「정보시스템연구」 제30권 제1호
한국정보시스템학회
2021년 3월, pp. 45~64
http://dx.doi.org/10.5859/KAIS.2021.30.1.45

Word2Vec를 이용한 토픽모델링의 확장 및 분석사례*

윤상훈** · 김근형***

〈목 차〉

Ⅰ. 서 론

스마트기기 및 SNS의 대중화로 온라인리뷰(Online reviews)가 넘쳐나고 있다. 온라인리뷰는 상품 및 서비스에 대한 고객의 품평후기를 기록한 디지털 텍스트 데이터이다. 설문조사 방식으로 수집한 고객 데이터는 비자발적으로 생성되는 반면, 온라인리뷰는 고객이 자발적으로 만들어내는 것이기 때문에 고객의 진성성이 담겨있는 데이터라고 할 수 있다(이창용, 김근형, 2019). 고객의 생각을 담고 있는 온라인리뷰에 대한 분석정보는 신제품기획이나 마케팅전략 수립에서 중요한 정보를 제공할 수 있다.

토픽모델링(Topic modeling)은 텍스트데이터를 분석하는 기법으로서, 텍스트 형태의 온라인리뷰를 분석할 때도 효과적으로 활용할 수 있다. 토픽모델링 알고리즘은 이미 1999년에 호프만에 의하여 제안되었지만, 최근에 더욱 주목을 받는 이유는 컴퓨팅환경의 발전 덕분이다. 토픽모델링 기법이 처음으로 발표될 당시에는 컴퓨팅파워가 좋지 않아 대량의 텍스트 데이터를 제대로 처리하기가 어려웠으며 효과적인 결과를 만들어내기에는 역부족이었다. 반면, 오늘날에는 개인용 컴퓨터를 활용해서도 대용량 텍스트 데이터를 무난하게 처리할 수 있게 되었다.

최근 들어 토픽모델링 기법을 활용하여 온라인리뷰를 분석하는 연구들이 많아졌다(정영진 등, 2017; 김정규, 정철, 2019; 남승주 등, 2020; 이병철, 김두산, 2020; 이혜진, 강영옥, 2020).

* 이 논문은 2020학년도 제주대학교 교원성과지원사업에 의하여 연구되었음.
** 제주대학교 대학원 경영정보학과, banktown1700@gmail.com(주저자)
*** 제주대학교 경영정보학과, khkim@jejunu.ac.kr(교신저자)

이러한 연구들은 일반적으로 토픽모델링의 LDA 알고리즘(Blei, D. et al, 2003)을 이용하고 있다. LDA 알고리즘에서는 하나의 텍스트 문서가 여러 토픽들로 구성된다고 가정하여 각 토픽에 관련 단어를 할당하는 방식으로 토픽추출을 한다. 최적의 단어-토픽 할당을 통해 각 토픽과 높은 관련이 있는 주요단어들을 찾아내어 토픽별 주요단어 목록을 생성한다. 추출된 각 토픽의 의미파악과 명칭부여는 토픽에 할당된 주요단어들을 참고하여 분석자가 수동적으로 진행한다. 토픽에 할당된 주요단어들은 다른 토픽에도 중복적으로 배정될 수 있다. 이것은 추출된 토픽들을 의미적으로 구분하는데 장애요인이 될 수 있다. 특히, LDA의 경우 온라인리뷰 개수가 1000건 미만일 경우에는 잘 작동하지 않은 것으로 알려져 있다(Qian et al., 2017). 하지만 현실에서 분석 필요성이 있는 온라인리뷰는 1000건 미만으로 이루어진 경우도 다수 존재하기 때문에 소량의 온라인 리뷰의 분석에 적용할 수 있는 토픽모델링 방법이 필요하다.

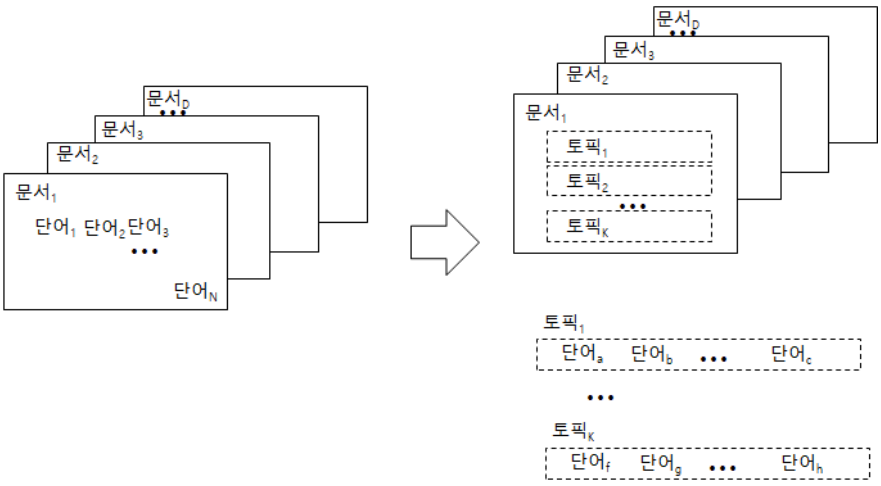
본 논문에서는 LDA 알고리즘을 개선한 확장 알고리즘을 제안한다. Word2Vec(word embedding to vector) 기법((Mikolov et al., 2013)을 LDA 알고리즘과 접목하여 토픽의 의미적 구분을 보다 명확히 할 수 있는 방법을 제안하고자 한다. Word2Vec은 인공신경망(Artificial Neural Network)기법을 이용하여 의미적으로 유사한 단어들을 추출해주는 학습모델로서, LDA에 Word2Vec이 적용되면 토픽들의 의미적 구분을 강화시킬 수 있다. 또한, 우도관광지의 온라인리뷰를 분석하는 사례를 제시하면서 확장알고리즘의 효과성을 확인하고자 한다. 트립어드바이저 웹사이트(www.tripadvisor.com) 에서 우도관광지에 대한 온라인리뷰는 1000건 미만이며, 온라인리뷰 내에 포함된 토픽들도 관광도메인으로의 수렴 가능성 때문에 토픽 구분이 명확치 않을 수 있어서 확장알고리즘의 성능을 시험하는데 적합할 수 있다. 본 논문에서 제안하는 확장알고리즘은 적은 분량의 온라인리뷰에 대하여 기존의 토픽모델링 기법보다 효과적인 결과를 제공할 수 있을 것이라는 측면에서 온라인리뷰 분석의 또 다른 방법이 될 수 있다.

Ⅱ. 이론적 배경

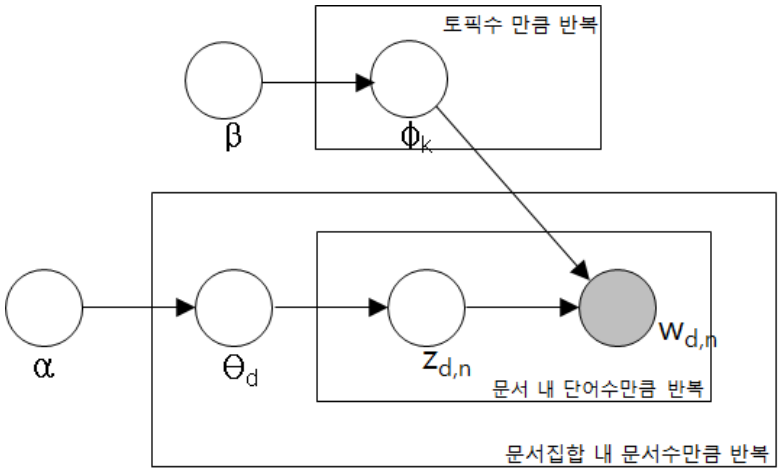
2.1 토픽모델링과 LDA

토픽모델링은 텍스트문서 집합으로부터 주로 언급되는 핵심주제 즉, 토픽(Topic)들을 추출할 수 있도록 한다(Hofmann T., 1999). 토픽모델링 기법은 <그림 1>에서 보여주는 것처럼, 단위문서(Document)들로 구성된 대량의 문서집합(Corpus)에서 단어(word)들의 분포 등을 확률적으로 계산하면서 여러 개의 토픽들을 추출하고 각 문서 내에 포함된 토픽비율 등을 계산한다. 각 토픽에는 주요단어들이 할당되어 토픽들을 의미적으로 구분할 수 있고 명칭을 부여할 수 있다.

