의 영향력이 크면 클수록 결과로 나타나는 이슈의 사회적인 파급력이 커지게 된다. 또한 원인 요인들의 영향력이 시간이 흐르면서 줄어들면 결과로 나타나는 이슈의 파급력은 줄어들게 된다. 이와 같이 사회이슈는 계속해서 생성되고 변화하고, 소멸되는 일종의 생명주기를 가지게 된다[5. 6]. 따라서 사회적 이슈를 보다 정확하게 분석하기 위해서는 현재 사회이슈의 원인요인을 탐색하고 탐색한 결과를 바탕으로 해당 이슈가 어떠한 생명주기 하에 있는지를 확인하고, 해당 이슈가 앞으로 어떠한 단계로 변화할지에 대한 예측이 가능해야한다[7].

따라서 본 연구에서는 이슈의 원인 및 결과 요인을 탐색하는 방법을 제안하고, 탐색된 인과요인을 바탕으로 해당 이슈의 시간에 따른 변화를 탐색하는 방법론을 제시하고자 한다. 이를 위해 본 연구에서는 세계경제포럼(WEF)에서 매년 발표하는 글로벌 리스크를 바탕으로 시드 단어(이슈)를 선정하고, 선정된 시드 단어에 대한 2017년부터 2019년까지의 3년 동안의 뉴스기사를 수집하였다. 이후 뉴스기사에서 해당 시드 단어가 포함된 문장만을 추출하여 말뭉치(corpus)를 구축하였다. 구축된 말뭉치를 바탕으로 원인과 결과를 포함하는 인과문형 패턴을 연결어미 등의 형태소 분석을 통해 도출하였다. 도출된 인과문형에 따라 시드 단어를 결과로 하는 원인 요인을 탐색하고, 이를 네트워크로 도식화하여 특정 이슈의 발생에 어떠한 요인의 영향력이 높은 지 판단할 수 있는 방법을 제안하였다.

2. 이론적 배경

우리 사회에서 나타나는 이슈들을 분석하고자 하는 연구는 다양한 분야에서 이루어져 왔다. 이들 이슈 분석 연구의 목적은 사회에서 다루는 주요 주제를 파악하고 이를 해결하기 위한 대응 방안을 마련하고자 하는 것이 목적이다. 이와 관련하여 기존의 연구들을 살펴보면 크게 전문가 중심의 top-down 방식과 데이터 기반의 bottom-up 방식의 2가지의 방식으로 사회이슈를 분석하고자 하였다. 전문가 중심의 top-down 방식에서는 주로 AHP 기법, 델파이 기법 등을 활용하여 사회이슈를 도출하고 분석하였다[8]. 하지만 전문가 중심의 분석 방법은 전문가가 보유한 지식의 수준이나 접근 가능한 정보의 종류와 양에 따라 결과가 전문가별로 다르게 나타날 수 있고, 전문가가 수동으로 분석하기 때문에 분석해야 하는 데이터의 양이 많은 경우 분석까지의 시간의 오래 걸릴 수 있다. 이로 인해 때로는 분석이 불가능한 경우도 발생할 수 있다[5, 9]. 따라서 객관적인 분석 및 시급성을 필요로 하는 경우에는 전문가 기반의 사회 이슈 분석 방법은 한계를 가진다.

이러한 전문가 중심의 Top-Down 방식의 사회 이슈 분석 방법이 가지는 한계점을 극복하기 위해서 최근에는 빅데이터 분석 기법을 기반으로 한 Bottom-Up 방식의 자동화된 사회이슈 분석 연구들이 활발하게 이루어져왔다. 이들 연구에서는 여러 사람들이 SNS 등의 온라인에 남긴 글을 분석하거나, 뉴스데이터를 활용하여 현재 사회에서 주요하게 나타나는 이슈들을 분석하고자 하였다. 배정환(2013)의 연구에서는 사회적 이슈를 포착하기 위해 토픽모델링 분석 기반의 실시간 트위터 트렌드 마이닝

