「정보시스템연구」 제33권 제3호
한국정보시스템학회
2024년 9월, pp. 105～121

http://dx.doi.org/10.5859/KAIS.2024.33.3.105

버토픽과 텍스트랭크의 융합을 통한 토픽모델링의 개선 및 사례 분석*

김근형** · 강재정***

〈목 차〉

Ⅰ. 서 론

토픽모델링(Topic modeling)은 대량의 텍스트 데이터로부터 주요 토픽 즉, 핵심 주제들을 추출할 수 있도록 하는 텍스트마이닝 기반의 데이터 분석 기법이다. 토픽모델링을 통하여 텍스트 데이터로부터 인사이트를 도출할 수 있기 때문에 비즈니스 및 사회과학 연구 등 다양한 도메인에서 많이 응용되고 있다.

토픽모델링에 대한 수요가 많아짐에 따라, 토픽 추출의 정확성을 개선할 목적으로 많은 알고리즘들이 개발되었다. 토픽모델링을 위한 주요 알고리즘들은 LDA(David M Blei et al., 2003; 김근형, 2022) 및 NMF(Cédric Févotte and Jérôme, 2011)와 같이 단어주머니(Bag-of-words) 기반으로 토픽을 추출하는 유형과 BERT(Bidirectional Encoder Representations from Transformers) 등과 같이 임베딩 기술 기반으로 토픽을 추출하는 유형이 있다. 단어주머니 기반의 토픽모델링 알고리즘은 문서 안에 여러 잠재 토픽들이 포함되었다는 가정하에, 해당 토픽과 관련된 단어의 출현확률을 계산하여 확률이 높은 단어들을 해당 토픽에 할당한다. 이러한 알고리즘은 단어의 출현빈도에 따른 해당 토픽으로의 출현확률을 계산하는 방식이므로, 단어 간의 의미적 관계 및 문장의 맥락을 고려하지 못하는 한계가 있다.

반면, 워드임베딩 기반의 토픽모델링 알고리

* 이 논문은 2024학년도 제주대학교 교원성과지원사업에 의하여 연구되었음.
** 제주대학교 경영정보학과 교수, khkim@jejunu.ac.kr(주저자)
*** 제주대학교 경영학과 교수, jaejung@jejunu.ac.kr(교신저자)

즘은 단어 간 의미적 관계를 고려하면서 토픽을 추출한다. 워드임베딩(Word embedding)은 단어 및 문장을 다차원 벡터 공간 상의 벡터값으로 표현함으로써, 단어 및 문장 간의 의미적 관계와 맥락을 고려할 수 있도록 한다. BERT는 단어 및 문장을 벡터공간에 표현하여 의미적 단어검색 뿐만 아니라 맥락이 고려된 문장을 생성하는데 훌륭한 결과를 보여 주었다(Devlin et al., 2018; 김효곤, 유동희, 2022). 버토픽(Bertopic)은 BERT를 토픽모델링 기법에 적용함으로써 토픽 추출의 정확성을 높였다. 버토픽은 문서집합의 각 문서가 하나의 토픽을 포함하고 있다고 가정하여, 먼저 임베딩 기반의 군집화를 수행하고 각 군집을 토픽으로 간주하면서 각 토픽에 대한 토픽표현(Topic Representation)을 생성한다(Maarten Grootendorst, 2023). 토픽표현은 각 토픽을 대표하는 핵심단어들과 문장들로 구성된다. 버토픽에서는 TF-IDF의 변형인 c-TFIDF 기법을 사용하여 각 토픽에 대한 핵심단어를 추출한다. c-TFIDF는 해당 군집에는 자주 출현하지만 다른 군집에는 덜 출현하는 단어 및 문장에 중요 가중치를 부여한다. 다른 토픽과의 차별성을 높이는 단어를 핵심단어로 선택하려는 전역적 접근 방법이라 할 수 있다. 이러한 방식은 토픽 내에서는 맥락적으로 핵심 단어일지라도, 다른 토픽과의 차별성을 높이지 못하면 중요 단어로 선택되지 못하게 함으로써, 결국 토픽표현이 왜곡될 수 있다. 이러한 현상을 보완하기 위하여 추가적인 토픽표현을 생성할 필요가 있다. 즉, 다른 토픽과는 독립적인 관점에서 해당 토픽의 중요 단어 및 문장들을 중심으로 한 요약텍스트를 각 토픽 별로 도출하여 추가적인 토픽표현을 생성하는 것이다. 다른 토픽과 독립적으로 해당토픽의 문서들만을 고려하여 핵심단어 및 문장을 선택하는 지역적 접근 방법을 도입할 필요가 있다.