LDA(Latent Dirichlet Allocation)는 토픽모델링을 위한 대표적인 알고리즘이다(Blei, D. et al., 2003). LDA에서는 토픽의 확률 분포와 단어 확률 분포를 추정하기 위한 사전 분포로 디리클레 분포를 사용한다. 토픽 모델링에서 디리클레 분포가 사용되는 이유는 디리클레 분포가 다른 다항 분포함수와 곱하면 다시 디리클레



<그림 1> 토픽모델링의 개념



<그림 2> LDA 아키텍처

분포 형태가 되어 관찰된 단어를 활용한 사후 분포를 만드는데 계산의 편의성이 있기 때문이다. <그림 2>는 LDA의 추론 과정에 대한 아키텍처를 나타내고 있다. LDA는 처음에 문서의 각 단어에 임의적으로 토픽을 할당한 후, 이를 실제 데이터와 반복적으로 대조하면서 사후 분포의 확률을 구하는 방식으로 작동된다. α와 β는 분석자가 수동적으로 설정하는 하이퍼 파라미터로 디클레 분포의 특성을 조정하는 역할을 한다. ϕk 는 k번째 토픽에 해당하는 벡터로 β의 영향을 받는다. ϕk 는 사전(priori)에 임의적으로 설정되지만 사후적으로 토픽수 만큼 반복되

면서 실제 문서를 반영하는 값으로 추론되어 변환된다. θ_d 는 d번째 문서가 가진 토픽의 비중을 나타내는 벡터로서 α의 영향을 받는다. θ_d 도 처음에는 임의적으로 설정되지만 사후적으로 문서집합 내 문서 수만큼 반복되면서 실제 문서를 반영하는 값으로 추론되는 과정이 반복된다. $z_{d,n}$ 은 d번째 문서의 n번째 단어가 어떤 토픽에 해당하는지 할당해주는 역할을 하는 파라미터이다. $z_{d,n}$ 도 처음에 임의적으로 설정되지만 사후적으로 문서 내 단어 수만큼 반복되면서 실제 문서를 반영하는 값으로 추론되어 변환된다. $w_{d,n}$ 은 d번째 문서의 n번째 단어를 의미한다.

하이퍼 파라미터(Hyper parameters)인 α와 β, 문서집합으로부터 관찰 가능한 $w_{d,n}$ 을 제외한 모든 변수는 미지수이다. 결국 아래 수식의 p(zd,i=j|z-i,w)를 최대로 만드는 z, ϕ, θ 를 구하는 것이 LDA 알고리즘의 목표가 된다.

$$p(z_{d,i}=j|z_{-i},w)=\frac{n_{d,j}+\alpha_j}{\sum_{i=1}^{K}(n_{d,i}+\alpha_i)}\times\frac{v_{k,w_{d,n}}+\beta_{w_{d,n}}}{\sum_{j=1}^{V}(v_{k,j}+\beta_j)}=A\times B$$

즉, 위 수식은 단어 w가 주어지고 w의 소속 토픽인 i를 일시적으로 지웠을 때(z-i의 의미) A×B를 최대로 하는 새로운 소속토픽 j를 계산한다. 위 수식의 각 기호의 의미는 <표 1>과 같다.

LDA 알고리즘에서 토픽 수 K는 수동적으로 설정된다. 토픽수를 너무 높게 설정하면 의미없는 주제들이 도출될 수 있으며, 너무 적게 설정하면 토픽에 너무 많은 단어가 포함되어 토픽 구분이 어려워질 수 있다(이병철, 김두산, 2020). 최적의 토픽 수는 Perplexity값과 Coherence 값을 사용하기도 한다. Perplexity 값은 특정 확률 모델이 실제로 관측되는 값을 얼마나 잘 예측하는지 평가할 때 사용된다. Perplexity가 작을수록 토픽모델링 결과가 실제 문서집합을 잘 반영한 것으로 본다(Chang et al., 2009). Perplexity가 낮다고 해서 항상 용이한 것은 아니다. Coherence는 한 토픽 안에 의미론적으로 유사한 단어가 얼마나 많이 모여 있는지를 나타내는 척도이다(Newman et al., 2010). Coherence값이 높을수록 주제 내 단어들 사이의 의미론적 일관성은 높다고 본다.

<표 1> 수식의 기호 의미

기호	의미
$n_{d,k}$	k 번째 토픽에 할당된 d번째 문서의 단어 빈도
$v_{k,w_{d,n}}$	문서집합에서 k번째 토픽에 할당된 단어 $w_{d,n}$의 빈도
wd,n	d 번째 문서에 n번째로 등장한 단어
α	문서의 토픽 분포 생성을 위한 디리클레 분포 파라메터
β	토픽의 단어 분포 생성을 위한 디리클레 분포 파라메터
K	토픽의 단어 분포 생성을 위한 디리클레 분포 파라메터
V	사용자가 지정하는 토픽 수
A	문서집합에 등장하는 전체 단어 수
B	d 번째 문서의 n번째 단어$w_{d,n}$이 k번째 토픽과 맺고 있는 연관성 정도

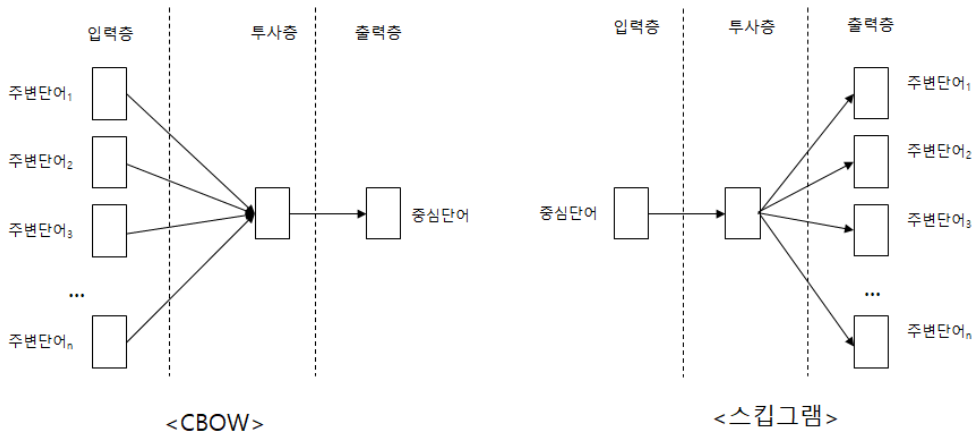
2.2 Word2Vec

단어임베딩(Word Embedding)은 문서집합에서 유사단어들을 찾아주는 기술이다. 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. Word2Vec의 학습방법은 크게 CBOW(Continuous Bag of Words)와 스킵그램(Skipgram)으로 나눌 수 있다. CBOW는 주변단어들을 통해 중심단어를 맞추는 방식이다. 예를 들어, “철수는 __에 간다” 라는 문장에서 “철수는”, “에”, “간다”를 활용하여 중심에 있는 단어를 맞추도록 학습시키는 방식이다. 스킵그램은 “___ 학교에 __”라는 문장에서 중심단어인 “학교에”를 가운데 놓고 주변단어를 학습시켜 유사단어를 예측하도록 하는 방식이다. <그림 3>은 Wrod2Vec의 CBOW와 스킵그램의 신경망 모형을 나타내고 있다.