시스템을 개발하였다[10]. 하지만 이슈 추출 및 유형 분류의 경우 인위적 방법에 의존하며 개별 소셜 빅데이터와의 상관성 분석이 부족하고, 신조어 등 마이크로블로깅에서 사용되는 용어나 문맥이 반영되기 어렵다는 한계를 가진다. 정다미(2013)의 연구에서는 사회적 이슈를 다루고 있는 대용량의 뉴스기사를 수집하고 통계적인 기법을 통하여 키워드를 추출하는 시스템의 개발을 제안하였고, 토픽모델링 기법을 활용하여 사회이슈를 추출하고자 하였다[11]. 허정(2014)의 연구에서는 소셜미디어 분석과 분석보고서 생성의 문제점(분석의 고립성, 전문가의 주관성, 고비용에 기인한 정보의 폐쇄성)을 해결하기 위한 소셜 빅데이터 마이닝에 기반 한 이슈 분석보고서 자동 생성 시스템 개발하였으며, 특히 이 연구에서는 이슈와 감성 사이의 상관성을 분석하고자 하였다[9]. 하지만 분석하는 이슈가 특정 상품을 대상으로 한 소비자의 반응을 분석한 연구이기 때문에 사회이슈로 확장하는 것에는 한계가 있다. 이새미(2020)의 연구에서는 사회적 정책 의제를 도출하기 위해서 특허와 뉴스 기사를 토픽모델링 기법을 이용하여 분석하였다 [12]. 또한 기재홍(2020)의 연구에서 역시 특정 이슈에 대해 토픽모델링 기법을 활용하여 해당 이슈를 분석하였다 [13]. 이수련(2020)의 연구에서도 특정 사회이슈에 대해서 텍스트 마이닝 분석 연구를 수행하였다[14]. 해당 연구에서는 감성분석, 연관분석, 군집분석의 방법을 이용하여 해당 이슈와 연결된 단어들을 분석하여 해당 이슈에 대한 변화 추이 등을 분석하였다. 이와 같이 최근의 많은 연구에서 빅데이터 분석 기법을 이용하여 사회이슈 및 사회적 의제를 도출 및 분석하는 연구들이 이루어져오고 있다. 하지만 기존의 연구들이 가지는 한계점은 나타난 이슈를 찾는 것에 집중하거나 이슈와 연결된 단어들을 바탕으로 이슈를 이해하는 것에 집중한 연구들이라는 것이다. 즉, 현재 사회적으로 어떠한 일들이 벌어지고 있는지를 확인하는 것에 초점이 맞춰져 있다는 것이다. 사회적 문제 이슈를 해결하기 위해서는 원인을 정확하게 파악하고 원인에 따른 대응책을 마련해야 한다. 하지만 기존의 연구들은 이슈의 원인을 파악하기보다는 이슈를 설명하는 것에 그치고 있다. 이러한 한계점을 극복하고자 한 연구들 중 이민철(2019)의 연구에서는 토픽모델링 기법을 활용하여 뉴스에서 등장한 사건들의 선후 관계를 분석하고자 하였으나 토픽모델링 기법이 가지는 한계점인 분석 결과로 나타난 토픽들에 대해 연구자가 자의적으로 해석하여 인과관계를 파악해야 한다는 단점이 존재한다[15]. 이러한 한계점을 극복하기 위해서는 텍스트로부터 인과지식을 추출할 수 있어야 한다. 이와 관련하여 기존 자연어 처리분야에서는 인과관계마이닝(causality mining)에 대한 연구들이 이루어져왔다 [16]. 인과관계마이닝은 대량의 텍스트로부터 인과지식을 추출하기 위한 방법으로 추출된 인과지식은 문제해결을 위한 의사결정과정에서 필요한 정보를 제공해주는 역할을 한다[17]. 하지만 인과관계마이닝은 사용되는 언어 및 각 언어별 문법의 특성 등으로 인해 분석이 어려운 한계점이 존재한다[18]. 특히 한국어의 경우 인과관계마이닝에 대한 연구가 거의 이루어지지 않고 있다.

한편, 사회이슈는 기간에 따라 동적으로 변화하며, 이러한 변화가 다른 이슈들을 발생시키거나 소멸시키는 등 이슈들 사이의 연관성을 가진다[19]. 따라서 사회 이슈를 보다 정확하게 예측하기 위해서는 이슈들이 어떠한 원인에 의해 발생하였으며, 해당 이슈가 다른 이슈에 어떻게 영향을 미치는 지 등 이슈들의 생성, 병합, 소멸의 이슈 생명주기에 대한 분석이 함께 이루어져야 한다. 이러한 이슈의 생명주기가 의미하는 것은 어떠한 이슈의 원인이나 결과요인들이 사회적으로 드러나서 이슈화가 되지 않더라도 향후 미래에는 이러한 요인들로 인해 새로운 이슈가 발생할 수 있다는 것이다. 이러한 개념에서 이루어지고 있는 연구가 미래신호연구이다. 여기서 미래신호란 미래에 일어날 수 있는 변화에 대한 징후이다. 이러한 미래신호는 현재에는 확인이 잘 되지 않는 신호이나 향후 시스템의 변화를 유도하는 요인을 의미한다[20, 21]. 이와 같은 미래신호는 현재의 사회적 현상의 원인이나 사회적 현상이 만드는 결과가 될 수 있다. 따라서 사회이슈에 대한 정확한 분석 및 미래 대응을 위해서는 현재의 사회이슈에 대한 인과요인을 분석하는 것이 필요하다.

3. 인과문형 기반 사회이슈 원인 요인 추출방법

앞의 2장에서 언급한 것과 같이 사회이슈를 보다 정확하게 분석하고 적절하게 대응하기 위해서는 이슈의 발생원인을 보다 정확하게 파악할 수 있어야 한다. 따라서 본 연구에서는 사회이슈 관련 텍스트 데이터로부터 해당 이슈의 원인 및 결과가 되는 요인을 문장의 연결어미를 활용하여 탐색하는 방법을 제안하였다. 본 연구에서 이슈의 인과요인을 탐색하기 위한 프로세스는 다음과 같다.

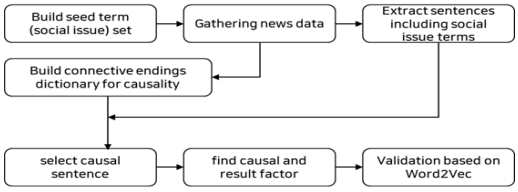

Fig. 1. Overall process for extract causality

3.1 데이터 수집을 위한 사회 이슈 정의 및 문장 추출

본 연구에서는 사회 이슈 데이터를 수집하기 위해서 뉴스데이터 검색을 위한 기본 검색 시드 단어(seed term) 세트를 세계경제포럼(WEF)에서 발표하는 글로벌 리스크(2009년부터 2018년)를 활용하여 총 30개의 검색 시드 단어를 선정하였다[22]. 아래의 Table 1은 선정된 30개의 단어들을 보여주는 그림이다.

