「정보시스템연구」 제31권 제4호
한국정보시스템학회
2022년 12월, pp. 25~45
http://dx.doi.org/10.5859/KAIS.2022.31.4.25

귀납적 사회과학연구 방법론을 위한 토픽모델링의 확장 및 사례분석*

김 근 형**

〈목 차〉

Ⅰ. 서 론

4차 산업혁명시대를 맞이하여 도처에서 많은 데이터들이 생성되고 있다. 기존의 정형 데이터뿐만 아니라 이미지 및 텍스트 등과 같은 비정형 데이터의 생산이 폭증하고 있다. 우리사회를 구성하는 사람들의 생각과 의견 등은 주로 텍스트 형태로 만들어진다. SNS 등을 통하여 다양한 주제에 대해서 수많은 사람들의 이채로운 생각과 의견들이 텍스트 형태로 생산되고 있다. 온라인리뷰(Online review)는 사회 구성원들의 생각과 의견이 인터넷 상에서 표출된 대표적인 텍스트 데이터이다. 수력발전이 물 자원을 효과적으로 관리하고 활용하여 소중한 전기에너지를 생산하듯이, 데이터자원을 제대로 관리하고 활용함으로써 창조적 가치를 창출하자는 개념이 '데이터 댐' 이다. 우리 사회 사람들이 자유롭게 생산하는 사진이미지 및 동영상뿐만 아니라 온라인리뷰 등과 같은 텍스트 데이터 등 빅데이터 자원을 제대로 분석하고, 이를 새로운 가치창출에 활용되도록 하는 것은 경제성장의 새로운 원동력이 될 수 있다.

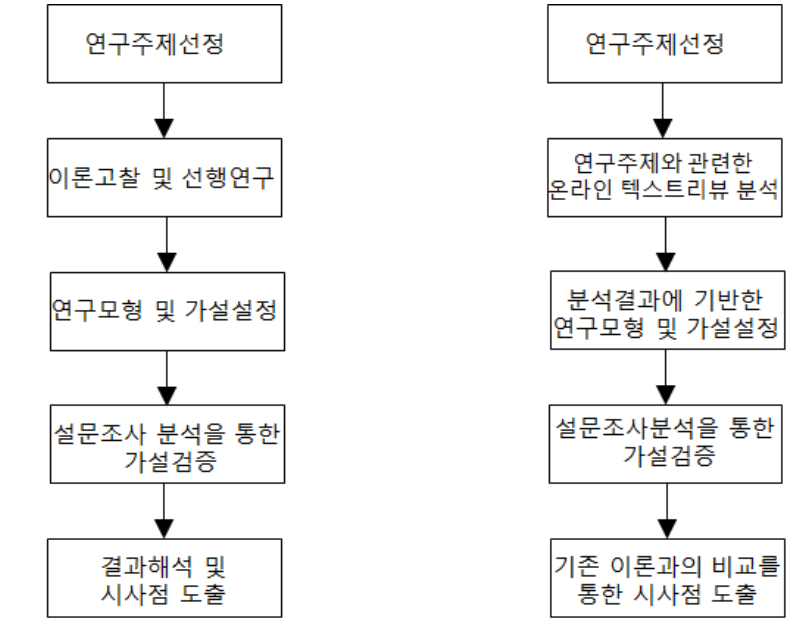
사회과학은 사람들로 구성된 우리 사회의 다양한 현상을 과학적으로 규명하여 새로운 지식

* 이 논문은 2022학년도 제주대학교 교원성과지원사업에 의하여 연구되었음
** 제주대학교 경영정보학과, khkim@jejunu.ac.kr(주저자 및 교신저자)

을 창출하는 학문분야이다. 기존의 사회과학 연구방법은 사람들의 의견과 생각을 알아내기 위하여 구조화된 설문지를 바탕으로 수집한 정형 데이터를 주로 활용하였다. 설문 데이터를 수집하기 전에 문헌고찰 등을 통한 연역적(deductive) 방법으로 연구가설을 도출하였다. 수립된 가설의 과학적 검증 차원에서 설문조사에 의해 수집된 데이터는 가설검증 목적으로 분석되었다. 연역적 방법에 의한 가설도출은 이론적 엄격성과 정합성의 관점에서는 강건할 수 있으나, 혁신적이고 다양한 가설을 도출하는 데는 한계가 있을 수 있다.

사람들의 생각이나 행동방식을 기록한 텍스트 데이터는 사회과학을 위한 중요한 원천자료이다. 온라인 쇼핑몰 등과 같은 웹사이트에 게시된 대량의 온라인리뷰들은 많은 리뷰어(reviewer)들이 자유롭게 작성한 텍스트 데이터의 일종이다. 특정 리뷰어가 작성한 온라인리뷰는 그 리뷰어에 대한 설문조사와 대응될 수 있다. 온라인리뷰와 같은 비정형 텍스트 데이터를 텍스트마이닝 등의 방법으로 분석하면 텍스트의 일관된 패턴을 발견할 수 있으며, 이를 바탕으로 사회현상을 나타내는 새로운 가설을 도출할 수도 있다. 데이터를 기반으로 가설을 도출하기 때문에 귀납적(inductive) 가설이라 할 수 있다. 이미 생성된 대량의 텍스트 데이터 안에 내재된 반복패턴으로부터 새로운 가설을 도출하는 개념이 본 논문에서 제안하고자 하는 귀납적 사회과학연구방법이다.

본 논문에서 제안하고자 하는 확장형 토픽모델링 기반의 새로운 사회과학연구방법은 먼저 연구주제와 관련한 텍스트를 수집하여 분석한



(a) 설문조사 기반의 연역적 사회과학방법론

(b) 토픽모델링 확장 기반의 귀납적 사회과학방법론

<그림 1> 사회과학연구방법의 비교

후 이를 기반으로 가설을 수립하는 형태의 귀납적 방법이다. 데이터가 넘쳐나는 시대에 텍스트 기반의 귀납적 연구방법이 필요하며 그 중요성이 커질 수 있다. <그림 1>은 전통적인 사회과학방법론과 본 논문에 제안하고자 하는 귀납적 사회과학방법론을 비교하여 나타내고 있다.

기존의 토픽모델링 기법은 텍스트문서집합에 대한 유용한 요약정보를 생성할 수 있음에도 불구하고, 사회과학 지식을 창출하기 위한 도구로서는 그 한계가 있다. 지식의 기본구조는 인과관계를 규명할 수 있는 것이어야 할 것이기 때문이다.

본 논문에서는 온라인리뷰와 같은 텍스트문서집합으로부터 인과관계를 내포한 정보를 생성할 수 있도록 토픽모델링 기법을 확장하고자 한다. 토픽모델링 기법을 확장하여 인과관계 분석을 할 수 있고 통계적 유의성을 확인할 수 있는 새로운 분석기법을 개발한 후, 그 효과성을 검증하기 위한 실제 분석사례를 제시하고자 한다. 이를 위하여, 기존 토픽모델링 방법 중의 하나인 LDA(Latent Dirichlet Allocation)알고리즘의 결과에 후처리(Post-processing)과정을 추가함으로써 토픽모델링을 확장한다. 토픽변수들 끼리 또는 토픽변수와 다른 외적변수와의 인과관계를 설정할 수 있도록 하며, 인과관계의 통계적 유의성을 검정할 수 있도록 하여, 새로운 가설을 수립할 수 있는 분석방법을 제시하고자 한다.

예를 들면, 수천 건의 성산일출봉 온라인리뷰 텍스트문서집합에서 '먹거리 관광'과 '경치 관광' 이라는 토픽이 추출되면, 각 온라인리뷰의 평점을 만족도로 설정하여 <그림 2>와 같은 형태의 연구가설을 설정할 수 있다.

<가설1> 성산일출봉에 대한 먹거리 관광은 관광객의 만족도에 유의한 영향을 미친다.
<가설2> 성산일출봉에 대한 경치관광은 관광객의 만족도에 유의한 영향을 미친다.