텍스트랭크(TextRank)는 페이지랭크(PageRank)에 기반한 텍스트요약 기법으로서, 대량의 텍스트로부터 핵심단어 및 문장을 추출하여 요약텍스트를 생성하는 기봅이다(Rada Mihalcea and Paul Tarau, 2004). 텍스트랭크 방법에 임베딩 기술을 적용함으로써 단어 및 문장 간의 의미적 관계가 고려된 요약텍스트를 도출할 수 있다. 텍스트랭크 기법을 각 토픽에 적용하여 지역적 접근 관점의 토픽표현을 생성할 수 있다.

한편, 토픽모델의 최적 토픽 수는 토픽 내의 응집도(Coherence)와 토픽 간의 다양도(Diversity)를 최대화할 수 있도록 선택되어야 한다(Maarten Grootendorst, 2023). 응집도는 토픽 내의 단어 및 문장들 간에 서로 연관성이 클수록 높으며, 다양도는 토픽 간에 핵심단어 및 문장들의 중복이 많을수록 낮다. 최적의 토픽 수는 최적의 토픽모델을 도출하는데 영향을 미치기 때문에 응집도와 다양도를 보다 정확하게 측정할 수 있는 방법이 필요하다.

본 논문에서는 버토픽에서 사용하는 전역적 접근의 한계를 극복하기 위하여 지역적 접근을 융합한 토픽모델링 방안을 제안하고자 한다. 버토픽 기반의 토픽모델링 과정에서 텍스트랭크 방법을 도입하여 토픽표현 및 다양도 측정의 정확성을 높임으로써 토픽모델링의 성능을 개선할 수 있는 방법을 제안하고자 한다. 또한, 토픽표현 과정에서 텍스트랭크 방법을 융합하여 별도의 요약텍스트를 생성함으로써, c-TFIDF

기반 토픽표현의 한계를 보완하고자 한다. 이를 위하여, 텍스트랭크 결과에 의한 토픽별 핵심 문장들을 기반으로 추가적인 다양도 지표를 계산하여 최적 토픽 수를 보다 정확하게 선택할 수 있도록 한다. 특히, 토픽표현 과정에서 ChatGPT를 활용함으로써 토픽레이블링 과정의 효율성을 높이고자 한다.

본 논문에서 제안하는 토픽모델링 방법을 사용하여 관광후기 텍스트를 분석하고 성능 평가를 수행함으로써, 새롭게 제안한 방법의 효과성을 확인하고자 한다.

Ⅱ. 이론적 배경

2.1 토픽모델링

토픽모델링은 대량의 텍스트로부터 공통 주제 또는 이야기 흐름 등을 추출할 수 있는 텍스트 분석 방법이다. LDA(Latent Dirichlet Allocation) 또는 NMF(Non-Negative Matrix Factorization) 등의 방법은 단어주머니 기반으로 공통 주제들을 추출하기 때문에 단어 간의 의미적 관계가 고려되지 않은 불명확한 토픽들이 추출되는 경우가 있다. 워드임베딩 기술은 텍스트의 단어 간 또는 문장 간의 의미적 관계를 고려하여 텍스트를 분석할 수 있게 한다. BERT는 임베딩 기술을 기반으로 텍스트 내 문장 또는 단어 간의 의미적 문맥(context) 관계를 고려하면서 텍스트 분석을 할 수 있도록 하는데, 버토픽은 BERT를 기반으로 하는 토픽모델링 방법이다(Maarten Grootendorst, 2023).