Table 1. Seed terms for gathering text data

Category	Risks
Environment	Extreme climate change related weather
	Droughts and desertification
	Water scarcity
	Cyclone
	Earthquake
	Flooding
	Air pollution
	Biodiversity loss
	Unprecedented geophysical destruction
	Rising greenhouse gas emissions
	Irremediable pollution
	Man-made environmental catastrophes
Social	Pandemic
	Infectious disease
	Chronic disease
	Migration
	Demographic challenges
	Economic disparity
	Food security
	Water security
	Mismanagement of population aging
	Food shortage crises
	Ineffective illicit drug policies
Technology	breakdown of critical information system
	Data fraud/loss
	Online data and information security
	Threats of new technology
	Massive digital misinformation
	Failure of intellectual property regime
	Cyber attacks

3.2 뉴스데이터 수집 및 정제

데이터 수집을 위한 시드 단어 세트가 완성되면 다음으로는 시드 단어와 연관된 뉴스데이터를 수집한다. 시드 단어의 원인 요인을 탐색하기 위한 인과문형 패턴 생성을 위해 시드 단어가 포함되지 않은 문장을 제외하여 분석을 위한 최종 데이터 셋을 구축한다. 시드 단어의 원인 요인을 탐색하기 위한 인과문형패턴을 탐색하는 데 있어 시드 단어와 관련 없는 문장이 다수 포함될 경우, 패턴 탐색에 어려움이 있기 때문에 사전에 시드 단어가 포함되어 있지 않은 문장은 분석을 위한 데이터셋에서 제외하였다.

3.3 인과문형패턴 탐색 및 인과단어 도출

다음으로 문장 내에 인과문형패턴을 탐색하는 과정을 거친다. 이를 위해서 본 연구에서는 문장 내 단어들의 형태소를 분석한 후 인과관계를 표현하는데 사용되는 연결어미가 포함된 문장만을 추출한 후 이를 분석하여 인과문형패턴을 도출하였다. 여기서 인과관계란 두 개의 사건이 원인과 결과의 관계로 묶이는 것을 말한다[23]. 더 구체적으로 살펴보면, 두 개의 사건 중 선행 사건이 후행 사건의 원인, 조건, 목적의 관계에 있는 것을 의미한다. 예를 들어, '중국에서 날아온 미세먼지로 인해 대기오염이 증가하고 있습니다.' 라는 뉴스기사의 한 문장을 분석해보면, 대기오염의 원인으로 미세먼지를 언급하고 있는 것을 알 수 있다. 이때 미세먼지가 원인임을 확인할 수 있는 중요한 단서는 '인해'이라는 단어이다. 해당 단어의 형태소를 살펴보면 '인하'라는 동사와 '어/아'라는 연결어미가 합쳐진 단어임을 확인할 수 있다. 따라서 어떠한 사건의 원인을 설명하는 동사와 선행절과 후행절 사이의 조건, 결과, 인과관계를 나타내는 연결어미가 조합된 단어를 문장에서 찾을 수 있다면 해당 문장 내에서 원인과 결과에 해당하는 단어를 탐색하는 것이 가능하다. 즉, 동사+인과연결어미 형태의 단어는 인과관계를 탐색하기 위한 단서단어의 역할을 한다.

연결어미는 문장 내 선행절과 후행절 사이의 의미관계를 나타내는 기능을 하며, 선행절과 후행절 사이의 조건, 양보, 대립, 목적, 결과, 인과, 나열, 선택, 시간, 상황, 전환 관계를 나타내는 데 사용된다[24, 25].

이를 바탕으로 본 연구에서는 인과문형패턴을 찾기 위해서 문장을 형태소 분석을 수행한 후 '명사(원인)', '동사+연결어미(단서단어)', '명사(결과)'로 구성된 벡터로 변환

하였다. 예를 들어, 위의 뉴스기사 문장의 경우에는 〈중국(명사), 날아온(동사+연결어미), 미세먼지(명사), 인해(동사+연결어미), 대기오염(명사), 증가하고(동사+연결어미)〉로 변환된다. 이후 해당 벡터 값을 〈선행단어, 동사+연결어미, 후행단어〉의 형태를 만족하는 부분만 추출하여 재구성하게 된다. 따라서 앞의 〈미세먼지, 인해, 대기오염, 증가하고〉 벡터는 〈미세먼지, 인해, 대기오염〉으로 재구성된다. 최종적으로 이렇게 만들어진 벡터를 바탕으로 중간 부분의 '동사+연결어미' 단어가 인과문형패턴을 찾는 중요한 단서 단어인지를 검증한다.

다만 연결어미 중 결과관계를 표현하는 연결어미 경우에는 선행단어가 후행단어의 결과가 되기 때문에 이 경우에는 선행단어와 후행단어의 위치를 교환하여 벡터를 완성한다. 다음의 그림은 위의 과정을 도식화한 그림이다.