<그림 2> 인과관계를 내포한 정보의 예

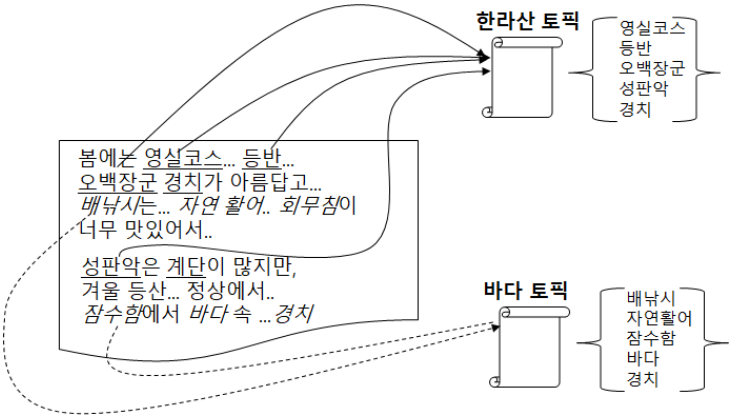
Ⅱ. 이론적 배경

2.1 토픽모델링

토픽모델링(topic modeling)은 대량의 문서집합 속에서 토픽(topic)을 찾는 텍스트분석 기법이다(Hofmann T., 1999). 텍스트분석의 관점에서 볼 때, 토픽은 단어주머니(bag-of-words)로 정의되며, 특정단어들이 반복적으로 함께 등장하는 패턴을 나타낸다. 토픽모델링의 결과물은 일반적으로 토픽을 구성하는 단어 목록이며, 이 단어 목록을 통해 문서 내에 존재하는 토픽들을 추론할 수 있다.

텍스트문서는 토픽의 혼합체(mixture of topics)로서 다양한 여러 토픽들로 구성되며, 각 토픽은 단어들로 구성되는 단어 혼합체이다.

<그림 3>은 제주관광에 대한 어떤 온라인리뷰 텍스트문서 내의 토픽들과 각 토픽 내의 단어목록을 나타내고 있다. 하나의 리뷰텍스트에는 '한라산 토픽'과 '바다 토픽'이 혼재되어 있다. '한라산 토픽'에 속하는 단어목록은 '영실코스', '등반', '오백장군', '경치' 등이며, '바다 토픽'은 '배낚시', '자연활어', '잠수함', '경치' 등의 단어를 포함한다. '경치' 와 같은 단어는 여러 토픽에 중복되어 나타나고 있다. 특정 단



<그림 3> 문서 내 토픽들과 단어목록

(1) 토픽의 개수 K 결정
(2) 문서 D에 포함된 각 단어 W를 임의적으로 K개 토픽 가운데 하나에 임의적으로 할당
(3) for(문서집합 내 각 단어 W에 대하여 토픽할당의 변화가 없을 때까지 반복)
  (3)-1. P(토픽T|문서D) 계산 // 문서D에서 토픽T가 차지하는 비율
  (3)-2. P(단어W|토픽T) 계산 //토픽T에 할당된 단어들 가운데 단어W가 차지하는 비율
  (3)-3. R = P(토픽T|문서D) × P(단어W|토픽T) 계산 //문서D에서 토픽T가 단어W를 생성한 비율
  (3)-4. 문서D에 포함된 각 단어W에 대하여 R의 확률로 소속 토픽 갱신
(4) 각 토픽에 할당된 단어들의 비율을 계산하여 주요단어 목록 추출
(5) 각 문서 내에서 각 토픽과 관련된 단어의 비율을 계산하여 문서 내 토픽들의 가중치 계산

<그림 4> LDA의 논리적 구조

어가 여러 토픽에 중복될 수 있다는 점은 토픽 모델링이 군집분석 개념과 다른 부분이다.

어떤 문서에 어떤 토픽들이 포함될지는 토픽의 확률분포에 따라 결정된다, 마찬가지로, 어떤 토픽에 어떤 단어들이 포함될지는 단어주머니에 포함된 단어의 확률분포에 따라 결정된다. 문서에 포함된 토픽들의 구성 비율 즉, 토픽의 확률분포와 각 토픽별로 특정 단어가 선택될 확률(즉, 토픽을 구성하는 단어의 확률분포)를 알고 있으면 특정 토픽들로 구성된 문서를 생성할 수도 있다.

LDA(Latent Dirichlet Allocation)는 대표적인 토픽모델링 알고리즘이다(Blei, D. et al., 2003). LDA를 이용한 토픽모델링은 문서집합 내 각 단어를 각 토픽에 할당하는 과정이라 볼 수 있다. 최적의 단어-토픽 할당을 통해 각 토픽과 높은 연관관계를 갖는 단어목록을 찾아내고 이를 통해 각 문서의 토픽구성을 파악한다. LDA에서는 토픽의 확률 분포와 단어 확률 분포를 추정하기 위한 사전 확률분포인 디리클레(Dirichlet)분포를 사용한다. 토픽 모델링에서 디리클레 분포가 사용되는 이유는 디리클레 분포가 다른 다항분포함수와 곱하면 다시 디리클레 분포 형태가 되어 관찰된 단어를 활용한 사후분포를 만들 때 유리한 점이 있기 때문이다. <그림 4>는 LDA의 논리적 구조를 나타내고 있다.

LDA 알고리즘에서 토픽 수 K는 수동적으로 설정된다. 토픽수를 너무 적게 설정하면 토픽에

너무 많은 단어가 포함되어 토픽 구분이 어려워질 수 있는 반면, 높게 설정하면 의미 없는 토픽들이 도출될 수 있다(이병철, 김두산, 2020). 최적의 토픽 수는 응집도(Coherence)와 혼란도(Perplexity)를 사용하여 결정할 수 있다. 응집도는 한 토픽 안에 의미론적으로 유사한 단어가 얼마나 많이 모여 있는지를 나타내는 척도이다(Newman et al., 2010). 응집도가 높을수록 토픽 내 단어들 사이의 의미론적 일관성은 높다고 본다. 혼란도는 특정 확률 모델이 실제로 관측되는 값을 얼마나 잘 예측하는지 평가할 때 사용된다. 혼란도가 작을수록 토픽모델링 결과가 실제 문서집합을 잘 반영한 것으로 본다(Chang et al., 2009).

2.2 선행 연구

온라인리뷰는 사회 구성원인 사람들의 생각과 의견이 인터넷 상에서 표출된 대표적인 텍스트 데이터이다. 온라인리뷰는 사용자의 솔직한 경험과 의견을 자발적으로 표현한 글이기 때문에 사회과학연구를 위한 중요한 데이터가 될 수 있다(김근형과 오성열, 2009). 온라인리뷰를 분석하기 위해서는 텍스트마이닝 기법이 주로 적용된다. 온라인리뷰는 텍스트 문서집합이기 때문이다. 텍스트마이닝의 다양한 영역 중에서 오피니언마이닝(Opinion mining)이나 의미연결망분석(Semantic Network Analysis), 토픽모델링(Topic modeling) 기법, Word2Vec 기법 등이 온라인리뷰 분석에 주로 활용되어져 왔다.

오피니언마이닝은 온라인리뷰에 포함된 형용사 및 동사 단어들의 긍정성과 부정성 등을 판별하고 그 빈도수 등을 계산하여 전체적인 긍정, 부정 정도를 판단하는 기법이다(김근형, 2011). 이창용(2021)의 연구에서는 오피니언마이닝을 확장하여 온라인리뷰의 평점과 길이와의 인과관계를 분석하고 있다. 그러나 한글문장에서는 동사 및 형용사에 대한 긍정성과 부정성을 정리한 단어사전이 제대로 구축되지 않아 오피니언마이닝의 정확성에 장애가 되고 있다.

의미연결망분석은 텍스트문장에서 동시 출현하는 단어들의 빈도를 계산하여 그 강도를 연결함으로써 테스트를 요약하고 이해하고자 하는 텍스트마이닝 기법중의 한 유형이다(Philipp Drieger, 2013). 의미연결망분석은 다른 텍스트마이닝 기법과 함께 보조적인 추가분석에 많이 활용된다.

토픽모델링은 온라인리뷰에서 주로 언급되는 주요 토픽들을 추출함으로써 문서집합 내에 내포된 핵심주제들을 파악할 수 있도록 한다(Blei, D. et al., 2003). 추출된 각 토픽들에는 주요단어들이 할당된다. 분석자는 각 토픽에 할당된 주요단어들을 바탕으로 토픽의 의미와 명칭 등을 결정한다. 그러나 주요단어들은 여러 개의 토픽에 중복 할당될 수 있어서 토픽의 의미를 파악하는데 장애가 될 수 있다.

Word2Vec(word embedding to vector)은 인공신경망(Artificial Neural Network)기법을 이용한 단어임베딩 학습모델로서, 단어들의 의미를 특정 차원의 벡터공간 모델에서 값으로 계산하고 표현하는 학습기법이다(Mikolov et al., 2013). 유사한 문맥분포를 가진 단어들은 유사한 의미를 가질 것이라는 아이디어를 바탕으로 문맥정보와 중심단어를 활용하여 유사단어를 학습하는 방식을 취한다. 다양한 도메인에서 Word2Vec를 활용하여 분석하는 많은 연구들

이 있다(심영석 등, 2018; 이민철과 김혜진, 2018).