2.3 관련연구 고찰

토픽모델링을 활용한 연구들은 최근 들어 다양한 분야에서 활발히 이루어지고 있다. 토픽모델링 분야의 연구주제를 크게 구분해보면, 기법을 활용하는 유형과 기법을 개선하는 유형으로 나눌 수 있다. 기법을 활용하는 유형은 토픽모델링 방법을 이용하여 특정 도메인의 텍스트데이터를 분석하는 연구들이다. 기법을 개선하는 연구는 토픽모델링 알고리즘의 성능을 개선하기 위하여 기술적으로 접근하는 연구들이다. 최근 들어서는 다양한 도메인의 관련 텍스트 데이터를 토픽모델링 기법으로 분석하는 연구가 주를 이루고 있었다.

차윤정 등(2015)의 연구에서는 삼성과 애플 스마트폰 신제품에 대한 트위터 데이터를 토픽모델링 기반으로 분석한 후 서로 비교함으로써 마케팅전략을 수립할 수 있는 정보를 제공하고 있다. 정영진과 조윤호(2017)의 연구에서는 토



<그림 3> Word2Vec의 신경망 모형

픽 모델링을 기반으로 도서 고객에게 적절한 추천을 제공할 수 있는 방법론을 제안하고 있다. 도서추천프로세스와 사용자 프로화일을 제안하고 있으며 국내 대형 온라인서점의 고객 트랜잭션 데이터를 대상으로 분석 응용하고 있다. 이진욱 등(2017)의 연구에서는 비정형 데이터의 대체 분석 방법을 선택하기 위한 목적으로 자동차 품질 검토 데이터를 분석하고 있다. 비정형 데이터를 분석하는 방법은 주로 비정형 데이터의 빈도를 기반으로 데이터 간 상관관계 정보를 활용하고 있다. 랜덤포레스트 방법이 가장 효과적인 것으로 나타났고 Word2Vec 방법은 자동차 부품들과 가장 관련성이 높은 데이터를 발견하는데 효과가 있는 것으로 나타났다. 심영석 등(2018)의 연구에서는 온라인 리뷰의 질적·양적 정보를 활용하여 의미론적 차원에서 관광지 개성을 확장하기 위해, 신경망 언어 모델인 Word2Vec를 활용하여 여행객 평점과의 영향관계를 추정하고 있다. 분석 결과, 관광지 개성은 기존 브랜드 개성의 구성개념과 달리 관광지 경험 후의 정서적인 감정을 표현하는 유사단어들이 도출되었다. 이민철과 김혜진(2018)의 연구에서는 Word2Vec와 LDA 토픽 모델링을 사용하여 방대한 양의 뉴스기사로부터 데이터를 추출하여 주요 사건을 감지하고, 사건들 간의 관련성을 판단하여 사건 네트워크를 구축함으로써 독자들에게 현시적이고 요약적인 사건정보를 제공하는 기법을 제안하고 있다. 김정규와 정철(2019)의 연구에서는 한국과 미국의 관광 관련 특허를 바탕으로 관광 분야의 기술 동향을 정량적으로 평가하고 있다. 이를 위해 LDA 토픽 모델링을 이용하여 관광 분야에서 활용되고 있는 세부 기술들을 도출하고 이를 바탕으로 현재 성장하는 기술과 쇠퇴하는 기술들을 파악하고 있다. 이병철과 김두산(2020)의 연구에서는 고객만족, 재방문 등 고객 행동에 영향을 미치는 결정요인으로 간주되는 호텔 서비스 품질의 구성 요소를 개발하는 목적을 갖고 온라인 호텔 리뷰를 토픽모델링 기법으로 분석하고 있다. 분석 결과 '가족 친화', '휴먼 서비스', '자원 접근성', '룸뷰', '편의 시설', '체크인 및 예약 편의성', '청결성' 등 7 가지 서비스 품질 주제가 추출되었다. 이혜진과 강영옥(2020)의 연구에서는 SNS 데이터 중 부산과 관련한 텍스트 데이터에 대하여 토픽모델링 기법으로 주요 관광 카테고리를 도출하고 있다. 분석결과는 시장/음식거리, 문화유산/역사명소, 전망/조망, 문화/축체, 공원/자연경관 등 9개의 카테고리를 생성하였다. 남승주 등(2020)의 연구에서는 쇼핑관광에 대한 선호 및 만족도 등에 대한 온라인리뷰에 대한 텍스트 분석을 통하여 인바운드 관광객들의 인식을 파악하고 있다. 야오즈옌 등(2020)은 온라인리뷰의 텍스트마이닝에 기반한 한국방문 외국인 관광객의 문화적 특성 등을 연구하였다.

토픽모델링 알고리즘을 개선하는 기술적 접근의 연구들도 있었다. Kyung Im Kim et al.(2009)의 연구는 LDA알고리즘 자체의 개선은 아니지만, 입력 텍스트를 전처리하는 과정에서 SVD(Singular Value Decomposition)방법을 도입함으로써 모호하고 중복적인 단어들을 처리할 수 있도록 하여 분석결과를 보다 의미있게 만들고 있다. Moody(2016)는 LDA알고리즘과 Word2Vec을 혼합하여 LDA2Word 알고리즘을 제안하고 있다. LDA2Word 알고리즘은 서로 연관된 단어들이 토픽 내 포함된 주요단

어들로 나타나도록 하여 토픽들 사이의 구분이 보다 명확할 수 있도록 한다. 그러나 이러한 기법도 토픽 내 주요단어들이 다른 토픽에도 할당되는 중복성은 여전히 존재하기 때문에 토픽수가 작은 코퍼스에 대해서는 여전히 성능이 좋지 않다. Won-joon Choi and Euhee Kim (2019)는 Word2vec기법과 클러스터링 기법, 토픽모델링을 혼합한 형태의 텍스트 분석방법을 제안하고 있다. Word2vec의 결과들을 K-평균군집화 기법으로 클러스터링하여 군집들을 도출한 후, 각 군집에 토픽모델링기법을 적용하여 2차적인 분석을 진행하는 방법을 제안하고 있다. 기존의 군집화기법에 토픽모델링을 적용하여 보다 세밀한 군집화를 도출하는 연구라고 할 수 있다.

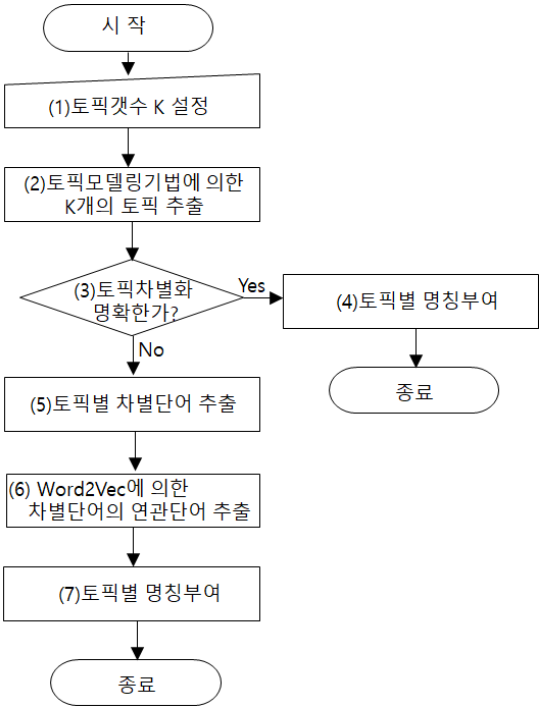
토픽모델링에 대하여 기술적으로 접근한 이러한 연구들은 토픽모델링의 성능을 일정부분 개선하는 효과가 있으나, 소량의 텍스트문서에 적용할 때는 그 효과가 크지 않았다. 본 논문에서는 소량의 텍스트문서를 가정하였을 때의 효과적인 텍스트분석 방법을 제안한다.