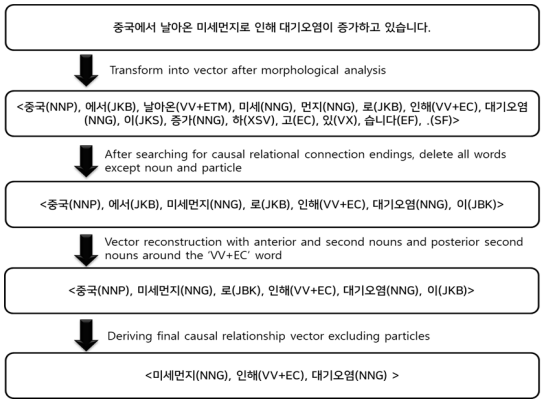

Fig. 2. Example process for building Causality vector

3.4 시드단어와 연관도가 낮은 인과관계 단어 필터링

다음으로는 word embedding 기법을 활용하여 시드단어와 도출된 인과단어를 필터링 하는 단계를 거친다. 이는 동사+연결어미 단어를 활용하여 도출된 인과단어의 정확성을 보다 높이기 위한 단계로 기존 텍스트 마이닝 분야에서 많이 사용되는 word2vec 기법을 활용하여 두 단어 사이에 연관성이 높은 단어인지를 확인하는 것이다.

본 연구에서는 word2vec기법을 활용하여 두 단어 사이의 연관도 값이 특정 값 이상으로 높은 경우 두 단어(시드단어와 도출된 인과단어)를 최종적으로 인과관계에 있는 단어로 확정하였다. word2vec 기법은 기존의 다양한 텍스트마이닝 분야의 연구에서 단어들 사이의 연관도를 측정하기 위해서 사용되어 온 대표적인 word embedding 기법으로 단어를 벡터로 표현한 뒤 특정 단어들 사이의 의미적 거리를 계산하여 단어 간 연관도를 유추하는 기법이다[26].

4. 실험 결과

4.1 데이터 수집 및 분석용 데이터 변환

본 연구에서는 시드단어들과 관련한 뉴스데이터 수집을 위해서 HTML Parser인 BeautifulSoup을 이용하였으며, python언어를 이용하여 뉴스데이터를 자동으로 수집하는 프로그램을 개발하였다. 본 연구에서는 제안하는 방법론의 가능성을 확인하기 위해서 환경 이슈가 되는 환경 문제인 '대기오염' 리스크에 대해 2017년 1월부터 2019년 12월까지의 3년 동안의 국내 주요 일간지 11개사의 뉴스기사 데이터(총 25,029 건)를 수집하였다. 뉴스데이터는 한국언론진행재단에서 제공하는 BigKinds 사이트에서 수집하였다. 아래의 그림은 뉴스데이터 수집을 위해 개발한 크롤러를 통해 수집되는 데이터의 일부 화면이다.

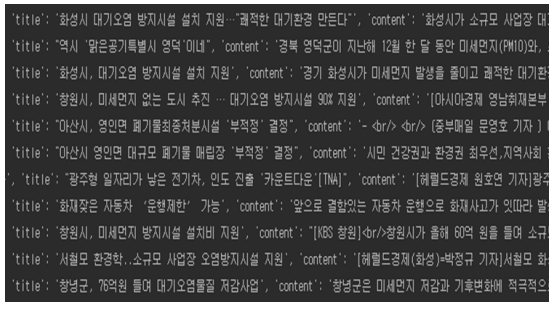

Fig. 3. Screen of news data crawler

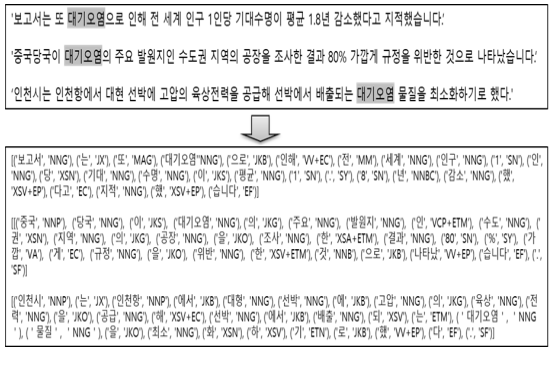

Fig. 4. Terms vector including seed term

데이터 수집 이후 특수문자, 기호 등 본 연구에 불필요한 불용어를 제거하는 과정을 거친 후 정제된 뉴스데이터 중 시드단어가 포함된 문장만을 추출하였다. 추출된 문장 데이터를 형태소 분석 라이브러리인 Mecab을 이용하여 각 단어별 품사를 태깅한 벡터의 형태로 변환하였다. 아래의 그림은 수집한 뉴스데이터 중 대기오염이 포함된 문장을 품사가 태깅(tagging)된 문장의 벡터로 변환한의 예이다. 여기서 등장하는 알파벳은 국어 품사를 의미하는 영어 약어이다.

4.2 인과문형패턴 기반 인과관계 단어 도출

각 단어별 품사가 태깅 된 문장단어벡터들을 바탕으로 본 연구에서는 인과관계에 있는 단어들을 추출하였다. 이를 위해서 앞의 3.3절에서 언급한 것과 인과, 결과, 목적의 관계를 나타내는 연결어미를 기반으로 문장 벡터에서 인과관계를 표현할 것으로 예상되는 부분을 추출하였다. 이때 연결어미는 동사와 함께 사용이 되기 때문에 동사+연결어미로 구성된 단어의 형태로 나타난다. 이를 본 연구에서는 인과관계를 나태나는 단서단어라고 정의하였다. 단서단어를 이용하여 아래 그림 5와 같은 인과문형패턴을 만들었다. 그림5에서 '인해', '따라', '의해' 등의 'VV(동사)+EC(연결어미)' 에 해당하는 부분의 단어가 단서단어이다.

```
re.compile("\(.*, 'NNG'\), .*, \('인해', 'VV\+EC'\), .*, \(.*, 'JKO'\)+")
re.compile("\(.*, 'NNG'\), .*, \('따라', 'VV\+EC'\), .*, \(.*, 'JKO'\)+")
re.compile("\(.*, 'NNG'\), .*, \('의해', 'VV\+EC'\), .*, \(.*, 'JKO'\)+")
re.compile("\(.*, 'NNG'\), .*, \('일으켜', 'VV\+EC'\), .*, \(.*, 'JKO'\)+")
re.compile("\(.*, 'NNG'\), .*, \('의해서', 'VV\+EC'\), .*, \(.*, 'JKO'\)+")
re.compile("\(.*, 'NNG'\), .*, \('의하여', 'VV\+EC'\), .*, \(.*, 'JKO'\)+")
```