다양한 도메인(Domain)에서 토픽모델링을 활용하여 텍스트 데이터를 분석한 많은 연구들이 있다. <표 1>은 토픽모델링 기법을 활용한 연구들을 나타내고 있다.

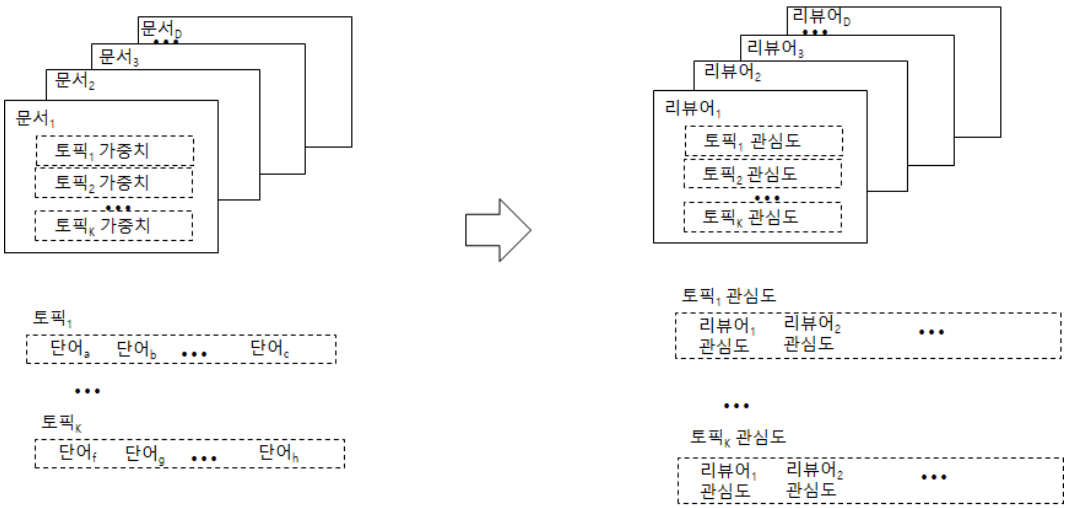
Ⅲ. 연구설계

3.1 토픽모델링의 확장개념

온라인 쇼핑몰 등과 같은 웹사이트에 게시된 대량의 온라인리뷰들은 많은 리뷰어들이 작성한 텍스트문서집합의 일종이다. 특정 리뷰어가 작성한 온라인리뷰 하나가 텍스트문서 하나에 대응된다. 리뷰어가 작성한 온라인리뷰 내 각 토픽 가중치는 그 토픽들에 대한 리뷰어의 관심도와 유사한 측면이 있다. 온라인리뷰들을 토픽모델링 기법으로 분석하여 도출한 각 리뷰 내 토픽들의 가중치는 그 토픽들에 대한 리뷰어의 관심도가 정량적으로 측정된 것이라 볼 수 있다.

<표 1> 토픽모델링 기반의 연구 동향 요약

논문	주요내용
코로나19 확산에 따른 외식에 대한 소비자 인식 분석: 토픽모델링 및 네트워크 분석의 활용(신서영과 이범준, 2021)	코로나19 시대의 외식에 대한 소비자 인식을 분석하기 위하여 토픽모델링과 의미연결망 분석을 실시함. 텍스트 내 주요 단어들에 대한 빈도분석 등이 주를 이루고 있으며 인과모형 분석 없음
암생존자 삶의 질 영향요인에 대한 연구동향: 텍스트 네트워크 분석과 토픽모델링(박진희 외, 2021)	의미연결망 분석 및 토픽모델링을 이용하여 암생존자 삶의 질 영향요인에 관한 연구의 핵심 주제어를 파악하고, 주요 토픽별 네트워크의 특성을 확인하여 암생존자 삶의 질 영향요인 연구의 지식구조를 탐색하고 있음
국내 복합리조트 서비스품질에 대한 토픽모델링 및 감성분석(신주하 외, 2021)	제주 복합리조트를 경험한 해외 관광객들이 직접 작성한 온라인 리뷰를 바탕으로 관광객들의 만족에 영향을 미칠 수 있는 중요한 서비스품질 요소들을 토픽모델링을 활용하여 도출하고 있으며, 온라인 리뷰에 담긴 관광객들의 감성을 긍정, 부정, 중립으로 구분하고 있음
토픽모델링 분석에 의한 글로벌 그린뉴딜 연구동향 분석(유재호 외, 2021)	한국의 Green New Deal, 유럽의 European Green Dea, 미국의 Green New Deal과 관련된 다양한 학술 연구에 대하여 토픽 모델링을 활용하여 비교 및 분석하고, 주요 연구동향을 도출하고 있음
인공지능과 인간안보 : 토픽모델링 기법을 활용한 유럽의 인공지능 안보화 분석(황원준, 2021)	인공지능에 대한 유럽 내 언론 기사들의 안보 담론을 토픽모델링을 활용하여 주요 토픽들을 도출하고 있음
토픽모델링 분석 기법을 활용한 국내외 금융보안 분야 연구동향 분석(채호근 외, 2021)	금융보안 분야의 핵심 연구분야를 도출하고, 방향성을 제시하기 위하여 토픽모델링을 통해 국내외 주요 연구 동향 비교분석을 수행하고 있음
'좋아요'와 '싫어요'같은 간접적 사회적 정보의 방향과 강도는 온라인 뉴스 콘텐츠 댓글의 숙의의 질과 어떤 관련이 있는가? 토픽모델링을 이용한 토픽 다양성 분석(민진영 외, 2021)	숙의의 질을 측정하는 대표적 지표인 다양성에 초점을 맞추고, 다섯 개 의 뉴스 사이트에서 수집한 기사 및 댓글에 토픽 모델링방법을 적용하여 분석하고 있음



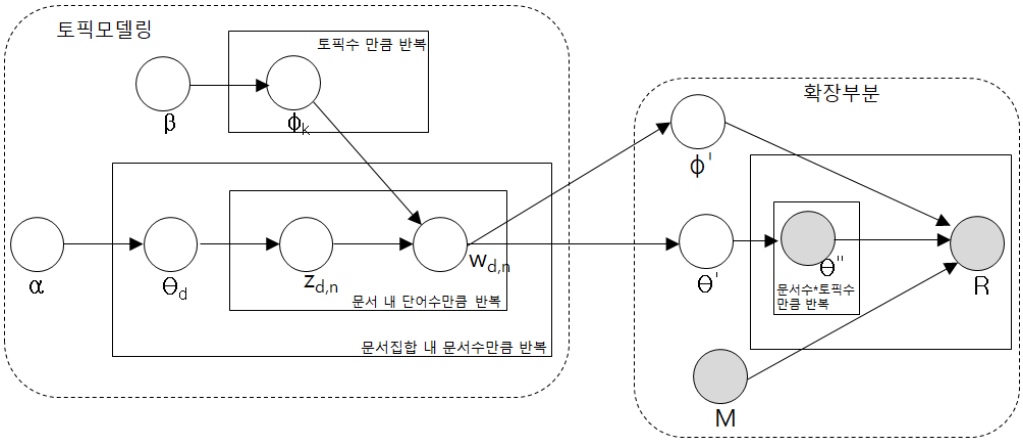
(a) 토픽모델링의 물리적 결과물 (b) 토픽모델링의 확장에 의한 결과물

<그림 5> 토픽모델링의 확장 개념

3.2 확장모형

토픽모델링 알고리즘의 확장 목적은 변수들 사이의 인과모형을 도출하고자 하는 것이다. 인과모형을 구성하기 위한 변수들은 토픽모델링의 결과물인 토픽들만으로 이루어질 수도 있고 추가적인 외적변수가 포함될 수도 있다. 인과모형의 독립변수는 토픽모델링에서 도출된 토픽들이 그 대상이 된다. 종속변수는 추가적인 외적변수가 도입되거나 기존 토픽들 중에서 선택될 수 있다. 외적변수는 텍스트 문서집합과 관련한 메타데이터(meta data)가 그 대상이 될 수 있다. 예를 들어, 온라인리뷰 집합의 경우, 각 텍스트리뷰의 평점 또는 텍스트리뷰의 작성날짜, 작성자의 인구통계 데이터 등을 외적변수로 설정할 수 있다. 독립변수와 종속변수가 결정되면 각 변수의 값들을 표준화된 데이터 규격으로 생성하여야 한다. 독립변수는 연속형 데이터이고 종속변수는 연속형 또는 범주형 데이터가 오도록 한다. 연속형 데이터는 5점 척도 데이터로 표준화한다. 토픽들에 대응하는 독립변수의 경우는 토픽 가중치값들이 5점 척도 데이터로 변환되어야 한다.