Ⅲ. 연구설계

3.1 토픽모델링의 확장모형

토픽모델링에 의하여 추출된 각 토픽들에는 일정 개수의 주요 구성단어들이 할당된다. 분석자는 각 토픽에 포함된 구성단어들을 참고하여 토픽의 의미를 파악하고 토픽명칭을 부여한다. 그러나 각 구성단어들은 여러 토픽에 중복적으로 할당될 수 있다. 이것은 단어에 의한 토픽구별을 불명확하게 하여 토픽명칭 부여를 어렵게 하는 원인이 된다. 각 토픽에 할당된 구성단어들은 중요도 가중치를 갖고 있으나 중복 할당된 구성단어들의 중요도 가중치 차이가 크지 않은 경우는 문제가 된다. 특히, 소량의 텍스트 문서에서는 각 토픽들 사이에 중복적으로 할당되는 단어가 많아져서 추출 토픽을 의미적으로 구분하는 것이 어렵게 된다.

본 논문에서는 추출된 토픽들에 대한 명칭부여를 보다 명확히 할 수 있도록, Word2Vec 기법을 토픽모델링과 융합한 형태의 토픽모델링 확장모형을 제안한다. <그림4>는 토픽모델링의 확장모형을 알고리즘 형태로 나타내고 있다. <그림 4>에서, (1)과 (2)의 단계에서는 전통적인 토픽모델링 기법에 의하여 K개의 토픽을 추출한다. (3)의 단계에서는 추출된 각 토픽들의



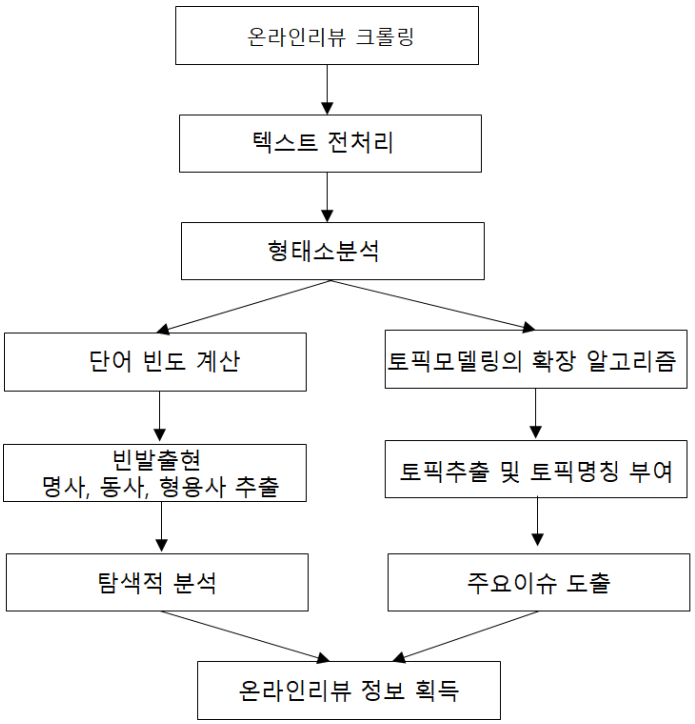
<그림 4> 토픽모델링의 확장 알고리즘

구성단어들인 주요단어들을 검토하여 각 토픽들을 명확하게 구분할 수 있는지 체크한다. 각 토픽들이 명확하게 구분될 수 있으면 (4)단계에서 각 토픽별로 적절할 명칭을 부여하여 분석작업을 종료한다. 각 토픽들의 차별화가 명확하지 않을 경우, (5)단계에서 각 토픽별로 차별단어(discriminatory words)를 추출한다. 차별단어들은 토픽별로 존재할 수 있는데, 해당 토픽에만 포함되고 다른 토픽에는 포함되지 않는 구성단어이다. (6)의 단계에서는 각 토픽의 차별단어들에 대한 연관단어들을 추출하기 위하여 Word2Vec 기법을 적용한다. (7)단계에서는 (6)단계의 결과와 (2)단계의 결과를 바탕으로 각 토픽들의 명칭을 부여한다.

3.2 분석방법

본 논문에서는 온라인리뷰에 포함된 주요단어들에 대한 빈도분석과 토픽모델링을 병행하여 분석하는 것을 제안한다. 토픽모델링의 확장모형을 사용하여 <그림 5>와 같은 방법으로 온라인리뷰를 분석하고자 한다. 주요단어들에 대한 빈도분석을 통하여 온라인리뷰에 대한 탐색적인 분석을 할 수 있으며, 토픽모델링을 통해서는 온라인리뷰에서 주로 언급되는 이슈들을 파악할 수 있다.

온라인리뷰는 크롤러(Crawler)에 의하여 인터넷 웹사이트 등으로부터 자동 수집된다. 수집된 온라인리뷰에서 불필요한 단어들을 제거하는 텍스트 전처리과정이 이루어진다. 텍스트 전



<그림 5> 토픽모델링의 확장알고리즘을 포함한 온라인리뷰 분석흐름도

처리가 적용된 온라인리뷰의 각 문장들은 토큰화(Tokenization) 기법이 적용되어 의미를 갖는 최소단어인 형태소들로 분할된다. 형태소 단어들로부터 명사, 동사, 형용사 품사에 해당하는 단어들만 추출한 후 출현빈도를 계산한다. 형태소 단어들로 변환된 온라인리뷰들은 벡터화(Vectorization) 과정과 확장된 토픽모델링 과정을 거쳐 주요 토픽추출에 활용된다. 온라인리뷰에 나타나는 빈발출현 단어들을 확인함으로써 온라인리뷰에 대한 전체적인 특징을 파악할 수 있다. 추출된 주요 토픽들을 통하여 온라인리뷰에서 언급되는 주요 이슈들을 파악할 수 있다. 온라인리뷰의 전체적인 특징과 주요 이슈들을 확인함으로써 온라인리뷰에 대한 정보를 획득할 수 있다.

Ⅳ. 분석사례

4.1 데이터수집

본 논문에서는 <그림5>에서 제안한 토픽모델링의 확장 알고리즘을 포함한 분석흐름도에 따라, 제주도 우도관광지에 대한 온라인리뷰를 분석하고자 한다. 우도 관광지에 대한 온라인리뷰는 파이썬 크롤러(crawler)를 개발하여 수집하였다. 크롤러는 파이썬 아나콘다 환경에서 셀레니움(Selenium)라이브러리를 이용하여 개발하였다. <표 2>는 수집한 온라인리뷰의 크기 및 데이터구조를 나타내고 있다. 온라인리뷰는 2011년~2019년 사이에 등록한 글인 것으로 나타났다.

<표 2> 온라인리뷰 데이터 구조

리뷰 수	작성기간	데이터구조
521	2011년 ~ 2019년	내용:텍스트(문자형) 평점:숫자형 날짜: 날짜형

4.2 데이터 전처리

온라인리뷰의 내용 열(column)에는 우도에 대한 품평이 한글문장으로 표현되어 있다. 한글문장에는 한글단어들과 함께 은어나 특수문자 등이 포함되어 있다. 한글문장의 전처리과정을 통하여 특수문자, 은어 등의 불용어를 제거하였다. 온라인리뷰의 한글문장들에 대한 형태소를 추출하기 위해서 코엔엘파이(KoNLPy)를 이용하였다. 추출된 형태소 단어들에 대한 빈도계산 및 토픽모델링 전에 의미적 가치가 높지 않다고 판단되는 단어들을 제거하였다.