Fig. 5. Causality patterns

이후 인과문형패턴을 이용하여 [단어 + 단서단어 + 시드단어]의 형태이면 시드단어의 원인 단어, 그리고 [시드단어 + 단서단어 + 단어]의 형태이면 시드단어의 결과 단어로 판단하여 시드단어의 후보 인과관계 단어세트를 구성하였다. 아래의 Table 2는 분석 결과 도출된 후보 인관단어 세트의 일부이다.

Table 2. Sample Word association values

Causal Factors	Clue words	Results Factors
pollution, polluter, polluter pays principle	따라	household waste, sewer, traffic congestion, air pollution, causing pollution, tourist, environmental conservation, contribution
the latest, ultrafine dust	인해	air pollution, concerns of pollution, event, green area
world, climate change	인해	Air pollution, records, flood, long term, drought, forest fire, foods, problems, grains, prices, virus, spread, humanity, reduction in life, politics, stability, riot
air pollution	의해	death, outdoors, indoors, pollutant
global warming	따라	heat wave, heavy rain, heavy snow, drought, flood, suffering, water, air pollution, world, human, life
Atomic electricity	인해	air pollution, deterioration of pollution, used nuclear fuel, neglect, safety of the people
company, collaboration	따라	electricity, truck, adoption, diesel truck, generated nitrogen, nitrogen oxide, downtown, air pollution
air pollution	따라	Replacement of boiler, replacement project, atmospheric environment, environment improvement, expectation effectiveness

4.3 Word2vec을 이용한 인과단어 필터링

위의 Table 2에서 나타난 것과 같이 후보인과관계단어는 인과문형패턴을 기반으로만 추출한 단어이기 때문에 실제 시드단어와 의미적으로 연관성이 낮은 단어들이 포함될 가능성이 있다. 예를 들어, 위의 Table에서 '환경오염', '원인자', '부담원칙'의 단어는 인과문형패턴에서는 원인 단어로 나타났으나, 이중 '원인자', '부담원칙'의 단어는 대기오염과는 의미적으로 연관성이 낮은 단어들이다. 따라서 본 연구에서는 인과문형패턴으로 나타난 단어들을 바탕으로 실제 의미적 연관성이 있는지를 확인하기 위해서 word2vec 알고리즘을 이용하여 시드단어인 '대기오염'과의 연관도점수를 계산한 후 최종 인과단어로 확정하였다. 실제 word2vec 분석 결과 '원인자', '부담원칙'의 단어의 연관도 점수 값은 각각 0.220, 0.287로 '환경오염'의 0.665보다 낮게 나타나 '대기오염'에는 상대적으로 연관도가 낮은 것으로 판단된다. 여기서 연관도 점수 값은 두 단어 사이의 의미적인 거리를 나타내는 점수로 -1에서 +1 사이의 값을 가질 수 있다. +1에 가까울수록 두 단어가 서로 의미적으로 연관이 되어 있음을 나타내는 값이며, 0은 서로 관련이 없음을 의미하고, -1은 서로 반대의 의미를 가지는 단어를 나타낸다.

아래의 Table 3은 실험에 이용한 시드단어인 '대기오염'과 인과문형패턴을 통해 나타난 후보인과단어들 사이의 연관성 점수를 계산한 결과 연관도 점수가 높은 상위 5개의 단어와 상대적으로 점수가 낮은 하위 3개의 단어를 나타낸 Table이다. 아래의 Table에서 '초미세먼지', '고밀도', '기후변화' 등의 단어들은 대기오염 단어와 상당히 높은 연관도를 나타내는 단어임을 의미한다.

Table 3. Sample Word association values

Causes	Results
Environmental pollution(0.665)	atmosphere environment(0.539)
atmosphere environment(0.585)	ozone(0.532)
ultrafine dust(0.557)	pollutant(0.521)
high desity(0.494)	density(0.442)
climate change(0.479)	carbon dioxide(0.435)
...	...
polluter pays priciples(0.287)	replacement project(0.338)
polluter(0.220)	indoor(0.227)
replacement of boiler(0.132)	Replacement of boiler(0.132)

이와 같이 인과문형패턴으로 도출된 인과 단어들에 대해서 word2vec 기법을 적용하여 서로 연관성이 낮은 단어들을 필터링 하였다.

최종적으로 본 연구에서는 나타난 원인, 결과 단어들의 word2vec 연관도 점수를 바탕으로 아래와 같이 시드단어와 연결된 원인 단어, 결과 단어 네트워크를 도출하였다. 이를 바탕으로 특정 사회이슈가 어떤 원인에 의해 나타나고, 해당 사회이슈의 결과가 어떻게 나타나는지를 쉽게 파악할 수 있다. 아래 네트워크에서 시드단어와의 거리는 시드단어와의 인과관계의 크기를 의미한다.