<그림 6>은 토픽모델링 아키텍처를 확장한 전체적인 구조를 나타내고 있다. <그림 6>에서 왼쪽 영역은 기존 토픽모델링 아키텍처(윤상훈과 김근형, 2021)이고 오른쪽 영역은 인과모형 도출을 위한 추가적인 처리 부분이다. 토픽모델링의 결과물로 Φ‘과 Θ‘을 얻을 수 있다. Φ‘은 k개의 토픽과 각 토픽별 관련 단어목록을 포함하는 2차원 벡터 데이터이다. Φ‘은 인과모형의 독립변수 또는 종속변수 역할을 할 토픽들의 목록을 포함하고 있다. M은 문서집합의 메타데이터로부터 도출한 외적변수로서, 종속변수를 설정하기 위하여 추가적으로 도입한 변수이다. 결국, 인과모형의 독립변수와 종속변수를 설정



<그림 6> 토픽모델링의 확장 아키텍처

<표 2> 수식의 기호 의미

기호	의미
Φ‘	토픽모델링의 결과로 도출된 2차원 토픽벡터로서, k개의 토픽과 각 토픽별 관련단어목록 포함
Θ‘	토픽모델링의 결과로 도출된 2차원 문서벡터로서, d개의 문서와 각 문서별 토픽비중을 나타내는 벡터 데이터 포함
Θ“	Θ‘를 5점척도 데이터로 변환한 2차원 벡터 데이터
M	온라인리뷰의 주변변수(평점, 날짜 등)
R	독립변수 목록과 종속변수(주변변수) 사이의 회귀방정식

하기 위한 변수들은 Φ‘과 M으로부터 선택할 수 있다. Θ‘은 문서집합의 각 문서에서 토픽들이 어떤 비중으로 포함되어 있는지에 대한 데이터를 포함하고 있는 2차원 벡터 데이터다. 특정 문서 내 각 토픽들의 비중 값들은 각 토픽변수의 변수값들이 된다. 특정 문서를 작성한 문서 작성자가 각 토픽들에 대한 관심도는 문서 내 그 토픽의 비중값과 대응된다. 결국, Φ‘ 이 변수 목록을 나타낸다면 Θ‘은 변수 값들을 포함하고 있다.

3.3 확장 알고리즘

아래 <그림 7>의 인과모형 도출알고리즘은 <그림 6>의 확장부분에 대한 처리과정을 구체화하여 알고리즘 형태로 나타내고 있다. 알고리즘의 입력 데이터는 2차원 벡터 데이터인 Φ’과 Θ’이고 출력 데이터는 인과모형을 나타내는 회귀방정식이다.

입력 데이터에서 Φ’p 는 2차원 벡터 Φ’의 1차원 구성요소를 의미한다. 즉, Φ’p는 k개의 토픽중의 하나인 토픽 p를 의미하는 것이며 wpi는 토픽 p가 포함하는 주요 단어들 중의 하나를 의미한다. 마찬가지로, Θ’q 는 Θ’의 1차원 구성요소를 의미하며 tqi는 문서 q 안에서 토픽 i의 가중치 값을 나타낸다. 각 문서 안에 포함된 토픽들의 가중치 값들을 모두 합하면 1이 되어야 한다. M은 필요한 경우 입력되는 데이터로서 문서집합의 메타데이터 중의 하나인 외적변수를 의미한다. 문서집합이 온라인리뷰인 경우, 리뷰문서와 대응되는 평점, 리뷰작성 날짜, 리뷰 작성

<그림 7> 인과모형 도출 알고리즘

입력: Φ‘, Θ’, M
Φ’ : 2차원 토픽벡터
Φ’ = {Φ‘1, Φ’2, Φ‘3, ... ,Φ’K}, K는 토픽 수
Φ’p = {wp1, wp2, wp3,..., wpn}, wpi는 토픽 p에 속하는 i번째 단어

Θ’ : 2차원 문서벡터
Θ’ = {Θ’1, Θ’2, Θ’3, ... ,Θ’d}, d는 문서 수
Θ’q = {tq1, tq2, tq3,..., tqn}, tqi는 q번째 문서에서 토픽 i의 비중

$$\sum_{i=1}^{K} t_{qi} = 1$$

M : 문서집합의 주변변수
: 평점, 날짜,

출력: R
R : 회귀방정식
R = Formula(M ~ Φ“) = (M ~ α1*Φ‘1 + α2*Φ‘2 + α3*Φ‘3+..., αK*Φ‘K+ β)

Begin
(1) Θ“ 초기화; // 문서별 토픽들의 5점척도 점수값들을 저장할 목적의 2차원 벡터 변수
x’ 초기화 ; // x’은 문서 내 토픽별 5점척도 값 저장 목적의 1차원 벡터 변수
y’ 초기화 ; // y’은 토픽 별 5점척도값 저장 목적의 스칼라 변수
(2) for x in Θ’ {
(3) interval = $\frac{\max(t_{ij}) - \min(t_{ij})}{5}$;
(4) for y in x {
if (y > 5*interval) then y’=5
else if (y > 4*interval) then y’=4
else if (y > 3*interval) then y’=3
else if (y > 2*interval) then y’=2
else y’ = 1;
(5) x’.insert(y’) ;
(6) Θ“.insert(x’)
(7) M’ = transform(M) ; // 인과모형의 종속변수인 주변변수의 5점척도 또는 범주화
(8) Formula = (Θ“, M’ ~ Φ‘1 + Φ’2+ Φ‘3+, ... , + Φ’K);
(9) R = Regression(Formula); // Formula를 회귀분석 함수에 적용
(10) return R;
End

자 성별의 성별 및 나이 등이 될 수도 있다.

출력 데이터는 독립변수 Φ“과 종속변수 M사이의 인과관계를 나타내는 회귀방정식으로서 회귀식의 회귀계수 값들이 그 대상이 된다. 독립변수는 토픽들 중에서 선택되며 종속변수는 외적변수 M이거나 토픽목록 중에서 선택될 수 있다. Φ“은 입력 데이터 Φ’에서 추출된 토픽들과 5점 척도로 변환된 토픽비중 값들을 포함한다.

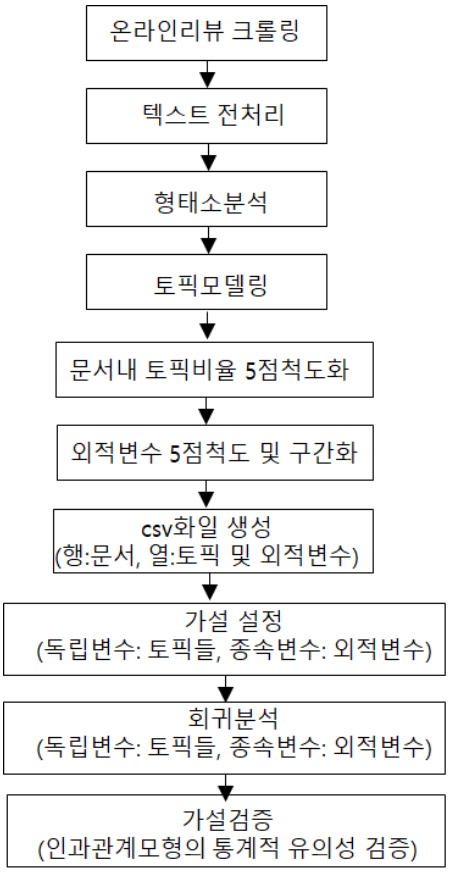
알고리즘의 처리내용 중에서 (1) 부분은 중간처리 값들을 저장할 용도의 변수들을 설정하고 초기화시키고 있다. (2)~(6) 부분은 문서 내 토픽비중 값들을 5점척도 값으로 표준화시키는

과정을 나타내고 있다. 이를 위해서 (3) 부분에서는 각 문서의 토픽비중 값 분포 범위를 5개 구간으로 나누고 있다. (4)~(5) 부분은 토픽비중 값을 5점척도 값으로 변환하여 x'변수에 추가하고 있다. (2)~(6) 부분의 반복처리 과정이 완료되면 문서집합 내 각 문서의 토픽비중 값들이 5점 척도 값으로 변환되어 Φ"에 들어간다. (7)부분에서는 주변변수 M을 5점 척도 기반으로 표준화시킨다. (8)부분에서는 Φ" 데이터를 대상으로 인과모형을 구성하고 있다. 독립변수는 Φ'1, Φ'2, Φ'3, ... , Φ'K 이고 종속변수는 M으로 설정하고 있다. (9)부분에서는 인과모형을 Φ"에 적합시켜 회귀계수들을 도출하고 객체변수 R에 보관하고 R을 리턴하게 된다.