4.2 단어 빈도분석 및 토픽모델링의 확장 분석

전처리가 적용된 데이터에 대하여 명사, 동사, 형용사 단어의 출현빈도를 계산하고 토픽모델링의 확장 알고리즘을 적용하였다. 온라인리뷰 상에서 추출된 단어들을 단어-문서행렬(Term-Document Matrix)로 벡터화(Vectorization)한 후, 명사, 동사, 형용사 별로 상위 10개의 빈발출현 단어들을 파이썬의 matplot 시각화 라이브러리모듈을 사용하여 막대그래프로 출력하였다. 파이썬의 젠심(gensim) 라이브러리모듈을 사용하여 정수인코딩 형태로 변환된 후, LdaModel()함수에 의하여 토픽 추출에 적용하였다. 적절한 토픽수를

설정하기 위하여 젠심모듈의 CoherenceModel 클래스와 log_perplexity() 등을 사용하여 Coherence와 Perplexity 값을 계산하였다. 또한, 각 토픽별로 차별단어들을 추출하였으며, 차별단어들 각각에 대하여 젠심모듈의 Word2Vec() 함수에 의한 연관단어들을 추출하였다.

4.4 분석 결과

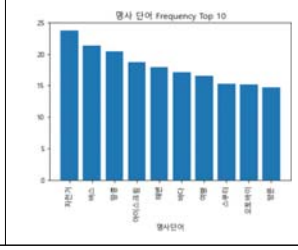
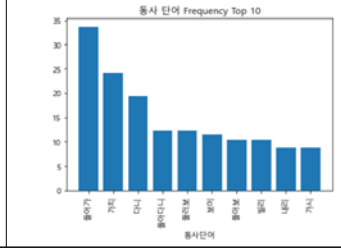
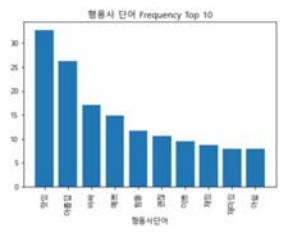
4.4.1 빈발출현 단어 분석

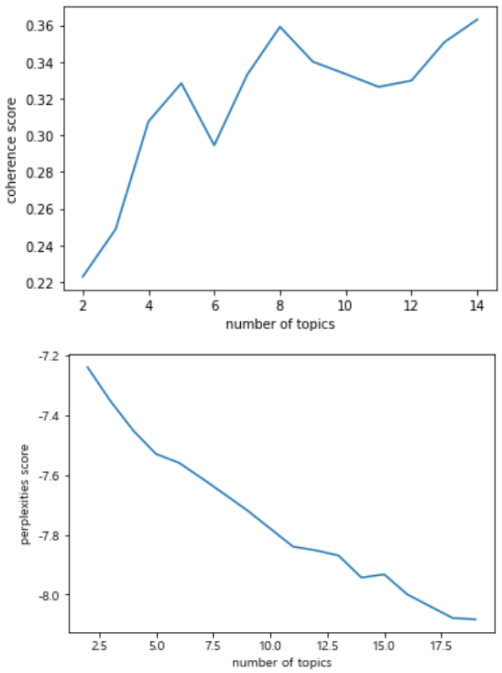
<표 3>은 우도 리뷰에서 자주 출현한 명사, 동사, 형용사 단어들을 나타내고 있다. 명사 품사에서는 '자전거'라는 단어가 가장 자주 출현하였고, 동사 품사에서는 '들어가'('들어가다'의 어간), 형용사 품사에서는 '맛있'('맛있다'의 어간)라는 단어가 빈발하게 출현하고 있었다. '자전거', '버스', '땅콩', '아이스크림' 등의 명사단어가 자주 출현하는 것으로 보아, 우도를 방문하는 관광객들은 버스 관광 또는 자전거 체험을 선호하며, '아이스크림'과 '땅콩' 등을 주요 먹거리로 선호하는 것을 알 수 있다. 특히, 형용사 단어들을 눈여겨 보면 우도관광 패턴을 개략적으로 예측할 수 있다. '맛있'이라는 단어는 먹거리 관광패턴과 관련되어 있으며, '아름답'과 '예쁘'라는 단어는 경치를 즐기는 관광, '재밌'은 체험관광 등의 관광패턴과 관련되어 있는 것으로 예측할 수 있다. '빌리' 라는 동사 단어는 자전거나 오토바이 등의 탈것들을 대여하는 상황을 나타내는 것으로 볼 수 있지만, 전반적으로 동사 단어들은 의미있는 정보를 제공하지 않았다.

4.4.2 토픽모델링 분석

<그림 6>은 토픽모델링 분석과정에서 적절한 토픽 개수를 결정하기 위한 Coherence값과 Perplexity값의 계산 결과를 나타내고 있다. X축에는 토픽개수, Y축에는 Coherence값과 Perplexity값이 대응하여 나타나고 있다. 적합한 토픽 개수는 Coherence값이 클수록 좋고 perplexity값이 작을수록 좋다. <그림 6>에서 토픽수가 5, 8, 15일 때의 Coherence값은 대체적으로 큰 값에 대응하고 있다. Perplexity값은

<표 3> 우도 온라인리뷰에 대한 단어 출현빈도

구분	명사	동사	형용사
단어 목록	자전거,버스,땅콩,아이스크림,해변,바다,여행,스쿠터, 오토바이,방문,추천,투어,전기,하루,생각,날씨,페리, 풍경,구경,사진	들어가,가지,다니,돌아다니,둘러보,보이,돌아보,빌리,내리,가시,걸리,보내,즐기,달리,나오,바라보,느끼,올라가,만들,들리	맛있,아름답,비싸,예쁘,힘들,괜찮,이쁘,재밌,재미있,아쉽
빈도수			



<그림 6> 우도 온라인리뷰의 Coherence와 Perplexity

토픽 수가 많아질수록 지속적으로 반비례하여 작아지고 있다. 일반적으로 토픽 수가 많아질수록 토픽내용은 문서내용을 제대로 반영할 가능성이 높은 반면, 토픽추출의 의미는 퇴색될 수밖에 없다. Perplexity값이 토픽 개수에 반비례하는 상황은 적합한 토픽 개수를 결정하기 위한 좋은 정보가 아닌 것이다. 따라서 여기서는 Coherence값만 고려하여 토픽 수를 2, 3, 5, 8로 결정하였다. 주요 토픽을 응축하기 위하여 2개 또는 3개의 토픽들도 추출하였다.

<표 4>에서는 토픽수가 2, 3, 5, 8 인 각 경우에 대한 토픽모델링 결과를 나타내고 있다. 각 토픽에는 주요단어와 차별단어가 대응되고 있다. 주요단어는 토픽모델링 과정에서 각 토픽과 연관되어 할당된 단어들이지만 다른 토픽들에도 중복하여 나타날 수 있다. 추출된 토픽들의 의미는 주요단어들을 통하여 파악될 수 있지만, 주요단어들의 일부가 다른 토픽에도 중복적으로 나타나고 있어서 각 토픽별 구분, 의미파악, 명칭부여 등의 작업이 쉽지 않음을 알 수 있다. 차별단어는 주요단어 중에서 해당 토픽에만 포함된 단어들이며 다른 토픽에는 포함되지 않는다. 차별단어를 활용하여 토픽별 의미구분과 명칭부여 작업을 진행하는 것이 보다 효과적일 것으로 보인다.