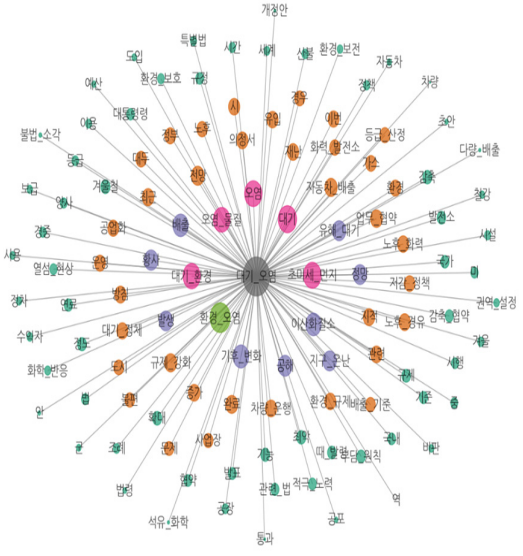

Fig. 6. Cause terms network

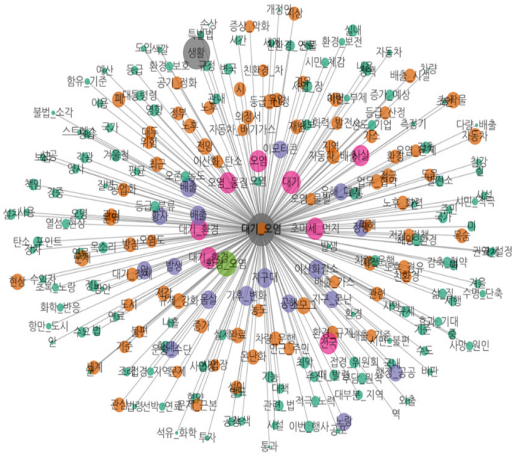

Fig. 7. Result terms network

4.4 시간에 따른 원인단어의 변화

본 연구에서는 위의 방법을 이용하여 2017, 2018, 2019년 3개년의 시간에 따른 인과요인의 변화를 비교해 보았다. 아래의 Table 4는 각 연도별 '대기오염'의 원인 단어를 비교한 Table이다.

Table 4. Cause factors by year

2017년도		2018년도		2019년도	
단어	연관도	단어	연관도	단어	연관도
온난	0.529	규제_강화	0.498	배출	0.547
공업화	0.495	유종	0.492	대기_환경	0.434
황사	0.451	다량_배출	0.477	정망	0.424
정차	0.433	대두	0.461	자동차_배출	0.375
기후_변화	0.414	국내_대기	0.450	문제	0.350
규제	0.382	과밀화	0.447	업무_협약	0.337
노후_화력	0.380	황사	0.419	환경	0.335
비판	0.363	유해_대기	0.416	저감_정책	0.325
노후	0.358	오염_물질	0.413	이산화질소	0.325
역	0.354	협약	0.403	증가	0.320
협약	0.350	초미세_먼지	0.401	열섬_현상	0.319
감축	0.342	공해	0.396	운영	0.311
화력_발전소	0.317	대기	0.395	시행	0.311
확대	0.316	환경_규제	0.394	사업장	0.292
개정안	0.314	오염	0.393	의정서	0.282
노력	0.313	기후_변화	0.385	차량_운행	0.281
국가	0.306	규정	0.367	도시	0.272
적극	0.296	유입	0.366	확대	0.268
지구	0.285	산불	0.362	가스	0.263

2017년도에는 대기오염의 증가 원인으로 '황사, '화력발전소', '노후' 등의 단어가 나타났으며, 2018년도에는 대기오염 발생원인 뿐만 아니라 대기오염을 감소시키는 원인 단어인 '규제강화', '협약', '규정' 등의 단어가 새롭게 등장한 것을 확인할 수 있었으며, 2017년도에는 없었던 '초미세먼지' 단어가 등장한 것을 확인할 수 있었다. 2019년도에는 국가 및 외교 정책과 관련한 단어인 '저감정책', '의정서', '업무협약' 등의 단어가 새롭게 등장하였고, 정책에 따라 시행되는 내용인 '차량운행', '운영', '자동차배출' 등의 단어가 새롭게 나타난 것을 확인할 수 있었다. 이 결과를 통해 대기오염이라는 사회이슈에 대해서는 시간 및 환경의 변화에 따라 해당 이슈의 인과요인에 대해서 다르게 인식하고 있는 것을 확인할 수 있었다. 이러한 결과는 어떠한 이슈가 현재 나타나고 있는가를 살펴보는 것도 중요하지만, 해당 이슈의 인과요인이 무엇인지를 함께 분석했을 때 보다 정확한 이슈에 대한 분석이 가능하며 이에 따른 효과적인 대응이 가능할 것이다.

5. 결론 및 제언

본 연구에서는 사회에서 나타나는 이슈를 보다 입체적으로 탐색하기 위해 특정 이슈의 원인 및 결과 요인을 뉴스데이터로부터 추출하는 방법론을 제안하였다. 이를 위해서 본 연구에서는 사회이슈와 관련한 뉴스데이터를 수집한 후 수집한 뉴스데이터의 각 문장들을 형태소 분석을 수행하여 인과요인을 담고 있는 문장들을 선별한 후 인과문형패턴을 도출하였다. 인과문형패턴은 인과관계를 나타내는 연결어미와 동사가 결합된 단서단어들을 바탕으로 도출하였으며, 도출된 인과문형패턴을 바탕으로 문장 내 원인 및 결과 요인을 추출하였다. 마지막으로 기존 텍스트마이닝 기법에서 많이 사용되는 word2vec 알고리즘을 이용하여 추출한 인과요인에 대한 보완적인 검증을 수행하였다.