Ⅳ. 사례분석

4.1 분석설계

앞에서 제안한 토픽모델링의 확장 알고리즘을 이용하여 실제 온라인리뷰를 분석한다. 온라인리뷰는 관광웹사이트인 트립어드바이저(www.tripadvisor.co.kr)의 관광상품 후기들을 사용한다. <그림 8>은 트립어드바이저의 온라인리뷰 분석과정을 나타내고 있다. 온라인리뷰는 크롤러(Crawler)에 의하여 트립어드바이저 웹사이트로부터 자동 수집되도록 한다. 파이썬 아나콘다 환경에서 기 개발된 웹크롤러를 활용하여 데이터를 수집한다. 수집된 온라인리뷰에서 불필요한 단어들을 제거하는 텍스트 전처리 과정을 수행한다. 텍스트 전처리가 적용된 온라인리뷰의 각 문장들은 토큰화(Tokenization) 기법이 적용되어 의미를 갖는 최소단어인 형태소들로 분할된다. KoNLPy 형태소분석기 등을 사용하여 형태소들을 생성한다. 형태소 단어들로 변환된 온라인리뷰들로부터 벡터화(Vectorization) 과정과 토픽모델링 과정을 거쳐 K개의 토픽들이 추출된다. 추출된 토픽들은 주요단어들의 해석을 바탕으로 토픽명칭이 부여된다. 문서별 토픽비율 데이터는 후처리 과정을 거쳐 통합가공데이터인 csv화일로 변환된다. 통합가공데이터의 토픽들을 독립변수로 하고, 외적변수를 종속변수로 하여 인과모형과 가설을 설정한다. 가설검증을 위하여 상관분석과 회귀분석을 실행한다.



<그림 8> 온라인리뷰 텍스트의 분석과정

구분	리뷰 수	작성기간	데이터구조
성산일출봉 리뷰	1000	2015년 ~ 2020년	번호:숫자형 내용:텍스트(문자형) 평점:숫자형 날짜: 날짜형

(a) 리뷰 구조

번호	내용	평점	날짜
0	입장료가격이 많이 오른듯해요 왕복 50분인데 5000원ㅜㅜ 무료관람가능한지역이	40	Aug-20
1	정상까지 올라가는건 조금 힘듭니다. 하지만 올라는 순간 그 힘듦이 순식간에 사라	50	Sep-19
2	코로나로 여름 여행을 어디로 갈까 하다가 제주도 다녀 왔습니다 가고 또 가도 볼	50	Jul-20
3	우뚝솟은 성산일출봉 근처에 올때부터 압도합니다. 아이가 힘들어해서 정상은 못	50	Jul-20
4	제주도하면 제일 먼저 떠오르는 곳은 한라산이지만 명소라고 하기에는 너무 특징	50	Jun-20
5	일출봉 정상까지 꼭 올라가 보세요. 어린아이들도 갈 정도로 그리 힘들지 않습니다	50	Apr-20
6	가족이 함께 오손도손 팔짱끼고 가며 이런저런 얘기도 나누고 앞으로의 계획도 약	40	Nov-19
7	제주도에 오면 그래도 한 번 정도 들러야 하는 곳 아닐까 합니다. 개인적으로는 우	40	Feb-20
8	예전에 여행이 활성화 되지 않았을 때나 성산 일출봉이 필수 코스였다. 그러나 한	30	Feb-20
9	일출을 보기 위해 새벽부터 서둘러서 해 뜨기도 전부터 산을 탔지요~~ 하지만 일	50	Jan-20
...	...	...	...

(b) 성산일출봉 온라인리뷰 예

<그림 9> 온라인리뷰 데이터

4.2 데이터수집

성산일출봉에 대한 온라인리뷰는 파이썬 크롤러(crawler)를 개발하여 수집하였다. 파이썬 크롤러는 파이썬 아나콘다 환경에서 셀레니움(Selenium)라이브러리를 이용하여 개발하였다. <그림 9>는 수집한 온라인리뷰의 크기 및 데이터구조와 실제 데이터를 나타내고 있다. 온라인리뷰는 2015년~2020년 사이에 등록한 글이다.

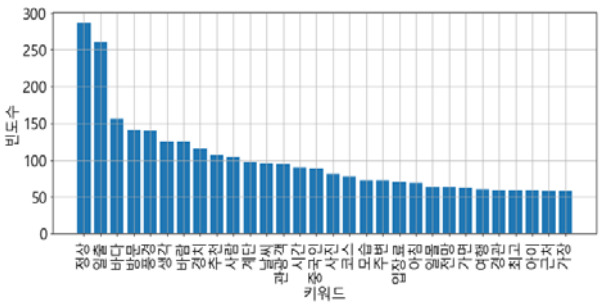
4.3 텍스트 전처리 및 기초분석

온라인리뷰의 내용 열(column)에는 성산일출봉에 대한 품평이 한글문장으로 표현되어 있다. 한글문장에는 한글단어들과 함께 은어나 특수문자 등이 포함하였다. 한글문장의 전처리과정을 통하여 특수문자, 은어 등의 불용어를 제거하였다.

국내에서 개발한 오픈소스 프로젝트인 KoNLPy 패키지(https://konlpy.org/ko/latest)를 사용하여 한글처리를 하였다. 추출된 형태소 단어들을 대상으로 빈도계산 및 토픽모델링을 적용하기 전에, <그림 10>에서 보여주는 바와 같이 의미적 가치가 높지 않다고 판단되는 단어들과 하나의 문자로 구성된 단어들을 제거하였다.

‘제주’, ‘제주도’, ‘성산’, ‘성산일출봉’, ‘정도’, ‘일출봉’, ‘정말’, ‘보기’, ‘한번’, ‘조금’, ‘보고’

<그림 10> 의미적 가치가 낮은 단어의 제거

빈발출현 단어목록(상위 30)	
정상, 일출, 바다, 방문, 풍경, 생각, 바람, 경치, 추천, 사람, 계단, 날씨, 관광객, 시간, 중국인, 사진, 코스, 모습, 주변, 입장료, 아침, 일몰, 전망, 가면, 여행, 경관, 최고, 아이, 근처, 가장	
	
막대그래프	워드클라우드

<그림 11> 성산일출봉 온라인리뷰의 빈발출현 단어

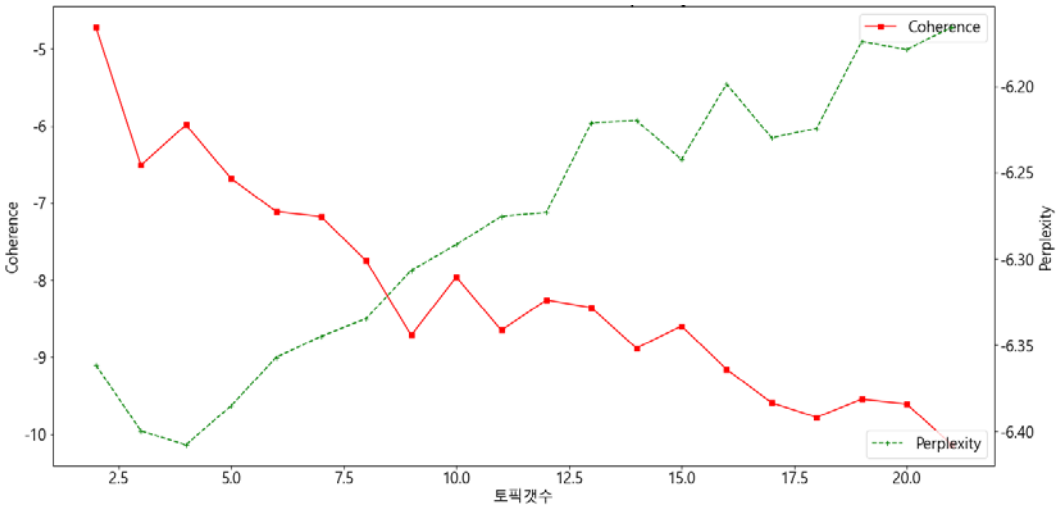
온라인리뷰 상에서 추출된 명사단어들은 단어-문서행렬(Term-Document Matrix)로 벡터화(Vectorization)한 후, 상위 30개의 빈발출현 단어들을 막대그래프 및 워드클라우드로 출력하였다. <그림 11>에서 볼 수 있는 바와 같이, '정상', '일출', '바다' 등의 단어들이 자주 출현하였다.