<표 4> 토픽모델링 결과 및 토픽별 차별 단어

토픽 개수	토픽 번호	주요단어	차별단어
2	토픽1	0.011*"땅콩"+ 0.011*"한바퀴" + 0.011*"자전거" + 0.011*"아이스크림" + 0.010*"바다" + 0.008*"해변" + 0.008*"추천" + 0.008*"스쿠터" + 0.007*"오토바이" + 0.007*"풍경	한바퀴, 스쿠터, 오토바이, 풍경
	토픽2	0.015*"버스" + 0.014*"자전거" + 0.012*"여행" + 0.011*"바다" + 0.011*"해변" + 0.010*"땅콩" + 0.010*"아이스크림" + 0.009*"투어" + 0.009*"추천" + 0.008*"방문"	버스, 여행, 투어, 방문
3	토픽1	0.019*"땅콩" + 0.019*"아이스크림" + 0.017*"버스" + 0.012*"해변" + 0.011*"투어" + 0.010*"생각" + 0.009*"바다" + 0.008*"자전거" + 0.008*"여행" + 0.007*"관광"	투어, 생각, 관광

	토픽2	0.017*"자전거" + 0.012*"바다" + 0.012*"여행" + 0.012*"스쿠터" + 0.010*"방문" + 0.009*"추천" + 0.009*"땅콩" + 0.008*"아이스크림" + 0.008*"버스" + 0.008*"해변"	방문
	토픽3	0.014*"자전거" + 0.012*"바다" + 0.010*"추천" + 0.010*"해변" + 0.010*"여행" + 0.009*"스쿠터" + 0.008*"오토바이" + 0.008*"전기" + 0.008*"버스" + 0.007*"바람	오토바이, 전기, 바람
5	토픽1	0.021*"땅콩" + 0.020*"아이스크림" + 0.017*"자전거" + 0.014*"스쿠터" + 0.012*"해변" + 0.010*"한바퀴" + 0.010*"오토바이" + 0.009*"전기" + 0.009*"바다" + 0.008*"여행"	한바퀴, 오토바이
	토픽2	0.021*"바다" + 0.013*"자전거" + 0.010*"해변" + 0.010*"버스" + 0.008*"투어" + 0.008*"코스" + 0.008*"생각" + 0.007*"여행" + 0.007*"땅콩" + 0.007*"자연"	코스, 자연
	토픽3	0.018*"여행" + 0.015*"방문" + 0.015*"추천" + 0.013*"버스" + 0.012*"자전거" + 0.010*"페리" + 0.010*"바다" + 0.009*"땅콩" + 0.009*"아이스크림" + 0.009*"풍경"	방문, 추천, 페리, 풍경
	토픽4	0.012*"스쿠터" + 0.009*"자전거" + 0.009*"날씨" + 0.008*"바다" + 0.008*"전기" + 0.008*"해변" + 0.008*"바람" + 0.008*"자동차" + 0.007*"아이스크림" + 0.007*"버스"	날씨, 바람, 자동차
	토픽5	0.024*"버스" + 0.016*"투어" + 0.012*"해변" + 0.011*"자전거" + 0.009*"바다" + 0.008*"하루" + 0.008*"카페" + 0.008*"관광" + 0.008*"생각" + 0.008*"여행"	하루, 카페, 관광
8	토픽1	0.020*"여행" + 0.014*"자전거" + 0.013*"페리" + 0.013*"하루" + 0.012*"생각" + 0.010*"땅콩" + 0.010*"추천" + 0.009*"바람" + 0.008*"바다" + 0.008*"사진"	생각, 바람
	토픽2	0.023*"자전거" + 0.015*"해변" + 0.011*"스쿠터" + 0.011*"전기" + 0.010*"풍경" + 0.010*"바다" + 0.009*"추천" + 0.009*"느낌" + 0.008*"버스" + 0.008*"페리"	풍경, 느낌
	토픽3	0.020*"오토바이" + 0.014*"추천" + 0.013*"아이스크림" + 0.013*"땅콩" + 0.012*"자전거" + 0.011*"바퀴" + 0.010*"해변" + 0.010*"스쿠터" + 0.009*"투어" + 0.009*"바다"	오토바이
	토픽4	0.019*"버스" + 0.015*"바다" + 0.014*"여행" + 0.012*"자전거" + 0.011*"땅콩" + 0.010*"투어" + 0.010*"아이스크림" + 0.009*"추천" + 0.009*"방문" + 0.009*"해변"	[]
	토픽5	0.023*"버스" + 0.015*"땅콩" + 0.014*"아이스크림" + 0.014*"투어" + 0.012*"해변" + 0.011*"자전거" + 0.011*"바다" + 0.010*"방문" + 0.009*"추천" + 0.008*"관광"	관광
	토픽6	0.012*"바다" + 0.009*"해변" + 0.009*"버스" + 0.007*"자전거" + 0.007*"투어" + 0.007*"여행" + 0.006*"스쿠터" + 0.006*"차량" + 0.006*"하루" + 0.006*"국제"	차량, 국제
	토픽7	0.017*"스쿠터" + 0.014*"바퀴" + 0.014*"아이스크림" + 0.013*"해변" + 0.012*"땅콩" + 0.011*"바다" + 0.011*"자전거" + 0.009*"버스" + 0.008*"여행" + 0.007*"이용"	이용
	토픽8	0.014*"아이스크림" + 0.013*"바다" + 0.012*"방문" + 0.011*"여행" + 0.011*"땅콩" + 0.010*"날씨" + 0.010*"전기" + 0.010*"사진" + 0.009*"자전거" + 0.008*"여름"	날씨, 여름

4.4.3 토픽모델링의 확장모형을 적용한 분석

<표 5>는 토픽모델링의 확장모형을 사용하여 우도 온라인리뷰로부터 2개의 토픽을 추출한 결과를 나타내고 있다. 2개 토픽에 대응하는 차별단어와 각 차별단어 별 연관단어들을 보여주고 있다. 연관단어들은 명사, 형용사, 동사를 포괄하고 있다. 토픽에 포함된 차별단어들과 각 연관단어들을 분석하여 해당 토픽의 의미를 효과적으로 파악하고 명칭을 부여할 수 있다. 토픽1에 포함된 차별단어들은 '한바퀴', '스쿠터', '오토바이', '풍경' 등이다. '한바퀴'는 우도 섬을 한 바퀴 돌면서 관광한다는 의미로 리뷰에 포함되어 있다. '한바퀴', '스쿠터'와 '오토바이' 등의 차별단어와 관련 연관단어들을 통하여 토픽1의 명칭을 '체험관광'(Experience) 으로 부여하였다. 토픽2의 명칭은 '버스', '여행', '투어' 등의 차별단어와 각 연관단어들을 바탕으로 '눈요기관광'(Sightseeing)으로 명명하였다.

<표 6>은 토픽모델링의 확장모형을 사용하여 우도 온라인리뷰로부터 3개의 토픽을 추출한 결과를 나타내고 있다. '눈요기관광'과 '체험관광'에 더하여 '방문느낌'을 언급하는 토픽이 추출되었다. '방문느낌' 토픽은 우도 방문에 대한 느낌을 표현하는 리뷰로서 '눈요기관광'과 '체험관광'을 아우르는 주제라고 할 수 있다.

<표 7>은 토픽모델링의 확장모형을 사용하여 우도 온라인리뷰로부터 5개의 토픽을 추출한 결과를 나타내고 있다. 거칠게 해석해 보면, <표 7>에서의 '체험관광'이 '전동차체험'과 '코스관광'으로, '눈요기관광'이 '휴양관광'과 '렌트카관광'으로 구분된 것으로 볼 수도 있다.