본 연구에서 제안한 방법론의 의의는 다음과 같다. 우선 기존의 전문가 중심 Top-down 방식의 한계점인 정성적인 사회이슈 도출 방법이 가지는 객관성의 결여 및 시간적 비효율성을 극복할 수 있다. 전문가 중심의 사회이슈 도출 방법은 전문가의 경험이나 전문 분야에 따라 도출된 사회이슈에 대한 분석 결과가 달라질 수 있다. 이러한 문제점을 해결하기 위해서는 최대한 많은 수의 전문가로부터 의견을 받아야하기 때문에 시간적인 비효율성이 높아질 수 있다. 본 연구에서는 이러한 한계점을 텍스트마이닝 기법을 활용하여 자동화된 사회이슈도출방법을 제안하여 이러한 한계점을 극복할 수 있었다.

둘째, 본 연구에서는 기존 국어학의 품사이론을 바탕으로 인과요인을 도출하는 새로운 방법론을 제안하였다. 사회과학분야의 많은 실증 연구에서는 사회과학모형을 만든 후 각 독립변인과 종속변인 간 인과관계를 분석하는 연구가 많이 이루어져왔다. 하지만 이러한 연구의 한계점은 설문응답자를 기반으로 분석데이터를 수집하기 때문에 분석데이터의 수집에 한계점이 있고, 적합한 독립변인 및 종속변인, 응답데이터를 사용하지 않을 경우 원하는 분석결과가 도출되지 않는 한계점이 있다. 또한 기존 텍스트마이닝 기법에서는 각 단어가 가지는 형태소를 분석하고 특정 품사만을 이용하여 단어의 사용빈도나 동시출현빈도 등의 단어의 등장확률을 기반으로 텍스트 분석을 수행하여 단어 간 연관성을 추론하는 방법을 이용한다. 하지만 이러한 방법의 한계점은 단어 간 상관성에 대한 분석은 가능하나 단어의 선후 관계 즉 인과관계를 추론하는 것에는 한계를 가진다. 사회이슈에 대한 분석은 해당 이슈의 발생원인 및 해당 이슈로 인해 영향을 받아 나타나는 결과에 대한 예측 등이 중요한 분야인데, 이러한 부분에 있어서 기존의 텍스트마이닝 기법의 적용은 한계를 가진다. 본 연구에서는 국어학 분야의 품사이론을 적용하여 이러한 한계점을 극복하고자 하였다. 이를 통해 기존 텍스트마이닝 기반의 사회이슈 탐색 방법이 가지는 한계점인 단편적인 이슈의 탐색이 아닌, 이슈를 분석함에 있어 해당 이슈의 원인과 결과를 탐색하여 살펴볼 수 있는 입체적인 이슈의 탐색이 가능하였다.

셋째, 미래 이슈에 대한 실무적 차원의 대응이 가능하다. 본 연구에서 제안한 방법론을 통해 이슈의 인과요인을 도출할 수 있기 때문에 사회이슈에 대응하는 사회 및 관련 기업의 전략 수립에 있어서 가이드라인을 제시할 수 있을 것으로 생각한다. 또한 인과요인의 탐색을 통해 새로운 연구 분야를 제시할 수도 있을 것이다. 이슈와 관련된 인과 요인 중 현재 연구가 잘 이루어지지 않은 요인의 발견이 가능하며, 이는 연구자들에게 새로운 연구 주제를 제시할 수 있을 것이다.

본 연구가 가지는 한계점은 인과문형패턴을 도출하기 위한 단서단어 사전을 보다 확장할 필요가 있다는 점이다. 본 연구에서는 사회이슈 중 '대기오염'이라는 이슈를 바탕으로 데이터를 수집하였고, 수집된 데이터를 바탕으로 단서단어 사전을 구축하였기 때문에 단서단어들의 종류가 제한적이었다. 본 연구에는 단서단어를 바탕으로 인과단어들을 추출하는데, 단서단어들을 보다 많이 찾아 단

서단어 사전을 확장한다면 지금보다 더 많은 인과단어들을 추출할 수 있을 것이다. 따라서 향후 연구에서는 다양한 사회이슈들을 대한 데이터를 추가적으로 수집하여 단서단어 사전을 확장할 필요가 있다. 또한 본 연구에서는 추출된 인과 단어들을 word2vec 알고리즘의 단어 간 연관도 점수를 바탕으로 검증을 하였는데, 보다 성능이 좋은 단어 임베딩 기법들을 활용하여 검증의 정확성을 높일 필요가 있다. 향후 연구에서는 FastText 및 BERT, CNN 등 다양한 단어 임베딩 기법들을 비교하여 인과단어 추출의 정확성을 높일 수 있는 보완적인 연구를 수행할 것이다. 마지막으로 본 연구에서는 국내 주요 일간지의 뉴스기사를 수집하여 연구를 수행하였다. 다만 뉴스기사 자체가 가지는 신뢰성에 대한 부분은 검증하지 못한 한계가 있다. 향후 연구에서는 뉴스기사의 신뢰성에 대한 부분도 함께 고려하여 데이터를 수집한 후 분석을 수행하여 보다 정교하게 인과요인을 도출할 예정이다.