4.4 토픽모델링 및 인과분석

토픽모델링 결과의 평가척도로 혼란도(Perplexity)와 응집도(Coherence)가 있다. 혼란도는 모델링 결과가 문서를 얼마나 잘 반영하는지를 나타내는 척도이다. 혼란도가 작을수록 토픽모델링 결과가 실제 문서 내 토픽패턴을 잘 반영한 것으로 볼 수 있다. 응집도는 모델링 결과의 각 주제 내 단어들 사이에 의미론적으로 얼마나 유사한지를 나타내는 척도이다. 값이 높을 수록 주제 내 단어들 사이의 의미론적 일관성은 높다고 본다. 응집도가 높을수록 특정 토픽 내 단어들 사이의 의미적 일관성은 높다고 볼 수 있다.

토픽모델링을 하기 전에 적절한 토픽수를 설정하기 위하여 혼란도와 응집도 분석을 하였다. 토픽수를 2에서 20까지 변화를 주면서 각 경우의 혼란도와 응집도를 계산하였다.

<그림 12>는 혼란도와 응집도의 계산 결과를 꺽은 선으로 표현한 그래프다. 점선과 '+'은 혼란도를 나타내고 실선과 '■'은 응집도를 나타낸다. 전체적으로 혼란도는 토픽수가 증가할수록 증가하고 있으며, 응집도 값은 토픽수가 증가할수록 낮아지고 있다. 토픽수가 2~4일 때 혼란도는 낮고 응집도는 높게 되어 최적 토픽수를 2~5로 설정하였다. 최적의 토픽 수를 결정하기 위하여 혼란도와 응집도와 같은 정량적인 수단만을 사용하는 것은 한계가 있을 수 있다. 다양한 토픽 수에 따른 토픽들을 추출한 후, 분석자가 각 토픽의 의미를 해석하면서 적합한 토픽들을 추출하는 정성적 접근도 병행하는 것이 좋다. 그러나 여기에서의 토픽추출은 확장된 토픽모델링 기법을 검증하는 목적이 크기 때문에 정량적인 척도만으로 최적 토픽수를 결정하였다.



<그림 12> 성산일출봉 리뷰에 대한 혼란도(Perplexity)와 응집도(Coherence)

<표 3> 성산일출봉에 대한 주요 토픽(2개)

토픽명	주요단어 (Top 20)
관광활동 (토픽1)	0.035*"정상" + 0.021*"일출" + 0.018*"사람" + 0.017*"계단" + 0.016*"바람" + 0.014*"경치" + 0.013*"바다" + 0.012*"입장료" + 0.012*"시간" + 0.012*"아침" + 0.011*"관광객" + 0.011*"날씨" + 0.010*"생각" + 0.010*"사진" + 0.008*"가면" + 0.008*"풍경" + 0.008*"방문" + 0.007*"길이" + 0.007*"코스" + 0.007*"중국"
체험활동 (토픽2)	0.025*"일출" + 0.017*"방문" + 0.017*"풍경" + 0.016*"바다" + 0.015*"정상" + 0.013*"추천" + 0.013*"중국인" + 0.012*"생각" + 0.011*"경관" + 0.011*"가장" + 0.010*"가족" + 0.010*"근처" + 0.010*"아이" + 0.009*"자연" + 0.008*"모습" + 0.008*"일몰" + 0.008*"한라산" + 0.008*"등반" + 0.008*"명소" + 0.007*"여행"

<표3>은 성산일출봉 리뷰에 대하여 2개의 토픽을 추출한 경우이다. 각 토픽은 그 토픽을 대표하는 주요단어들을 포함한다. 각 토픽에 포함되는 주요단어들을 바탕으로 토픽명칭을 부여할 수 있다. 성산일출봉 리뷰에서는 전체적으로 관광여행 등의 내용일 것이므로 토픽별로 명확한 구분이 쉽지 않지만, 대체적으로 관광활동(Sightseeing)과 체험활동(Experienc)으로 구분하여 보았다. 토픽1에 포함된 단어들 중에서 '입장료', '관광객, '사진', '코스' 등의 단어들을 바탕으로, 토픽1은 '관광활동'으로 명칭을 부여하였다. 토픽2는 '가족', '아이', '등반' 등의 단어를 바탕으로 '체험활동'으로 명명하였다.

<표 4>는 성산일출봉 리뷰에 대하여 3개의 토픽을 추출한 경우이다. 토픽1과 토픽2에 포함된 단어들은 중복되는 경우가 많아, '관광활동A'와 '관광활동B'로 명명하였다. 토픽3은 '가족', '아이' 등의 단어를 바탕으로 '가족체험'으로 명명하였다.

<표 5>는 성산일출봉 리뷰에 대하여 4개의 토픽을 추출한 경우이다. 토픽1은 '개인체험A',

토픽2는 ‘관광활동A’, 토픽3은 ‘관광활동B’, 토픽4는 ‘개인체험B’로 명명하였다.

<표 6>은 성산일출봉에 대한 각 리뷰(문서)에서 관련 토픽의 가중치를 5점 척도로 변환한 결과를 나타내고 있다. 2개의 토픽을 추출하였을 경우, 첫 번째 리뷰에서는 ‘관광활동’과 ‘체험활동’의 가중치값이 높아서 모두 5점으로 변환되었다. 3개의 토픽을 추출하였을 경우, 첫 번째 리뷰에서는 ‘관광활동A’ 토픽의 비중은 5점, ‘관광활동B’ 토픽의 비중은 1점, ‘가족체험’은 5점으로 변환되었다.

<그림13>은 <표 6>의 변환 데이터를 기반으로 분석된 각 토픽들 사이의 상관분석 결과를 나타내고 있다. 추출 토픽수를 3으로 하였을 경우, ‘관광활동A’와 ‘관광활동B’ 사이의 상관도가 상대적으로 높게 나타나고 있다. 추출 토픽수를 4로 하였을 경우, ‘개인체험A’와 ‘개인체험B’사이의 상관도가 상대적으로 높게 나타나고 있다.

<표 4> 성산일출봉에 대한 주요 토픽(3개)

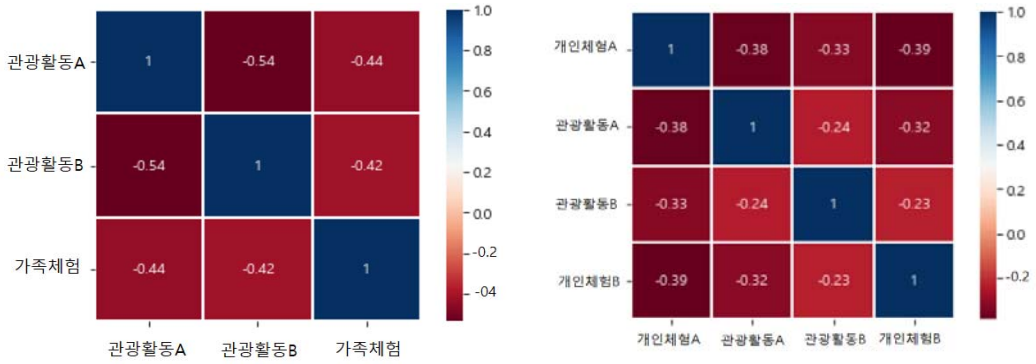
토픽명	주요단어 (Top 20)
관광 활동A (토픽1)	'0.040*"일출" + 0.028*"방문" + 0.024*"정상" + 0.018*"관광객" + 0.018*"바다" + 0.015*"풍경" + 0.015*"사람" + 0.014*"중국인" + 0.012*"전망" + 0.012*"시간" + 0.011*"계단" + 0.011*"바람" + 0.010*"추천" + 0.009*"중국" + 0.008*"느낌" + 0.008*"일몰" + 0.007*"꼭대기" + 0.007*"경치" + 0.007*"근처" + 0.007*"가장"
관광 활동B (토픽2)	'0.034*"정상" + 0.019*"코스" + 0.018*"생각" + 0.016*"바람" + 0.014*"사진" + 0.014*"주변" + 0.012*"일출" + 0.010*"바다" + 0.010*"경관" + 0.009*"입장료" + 0.009*"일몰" + 0.008*"주차장" + 0.008*"경치" + 0.008*"가장" + 0.007*"날씨" + 0.007*"최고" + 0.007*"풍경" + 0.006*"사람" + 0.006*"필수" + 0.006*"가면"
가족체험 (토픽3)	'0.022*"날씨" + 0.020*"정상" + 0.017*"경치" + 0.015*"풍경" + 0.014*"추천" + 0.013*"바다" + 0.012*"생각" + 0.011*"일출" + 0.011*"길이" + 0.010*"체력" + 0.010*"멀리" + 0.010*"계단" + 0.009*"시간" + 0.009*"입장료"+ 0.009*"절경" + 0.009*"등반" + 0.009*"중간" + 0.008*"주변" + 0.008*"아이" + 0.008*"가족"