<표 5> 추출된 2개 토픽에 대한 차별화 및 명칭부여

토픽 번호	토픽명	차별단어	연관단어(Top 10)
토픽1	체험 관광	한바퀴	'바다', '많다', '아름답다', '좋다', '카페', '자전거', '버스', '작다', '가다', '추천'
		스쿠터	'깨끗하다', '좋다', '오토바이', '해변', '다니다', '보다', '돌아다니다', '전기차', '여행', '들어가다'
		오토바이	'스쿠터', '맑다', '해변', '가보다', '좋다', '돌아다니다', '깨끗하다', '그렇다', '곳곳', '다니다'
		풍경	'보트', '장소', '예쁘다', '해변', '다니다', '버스', '추천', '돌다', '좋다', '카페'
토픽2	눈요기 관광	버스	'가다', '자전거', '추천', '보다', '좋다', '아름답다', '많다', '들어가다', '한퀴', '해변'
		여행	'땅콩', '추천', '다니다', '둘러보다', '해변', '아이스크림', '좋다', '관광객', '렌트', '자전거'
		투어	'이동', '바람', '사진', '좋다', '방문', '해변', '버스', '가다', '쉬다', '추천'
		방문	'바다', '한바퀴', '가보다', '서빈백사', '많다', '멋지다', '좋다', '가다', '이동', '남다'

<표 6> 추출된 3개 토픽에 대한 차별화 및 명칭부여

토픽 번호	토픽명	차별단어	연관단어(Top 10)
토픽1	눈요기 관광	투어	'이동', '바람', '사진', '좋다', '방문', '해변', '버스', '가다', '쉬다', '추천'
		생각	'다니다', '많다', '땅콩', '맛있다', '해변', '바다', '구경', '자전거', '추천', '한바퀴'
		관광	'조용하다', '느끼다', '자전거', '가족', '비싸다', '아름답다', '둘러보다', '보다', '타다', '투어'
토픽2	방문 느낌	방문	'바다', '한바퀴', '가보다', '서빈백사', '많다', '멋지다', '좋다', '가다', '이동', '남다'
토픽3	체험 관광	오토바이	'스쿠터', '맑다', '해변', '가보다', '좋다', '돌아다니다', '깨끗하다', '그렇다', '곳곳', '다니다'
		전기	'이용', '느끼다', '돌다', '장소', '해안', '기억', '하루', '성산', '느낌', '즐기다'
		바람	'이동', '투어', '사진', '방문', '좋다', '해변', '가보다', '최고', '서빈백사', '바다'

<표 7> 추출된 5개 토픽에 대한 차별화 및 명칭부여

토픽 번호	토픽명	차별단어	연관단어(Top 10)
토픽1	전동차 체험	한바퀴	'바다', '많다', '아름답다', '좋다', '카페', '전기', '버스', '작다', '가다', '추천'
		오토바이	'스쿠터', '맑다', '해변', '가보다', '좋다', '돌아다니다', '깨끗하다', '그렇다', '곳곳', '다니다'
토픽2	코스 관광	코스	'오다', '들어가다', '티켓', '보이다', '깨끗하다', '버스', '사진', '돌다', '좋다', '바람'
		자연	'날씨', '함께', '스쿠터', '다니다', '선택', '버스', '깨끗하다', '사진', '투어', '등대'
토픽3	방문 느낌	방문	'바다', '한바퀴', '가보다', '서빈백사', '많다', '멋지다', '좋다', '가다', '이동', '남다'
		추천	'가다', '해변', '바다', '여행', '좋다', '아이스크림', '남다', '보다', '보트', '버스'
		페리	'아름답다', '방문', '많다', '남다', '가다', '이동', '투어', '바다', '추천', '좋다'
		풍경	'보트', '장소', '예쁘다', '해변', '다니다', '버스', '추천', '돌다', '좋다', '카페'
토픽4	렌트카 관광	날씨	'함께', '다니다', '자연', '기억', '아이스크림', '스쿠터', '자전거', '버스', '등대', '여행'
		바람	'이동', '투어', '사진', '방문', '좋다', '해변', '가보다', '최고', '서빈백사', '바다'
		자동차	'가격', '크다', '경험', '보이다', '작다', '바다', '가다', '돌다', '보내다', '한바퀴'
토픽5	휴양 관광	하루	'도로', '이용', '땅콩', '느끼다', '여행', '대여', '추천', '보트', '전기', '아이스크림'
		카페	'한바퀴', '서빈백사', '풍경', '예쁘다', '바다', '좋다', '유명하다', '버스', '들어가다', '빌리다'
		관광	'조용하다', '느끼다', '자전거', '가족', '비싸다', '아름답다', '둘러보다', '보다', '타다', '투어'

<표 8> 추출된 8개 토픽에 대한 차별화 및 명칭부여

토픽번호	토픽명	차별단어	연관단어(Top 10)
토픽1	생각 바람	생각	'다니다', '많다', '땅콩', '맛있다', '해변', '바다', '구경', '자전거', '추천', '한바퀴'
		바람	'이동', '투어', '사진', '방문', '좋다', '해변', '가보다', '최고', '서빈백사', '바다'
토픽2	방문 느낌	풍경	'보트', '장소', '예쁘다', '해변', '다니다', '버스', '추천', '돌다', '좋다', '카페'
		느낌	'장소', '성산', '가다', '아이스크림', '추천', '등대', '돌다', '이용', '전기', '느끼다'
토픽3	전동차 체험	오토바이	'스쿠터', '맑다', '해변', '가보다', '좋다', '돌아다니다', '깨끗하다', '그렇다', '곳곳', '다니다'
토픽4	관광	[]	
토픽5	관광	관광	'조용하다', '느끼다', '자전거', '가족', '비싸다', '아름답다', '둘러보다', '보다', '타다', '투어'
토픽6	전동차 체험	차량	'어렵다', '전기차', '돌아다니다', '2시간', '구석구석', '느끼다', '도보', '개인', '괜찮다', '빌리다'
		국제	'보이다', '가격', '나오다', '자동차', '경험', '차갑다', '곳곳', '많아지다', '해변', '바람'
토픽7	관광 정보	이용	'전기', '하루', '아이스크림', '기억', '맛있다', '장소', '즐기다', '땅콩', '돌다', '느낌'
토픽8	날씨 정보	날씨	'함께', '다니다', '자연', '자전거', '아이스크림', '스쿠터', '기억', '버스', '여행', '가다'
		여름	'유명하다', '먹다', '해변', '멋지다', '여행', '추천', '돌아보다', '맛있다', '서빈백사', '좋다'

<표 8>은 토픽모델링의 확장모형을 사용하여 우도 온라인리뷰로부터 8개의 토픽을 추출한 결과를 나타내고 있다. 토픽1처럼 토픽에 적합한 명칭이 없을 경우, 차별단어의 조합으로 명칭을 부여할 수 있다. 토픽4의 경우는 차별단어가 존재하지 않고 있다. 이러한 경우는 주요 단어만으로 토픽명칭을 부여할 수밖에 없다. 8개의 토픽을 추출한 결과, 정보를 제공하는 토픽들이 추가적으로 도출되었다.

Ⅴ. 결 론

온라인리뷰는 상품이나 서비스의 사용 결과에 대하여 사람들이 생각한 바를 기록한 디지털 데이터이다. 스마트폰이나 SNS 등이 보편화됨에 따라 온라인리뷰의 생성은 더욱 증가되고 있으며, 컴퓨팅 기술의 발전으로 온라인리뷰의 분석환경도 좋아졌다. 기업이 생산한 상품이나 서비스 등에 대한 고객들의 생각을 기록한 데이터가 온라인리뷰이기 때문에, 온라인리뷰를 제대로 분석할 수 있으면 마케팅전략 수립 등

경영분야의 중요한 의사결정에 대한 유의미한 정보를 만들어낼 수 있다.