REFERENCES

[1] J. Y. Won & D. G. Kim. (2014). Deduction of Social Risk Issues Using Text Mining. *Crisisonomy*, 10(7), 33-52.

[2] Li. Z. Zhou, D. Juan. Y. F & Han. J. (2010, April). Keyword extraction for social snippets. I*n Proceedings of the 19th international conference on World wide web*, 1143-1144.

[3] Sakaki. T, Toriumi. F & Matsuo. Y. (2011). Tweet trend analysis in an emergency situation. *In Proceedings of the Special Workshop on Internet and Disasters*, 1-8.

[4] Wang. J, Liu. J & Wang, C. (2007, May). Keyword extraction based on pagerank. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining* , 857-864.

[5] W. J. Song. (2014). Significance and Tasks of R&D Projects for Solving Social Problems. *SCIENCE & TECHNOLOGY POLICY*, 24(2), 4-13

[6] M. S. Lim & N. G. Kim. (2014). Analyzing the Issue Life Cycle by Mapping Inter-Period Issues. *Journal of Intelligence Information Systems*, 20(4), 25-41.

[7] S. C. Back, S. H. Jo, N. H. Kim & K. S. Noh. (2017). A Study on the Process of Refining Ideas for Social Problem Solving Based on Design Thinking in Digital Convergence Era. *Journal of Digital Convergence*, 15(2), 155-163.

[8] H. J. Lee & J. S. Hwang. (2015). Service framework and process for solving social issues. *Review of Korean Society for Internet Information*, 16(1), 63-68.

[9] J. Heo, C. H. Lee, H. J. Oh, Y. C. Yoon, H. K. Kim, Y. H. Jo. & C. Y. Ock. (2014). Web Science : Automatic Generation of Issue Analysis Report Based on Social Big Data Mining. *Software and Data Eng*, 3(12), 553-564.

[10] J. H. Bae, N. G. Han. & M. Song. (2014). Twitter Issue Tracking System by Topic Modeling Techniques. *Journal of Intelligence and Information Systems*, 20(2), 109-122

[11] D. M. Jeong, J. S. Kim, G. N. Kim, J. U. Heo, B. W. On. & M. J. Kang. (2013). A Proposal of a Keyword Extraction System for Detecting Social Issues. *Journal of Intelligence Information Systems*, 19(3), 1-23.

[12] S. M. Lee & S. J. Ahn. (2020). Mass Media and Social Media Agenda Analysis Using Text Mining : focused on '5-day Rotation Mask Distribution System'. *Journal of The Korea Contents Association*, 20(6), 460-469.

[13] J. H. Ki & S, H, Ahn. (2020). Application of Sentiment Analysis and Topic Modeling on Rural Solar PV Issues : Comparison of News Articles and Blog Posts. *Journal of Digital Convergence*, 18(9), 17-27.

[14] S. R. Lee & E. J. Choi. (2020). Comparison of responses to issues in SNS and Traditional Media using Text Mining-Focusing on the Termination of Korea-Japan General Security of Military Information Agreement(GSOMIA)-. *Journal of Digital Convergence*, 18(20), 277-284.

[15] M. C. Lee & H. J. Kim (2018). Construction of Event Networks from Large News Data Using Text Mining Techniques. *Journal of Intelligence and Information Systems*, 24(1), 183-203.

[16] Li. P & Mao. K. (2019). Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, 115, 512-523.

[17] Girju. R. (2003). Automatic detection of causal relations for question answering. *In Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, 76-83.

[18] T. N. De Silva, X. Zhibo, Z. Rui & M. Kezhi. (2017). Causal relation identification using convolutional neural networks and knowledge based features. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 11(6), 696-701.

[19] M. S. Lim & N. G. Kim. (2016). Investigating Dynamic Mutation Process of Issues Using Unstructured Text Analysis. *Journal of Intelligence Information Systems*, 22(1), 1-18.

[20] Ilmola, L & Kuusi, O. (2006) Filters of weak signals hinder foresight: Monitoring weak signals efficiently in corporate decision making. *Futures*, 38(8), 908-924.

[21] Ansoff, I. H. (1975). Managing strategic surprise by response to weak signals, *California Management Review*, 18(2), 21-33.

[22] World Economic Forum Global Risks Report, http://reports.weforum.org/

[23] C. H. Lim. (2006). The formation process of causal relations and the connecting endings of Korean language. *In Proceedings of the conference on the discourse and cognitive linguistics society of Korea*, 151-164.

[24] E. G. Yi. (2000). *A Study on connective endings in Korean*. Thae Hak Sa.

[25] P. H. Yoon. (1992). *Study of Korean conjunctive endings*, Hansin MUNHWASA.

[26] Mikolov. T, Sutskever. I, Chen. K, Corrado. G. S & Dean. J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 3111-3119.

이 남 연(Namyeon Lee) [정회원]

· 2008년 2월 : 경희대학교 기술경영학과(경영학석사)
· 2013년 2월 : 경희대학교 국제경영학과(경영학박사)
· 2014년 4월 ~ 2017년 8월 : 성결대학교 파이데이아학부 조교수
· 2017년 9월 ~ 현재 : 한신대학교 IT경영학과 조교수
· 관심분야 : 빅데이터, 텍스트마이닝, 기술융합
· E-Mail : nylee@hs.ac.kr

이 재 형(Jae Hyung Lee) [학생회원]

· 2021년 3월 : 한신대학교 IT경영학과(학사과정)
· 관심분야 : 빅데이터, 텍스트마이닝, 데이터시각화
· E-Mail : eju2003@gmail.com