<표 5> 성산일출봉에 대한 주요 토픽(4개)

주제명	주요단어 (Top 10)
개인 체험A (토픽1)	'0.023*"경치" + 0.020*"바다" + 0.019*"일출" + 0.017*"추천" + 0.015*"정상" + 0.012*"날씨" + 0.011*"경관" + 0.011*"풍경" + 0.010*"중국인" + 0.010*"멀리" + 0.010*"생각" + 0.010*"주변" + 0.009*"가면" + 0.009*"꼭대기" + 0.009*"자연" + 0.009*"느낌" + 0.009*"방문" + 0.008*"가장" + 0.008*"가족" + 0.008*"바람"
관광 활동A (토픽2)	0.037*"정상" + 0.019*"바람" + 0.019*"계단" + 0.018*"여행" + 0.017*"바다" + 0.016*"일출" + 0.015*"전망" + 0.015*"풍경" + 0.012*"사람" + 0.012*"추천" + 0.011*"경치" + 0.011*"방문" + 0.010*"입장료" + 0.010*"중국인" + 0.009*"관광객" + 0.009*"주차장" + 0.009*"중간" + 0.009*"세계" + 0.009*"여기" + 0.009*"분화구"
관광 활동B (토픽3)	0.023*"관광객" + 0.021*"풍경" + 0.018*"일출" + 0.017*"정상" + 0.016*"중국" + 0.015*"방문" + 0.015*"코스" + 0.015*"중국인" + 0.013*"바다" + 0.011*"바람" + 0.011*"전망" + 0.011*"예전" + 0.010*"아이" + 0.010*"왕복" + 0.009*"사진" + 0.008*"요즘" + 0.008*"경관" + 0.007*"날씨" + 0.007*"위로" + 0.007*"산책"
개인 체험B (토픽4)	0.035*"일출" + 0.032*"정상" + 0.022*"생각" + 0.017*"시간" + 0.017*"방문" + 0.016*"사람" + 0.013*"사진" + 0.012*"계단" + 0.012*"일몰" + 0.010*"코스" + 0.010*"모습" + 0.009*"아침" + 0.009*"바람" + 0.009*"입장" + 0.009*"도착" + 0.008*"금방" + 0.008*"근처" + 0.008*"풍경" + 0.007*"날씨" + 0.007*"입장료"

<표 6> 문서 내 토픽비율 및 외적변수 5점척도화

리뷰 번호	토픽수=2		토픽수=3			토픽수=4				평점
	관광 활동	체험 활동	관광 활동A	관광 활동B	가족 체험	개인체 험A	관광 활동A	관광 활동B	개인 체험B	
1	5	5	5	1	1	1	1	5	1	4
2	1	5	1	1	5	1	1	5	1	5
3	1	5	5	2	1	5	1	1	1	5
4	1	5	1	5	1	1	5	1	1	5
5	5	1	1	5	1	5	1	1	1	5
...	...	...	...	...	...	...	...	...	...	
1000	1	5	1	5	1	5	2	1	1	4



(a) 토픽수 3 (b) 토픽수 4

<그림 13> 상관분석 결과

<표 7>은 성산일출봉 리뷰로부터 추출한 토픽들을 독립변수로, 평점을 종속변수로 설정하여 회귀분석을 실시한 결과를 나타내고 있다. 토픽 수를 2로 설정하여 추출한 토픽들인 '관광활동'과 '체험활동'은 '평점'에 유의한 영향을 미치지 않았다. 토픽 수를 3으로 설정하여 추출한 토픽들인 '관광활동A', '관광활동B', '가족체험'은 '평점'에 $p<0.01$ 수준에서 유의한 영향을 미쳤다. 회귀계수 값이 음수인 것으로 보아, 각 토픽들의 비중(가중치)는 평점에 부정적으로 영향을 미치고 있다. 즉, 성산일출봉 리뷰에서 3개의 토픽을 추출하였을 경우, 온라인리뷰 중에서 '관광활동'이나 '가족체험'과 관련한 토픽 비중이 높은 리뷰는 불평, 불만 등 부정적인 내용을 포함한 것임을 알 수 있다. 토픽 수를 4로 설정하여 추출된 토픽들인 '체험활동A', '관광활동A', '관광활동B', '개인체험B' 는 '평점'에 $p<0.05$ 수준에서 유의한 영향을 미쳤다. 회귀계수 값이 양수인 것으로 보아, 리뷰 내 각 토픽들의 비중은 평점에 긍정적으로 영향을 미치고 있음을 알 수 있다.

<표 7> 종속변수를 평점으로 한 회귀분석 결과

토픽수	독립변수 (토픽명)	회귀계수	표준오차 (SE)	t	유의확률	Adj R^2	F
2	(상수)	4.4912	0.114	39.356	0.000	0.003	2.445*
	관광활동	0.0179	0.018	0.994	0.321		
	체험활동	-0.0070	0.018	-0.387	0.699		
3	(상수)	5.1989	0.185	28.092	0.000	0.017	6.619***
	관광활동A	-0.1060	0.026	-4.065	0.000***		
	관광활동B	-0.0775	0.027	-2.849	0.004***		
	가족체험	-0.0925	0.026	-3.512	0.000***		
4	상수	3.9926	0.229	17.426	0.000	0.005	2.251*
	개인체험A	0.0639	0.029	2.208	0.027**		
	관광활동A	0.0738	0.030	2.500	0.013**		
	관광활동B	0.0473	0.031	1.547	0.122		
	개인체험B	0.0710	0.028	2.546	0.011**		

*:p<0.1, **: p<0.05, ***:p<0.01

추출 토픽 수에 따라, 관광활동 등의 토픽이 평점에 영향을 미치는 방향성이 양 또는 음으로 달라지는 현상은 토픽들 사이의 차별성이 낮아서, 결과적으로 부여된 토픽명칭의 정확성이 떨어지기 때문인 것으로 해석할 수 있다. 토픽별 차별성이 높지 않은 점과 함께, 전체적으로 회귀모형의 R2값과 F값이 낮은 점도 감안하여 해석할 필요가 있다. 결국, 토픽모델링으로부터 도출된 토픽들의 의미적 명확성이 높지 않으면, 회귀분석 결과의 정확성도 높지 않을 수 있다. 그럼에도 불구하고, 기술 발전에 의한 토픽모델링의 정확성이 높아진다는 전제하에, 토픽모델링의 확장분석을 통한 인과모형 도출과 귀납적 가설 설정의 가능성을 확인하였다는 점에서 그 의의가 있다.

Ⅴ. 결 론

기존의 사회과학연구는 연역적 방법으로서, 문헌고찰 등을 통하여 연역적으로 연구가설을 도출하고 설문조사 등의 방식으로 가설검증을 수행하는 방법이 일반적이었다. 연역적 방법에 의한 사회과학연구는 수립된 연구가설의 이론적 엄격성과 정합성의 관점에서는 강건할 수 있으나, 혁신적이고 다양한 가설을 도출하는 데는 한계가 있을 수 있다. 본 논문에서는 사회과학을 위한 연구가설을 수립할 때, 데이터 기반의 연구가설을 도출할 수 있도록 기존의 토픽모델링 기법을 확장하여 활용하는 방법을 제안하였다. 기존의 연역적 가설수립과 대비되는 개념으로, 일명 귀납적 가설수립이 가능할 수 있도록 토픽모델링 기법을 확장하였으며, 제안된 기법을 기반으로 한 성산일출봉 온라인리뷰의 분석 사례를 제시하였다.

확장된 토픽모델링 기법의 개념은 기존 토픽모델링에 의하여 도출된 토픽들과 외적변수들을 대상으로 후속적인 인과관계 등을 분석하는 것이었다. 도출된 토픽들은 연역적 연구방법에서의 연구변수 즉, 설문조사 항목들과 대응시킬

수 있었다. 각 온라인리뷰 내에서 특정 토픽의 가중치는 그 토픽에 대한 사람들의 관심도가 될 수 있다. 특정 토픽을 연구변수로 설정하였을 경우, 해당 연구변수에 대한 사람들의 관심도 또는 생각을 5점 척도로 변환시킬 수 있었다. 확장된 토픽모델링 기법을 통하여 각 토픽들을 연구변수로 설정할 수 있도록 데이터를 가공하였다.