토픽모델링은 온라인리뷰에서 주로 언급되는 주제들을 추출하는 기법으로서, 온라인리뷰를 분석하는 방법 중의 하나이다. 토픽모델링을 통하여 각 토픽에 할당된 주요단어들은 여러 개의 토픽에 중복 할당될 수 있기 때문에 토픽 구분과 명칭부여작업을 어렵게 하는 원인이 될 수 있었다.

본 논문에서는 기존의 토픽모델링 기법을 보완한 확장모형을 제안하였다. 토픽모델링 확장모형은 기존의 토픽모델링기법에 Word2Vec기법을 접목한 형태로서, 기존의 토픽모델링에 의하여 추출된 각 토픽들에 대하여 차별단어와 연관단어를 도출하고, 이를 바탕으로 각 토픽에 대한 의미 분석과 명칭부여 작업을 진행한다. 토픽모델링의 확장모형은 규모가 크지 않은 온라인리뷰에 대하여 전통적인 토픽모델링기법보다 토픽의 의미를 보다 명확하게 구분할 있도록 하였으며 토픽명칭을 보다 효과적으로 부여할 수 있도록 하였다.

본 논문에서는 토픽모델링의 확장모형을 포함한 온라인리뷰의 분석방법도 제안하였다. 분석방법은 단어의 출현빈도분석과 토픽모델링의 확장모형을 병행하는 방식이었다. 단어의 출현빈도 분석을 통하여 온라인리뷰의 전체적인 특징을 파악할 수 있었다.

토픽모델링 확장모형의 효과성을 확인하기 위하여 우도관광지 온라인리뷰의 분석사례를 제시하였다. 기존의 토픽모델링 결과로는 추출된 토픽들 사이의 의미구분과 명칭부여 작업이 어렵다는 것을 확인할 수 있었다. 토픽모델링의 확장모형을 사용하여 각 토픽별 차별단어와 각 차별단어의 연관단어들을 바탕으로 토픽명칭을 부여하면서 토픽추출의 효과성을 확인할 수 있었다.

본 논문에서 제안한 확장알고리즘은 경영정보학 분야에서 텍스트 기반의 빅데이터 분석연구의 정확성을 높이는데 기여할 수 있다. 제안한 확장모형은 기존의 토픽모델링보다 토픽구분을 쉽게 할 수 있는 효과를 제공할 수 있음에도 여전히 한계를 갖는다. 추출된 토픽을 의미적으로 구분하고 명칭을 부여하는 주체는 분석자이기 때문에 사람의 주관성이 개입될 수밖에 없다. 우도 온라인리뷰를 분석하는 과정에서 이러한 한계점에 봉착하기도 하였다. 이러한 한계점을 극복할 수 있는 추가적인 연구들이 일어나기를 기대한다.

참고문헌

김정규, 정철, “특허 정보를 활용한 한국과 미국의 관광 관련 기술 동향 분석 : 토픽모델링을 중심으로”, 관광학연구, 제43권, 제1호, 2019, pp.249-267.

남승주, 김준환, 유영준, “텍스트 분석 기반의 쇼핑 관광객 인식 분석”, 신산업경영저널, 제38권, 제2호, 2020, pp.3-21.

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

이민철, 김혜진, “텍스트 마이닝 기법을 적용한 뉴스 데이터에서의 사건 네트워크 구

축”, 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.

이진욱, 유국현, 문병민, 배석주, “감성분석과 Word2vec을 이용한 비정형 품질 데이터 분석”, 품질경영학회지, 제45권, 제1호, 2017, pp.117-127.

이창용, 김근형, “온라인리뷰의 랭킹모델링을 위한 양과 질의 인과모형 분석 ”, 정보시스템연구, 제28권, 제1호, 2019, pp.1-116.

이혜진, 강영옥, “토픽모델링과 LSTM기반 텍스트 분석을 통한 부산방문 외국인 관광객의 선호관광지 및 관광매력요인 분석”, 한국도시지리학회지, 제23권, 제3호, 2020, pp.61-70.

정영진, 조윤호,“온라인 구매 행동을 고려한 토픽모델링 기반 도서 추천”, 지식경영연구, 제18권, 제4호, 2017, pp.97-118.

차윤정, 이지혜, 최지은, 김희웅, “소셜미디어 토픽모델링을 통한 스마트폰 마케팅 전략 수립 지원”, 지식경영연구, 제16권, 제4호, 2015, pp.69-87.

야오즈옌. 김은미, 홍태호, “온라인리뷰의 텍스트마이닝에 기반한 한국방문 외국인 관광객의 문화적 특성 연구”, 정보시스템연구, 제29권, 제4호, 2020, pp.171-191.

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.

Kyung Im Kim, Nguyen Cao Truong Hai, and Hyuk Ro Park, “SVD-LDA: A Combined Model for Text Classification”, *JIPS(Journal of Information Processing Systems)*, Vol.5, No.1, 2009, pp.5-10.

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

Moody, C. E., “Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2vec”, 2016, arXiv Preprint arXiv: 1605.02019.

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.

Qiang, J., Chen, P., Wang, T., and Wu, X.,“Topic Modeling Over Short Texts by Incorporating Word Embeddings”, *In Pacific-asia Conference on Knowledge Discovery and Data Mining*, 2017, pp. 363-374.

Won-joon Choi and Euhee Kim, “A Large-scale Text Analysis with Word Embeddings and Topic Modeling, *Journal of Cognitive Science*, Vol.20, No.1, 2019, pp.147-187.

윤 상 훈 (Yoon, Sang Hun)

영동대학교 시각디자인학사와 제주대학교 석사학위를 취득하였다. 현재 제주대학교 대학원 경영정보학과 박사과정 중에 있다. 주요 관심분야는 텍스트마이닝, 토픽모델링 등이다.

김 근 형 (Kim, Keun Hyung)

서강대학교 컴퓨터학과에서 학사/석사/박사학위를 취득하였다. 현재 제주대학교 경영정보학과 교수로 재직하고 있으며, 주요 관심분야는 데이터마이닝, 텍스트마이닝 등이다.

<Abstract>

Expansion of Topic Modeling with Word2Vec and Case Analysis

Yoon, Sang Hun · Kim, Keun Hyung

Purpose

The traditional topic modeling technique makes it difficult to distinguish the semantic of topics because the key words assigned to each topic would be also assigned to other topics. This problem could become severe when the number of online reviews are small. In this paper, the extended model of topic modeling technique that can be used for analyzing a small amount of online reviews is proposed.

Design/methodology/approach

The extended model of being proposed in this paper is a form that combines the traditional topic modeling technique and the Word2Vec technique. The extended model only allocates main words to the extracted topics, but also generates discriminatory words between topics. In particular, Word2vec technique is applied in the process of extracting related words semantically for each discriminatory word. In the extended model, main words and discriminatory words with similar words semantically are used in the process of semantic classification and naming of extracted topics, so that the semantic classification and naming of topics can be more clearly performed. For case study, online reviews related with Udo in Tripadvisor web site were analyzed by applying the traditional topic modeling and the proposed extension model. In the process of semantic classification and naming of the extracted topics, the traditional topic modeling technique and the extended model were compared.

Findings

Since the extended model is a concept that utilizes additional information in the existing topic modeling information, it can be confirmed that it is more effective than the existing topic modeling in semantic division between topics and the process of assigning topic names.

Keyword: Topic Modeling, Word2vec, Online Reviews, Udo Island, Big Data

* 이 논문은 2020년 2월 24일 접수, 2020년 3월 9일 1차 심사, 2020년 3월 9일 게재 확정되었습니다.