본 논문에서는 토픽모델링의 확장 개념을 확장 아키텍쳐와 확장 알고리즘 형태로 형식화하였다. 확장 아키텍쳐 및 알고리즘은 기존 토픽모델링에 의하여 도출된 토픽들을 대상으로 문서 내 토픽 가중치에 근거한 데이터 가공방법과 가공 데이터에 대한 상관분석 및 회귀분석 등을 포함하도록 하였다. 또한, 본 논문에서는 토픽모델링 확장 알고리즘을 적용하여 성산일출봉에 대한 온라인리뷰의 분석 사례를 제시하였다. 텍스트 기초분석을 통하여 성산일출봉 온라인리뷰들에 대한 탐색적 분석을 하였다. 각 온라인리뷰 내 토픽 가중치 등을 바탕으로 상관분석과 회귀분석이 가능할 수 있도록 5점 척도 기반으로 데이터를 변환하였다. 도출된 토픽들을 독립변수로, 리뷰의 평점을 종속변수로 하여 회귀분석을 실시하였으며 이를 바탕으로 가설을 도출할 수 있도록 하였다.

본 논문의 사례분석에서는 도출된 토픽들의 차별성이 높지 않아서, 도출된 회귀모형의 모형적합도 또한 높지 않았다. 그럼에도 불구하고, 지속적 기술 발전으로 토픽모델링의 정확성이 높아진다는 전제하에, 토픽모델링의 확장분석을 통한 인과모형 도출 및 귀납적 가설 설정의 가능성을 확인하였다는 점에서 그 의의가 있다.

본 논문에서 제안한 토픽모델링의 확장기법은 다양한 방식으로 활용될 수 있다. 온라인리뷰 뿐만 아니라 인터넷검색 결과와 뉴스기사 등의 분석에도 활용될 수 있다. 분석대상 텍스트에 따라 텍스트 이외의 외적변수는 다양하게 존재할 수 있으며 토픽목록과 주변변수 사이의 다양한 인과모형을 구성할 수 있다. 인과모형을 바탕으로 새로운 사회현상을 설명하기 위한 가설을 설정하여 검증할 수 있다. 채택된 가설은 새로운 사회과학 지식으로 창출될 수 있다. 본 연구에서 개발한 새로운 텍스트분석방법으로 인하여 다양한 분야에서 생성되는 온라인리뷰 및 텍스트 데이터의 가치가 높아지는 계기가 될 것이다.

참고문헌

김근형, “개체연관망 모델에 의한 오피니언마이닝의 확장,” 한국정보처리학회논문지, 18-D권, 4호, 2011, pp.237-244.

김근형, 오성열, “온라인 고객리뷰 분석을 통한 시장세분화에 텍스트마이닝 기술을 적용하기 위한 방법론,” 한국콘텐츠학회논문지, 9권, 8호, 2009, pp.272-284.

박진희, 전미선, 배선형, 김희준, “암생존자 삶의 질 영향요인에 대한 연구동향: 텍스트 네트워크 분석과 토픽모델링,” 중앙간호연구, 제21권 제4호, 2021, pp.231-240.

신서영, 이범준, “코로나19 확산에 따른 외식에 대한 소비자 인식 분석: 토픽모델링 및 네트워크 분석의 활용,” 호텔경영학연구, 제30권, 제8호, 2021, pp.71-90.

심영석, 김홍봄, “온라인 리뷰 빅데이터 기반의 Word2Vec 기법을 활용한 관광지 개성과 여행객 평점 간 구조적 관계 분석,” 관광학연구, 제8권, 제164호, 2018, pp.165-189.

신주하, 임희진, 이병주, “국내 복합리조트 서비스품질에 대한 토픽모델링 및 감성분석,” 관광연구저널, 제35권, 제11호, 2021, pp.191-206.

유재호, 조연희, 전의찬, “토픽모델링 분석에 의한 글벌 그린뉴딜 연구동향 분석,” 한국기후변화학회지, 제12권, 4호, 2021, pp.289-298.

윤상훈, 김근형, “Word2Vec를 이용한 토픽모델링이 확장 및 분석사례,” 정보시스템연구, 제30권, 제1호, 2021, pp.45-63.

이민철, 김혜진, “텍스트 마이닝 기법을 적용한 뉴스 데이터에서의 사건 네트워크 구축,” 지능정보연구, 제24권, 제1호, 2018, pp.183-203.

이병철, 김두산, “OTA 리뷰를 사용하여 호텔 서비스 품질 결정 요인 도출-LDA 토픽 모델링,” 호텔리조트연구, 제19권, 제4호, 2020, pp.41-58.

이창용, 김근형, “온라인리뷰의 랭킹모델링을 위한 양과 질의 인과모형 분석,” 정보시스템연구, 제28권, 제1호, 2019, pp.1-16.

채호근, 이기현, 이주연, “토픽모델링 분석 기법을 활용한 국내외 금융보안 분야 연구 동향 분석,” 한국산업정보학회논문지, 제26권, 1호, 2021, pp.83-95.

민진영, 이애리, “‘좋아요’와 ‘싫어요’같은 간접적 사회적 정보의 방향과 강도는 온라인 뉴스 콘텐츠 댓글의 숙의의 질과 어떤 관련이 있는가? 토픽 모델링을 이용한 토픽 다양성 분석,” 정보시스템연구, 제30권, 4호, 2021, pp.303-327.

황원준, “인공지능과 인간안보 : 토픽모델링 기법을 활용한 유럽의 인공지능 안보화 분석,” 21세기정치학회보, 제31권, 2호, 2021, pp.55-82.

Blei, D., A. Ng.,and M. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, Vol.3, 2003, pp.993-1022.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M., “ Reading Tea Leaves: How Humans Interpret Topic Models,” In Advances in Neural Information Processing Systems, 2009, pp.288-296.

Hofmann T., “Probabilistic Latent Semantic Analysis”, Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 1999, pp.289-296.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. ,“Efficient Estimation of Word Representations in Vector Space,” 2013, arXiv preprint arXiv:1301.3781.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T.,“Automatic Evaluation of Topic Coherence”, In Human Language Technologies : The 2010 Annual Conference of the North American

Chapter of the Association for Computational Linguistics, 2010, pp.100-108.

Philipp Drieger, “Semantic Network Analysis as a Method for Visual Text Analytics,” Procedia - Social and Behavioral Sciences, 79, 2013, pp.4-17.

김 근 형 (Kim, Keun Hyung)

서강대학교 전자계산학사와 동 대학원 석사 및 박사학위를 취득하였다. 현재 제주대학교 경영정보학과 교수로 재직하고 있으며, 주요 관심분야는 빅데이터분석, 데이터마이닝, 정보시스템 등이다.

<Abstract>

Extension and Case Analysis of Topic Modeling for Inductive Social Science Research Methodology

Kim, Keun Hyung

Purpose

In this paper, we propose the method to extend topic modeling techniques in order to derive data-based research hypotheses when establishing research hypotheses for social sciences, As a concept in contrast to the existing deductive hypothesis establishment methodology for the social science research, the topic modeling technique was expanded to enable the so-called inductive hypothesis establishment methodology, and an analysis case of the Seongsan Ilchulbong online review based on the proposed methodology was presented.

Design/methodology/approach

In this paper, an extension architecture and extension algorithm in the form of extending the existing topic modeling were proposed. The extended architecture and algorithm include data processing method based on topic ratio in document, correlation analysis and regression analysis of processed data for topics derived by existing topic modeling. In addition, in this paper, an analysis case of the online review of Seongsan Ilchulbong Peak was presented by applying the extended topic modeling algorithm. An exploratory analysis was performed on the Seongsan Ilchulbong online reviews through the basic text analysis. The data was transformed into 5-point scale to enable correlation and regression analysis based on the topic ratio in each online review. A regression analysis was performed using the derived topics as the independent variable and the review rating as the dependent variable, and hypotheses could be derived based on this, which enable the so-called inductive hypothesis establishment.

Findings

This paper is meaningful in that it confirmed the possibility of deriving a causal model and setting an inductive hypothesis through an extended analysis of topic modeling.

Keywords: topic modeling, inductive research method, hypothesis establishment, expansion algorithm, case analysis, Seongsan Ilchulbong

* 이 논문은 2022년 8월 9일 접수, 2022년 9월 16일 1차 심사, 2022년 11월 11일 게재 확정되었습니다.