2.1.1 버토픽

버토픽은 대량의 텍스트로부터 3단계의 과정을 거쳐 토픽을 추출한다. 1단계에서는 사전학습 언어모델(Pre-trained language model)을 사용하여 텍스트의 각 문서들을 임베딩 벡터값으로 변환한다. 2단계에서는 임베딩 벡터값으로 변환된 문서들을 군집화한다. 군집화 성능을 높이기 위한 방편으로, 임베딩 벡터의 차원(Dimension)을 축소한다. 3단계에서는 각 군집들에 대한 토픽표현을 생성한다. 토픽표현의 수단인 토픽별 핵심단어들은 TF-IDF의 변형인 c-TFIDF(classic TFIDF)를 사용한다.

1단계의 문서 임베딩 과정에서는 텍스트의 각 문서들이 의미적으로 비교될 수 있도록 벡터공간의 다차원 벡터값으로 변환된다. 동일한 토픽에 포함되는 문서들은 의미적으로 유사할 것이라는 가정이 깔려있다. 버토픽에서는 SBERT(Sentence-BERT) 프레임워크를 사용하여 텍스트의 문서들을 벡터값으로 변환한다. SBERT는 텍스트의 문장들을 임베딩 벡터값으로 변환하기 위하여 사전학습 언어모델을 사용한다(Reimers and Gurevych, 2019). SBERT를 향상된 임베딩 기술로 대체하여 버토픽의 성능을 개선할 수도 있다.

2단계에서는 임베딩 값으로 변환된 문서들에 대하여 HDBSCAN(Hierarchical Density-based spatial clustering of applications with noise)를 사용하면서 군집화가 이루어진다(McInnes et al., 2017). HDBSCAN은 노이즈(noise) 값들이 이상치(outlier)로 군집화될 수 있도록 한다. 이러한 방법은 관련 없는 문서가 임의의 군집에 할당되는 것을 방지하여 보다 명확한 토픽 표현이 도출될 수 있도록 한다. 따

라서, 도출될 군집 수를 수동으로 설정하더라도 도출된 군집 개수는 수동 설정값 보다 적어질 수 있다. 버토픽에서는 군집화에 들어가기 전에 UAMP(Uniform Manifold Approximation and Projection for Dimension Reduction) 방법을 사용하여 임베딩 벡터의 차원을 축소한다 (McInnes et al., 2018). 임베딩 벡터의 차원이 증가함에 따라 가장 가까운 점까지의 거리와 가장 먼 점까지의 거리가 비슷해져 버리는 현상이 나타날 수 있다(Charu C Aggarwal et al. 2001). 따라서, 효율적인 군집화와 그 결과의 성능을 높이기 위해서는 임베딩 벡터의 차원을 축소할 필요가 있다.

3단계에서는 도출된 각 군집들로부터 핵심 단어들을 추출한다. 각 군집을 대표하는 핵심단어들은 해당 군집에서는 중요하지만 다른 군집에서는 덜 중요한 단어들이어야 한다. 버토픽에서는 이러한 단어들을 찾기 위하여 TF-IDF 방법의 변형인 c-TFIDF 방법을 사용한다. TF-IDF 방법은 문서 관점에서 차별화된 단어 즉, 중요 단어들을 찾는 반면, c-TFIDF 방법은 문서들로 구성된 문서군집 관점에서 차별화된 중요 단어들을 찾는다. 즉, c-TFIDF는 개별문서 단위가 아니라 문서군집 단위로 차별단어를 찾는다는 측면에서 TFIDF를 개량한 것이라 할 수 있다. 식(1) 은 c-TFIDF가 각 군집의 핵심단어를 찾는 모델을 나타낸다. Wt,c는 군집 c에 속하는 단어 t의 중요도를 의미한다. tft,c는 군집 c에 나타나는 단어 t의 출현 빈도를 나타낸다. A는 각 군집의 평균 단어 개수이다. tft는 단어 t가 전체 군집에서 출현하는 빈도수를 의미한다. 식(1)을 바탕으로 각 군집에 속하는 단어들의 중요도를 계산하여 중요도가 높은 순으로 핵심 단어들이 추출된다.

$$W_{t,c} = tf_{t,c} \cdot \log(1 + \frac{A}{tf_t}) \quad \text{----- 식(1)}$$

2.1.2 응집도와 다양도

토픽모델링에 의하여 생성된 토픽모델의 성능은 추출된 토픽들이 서로 중복되지 않으면서 다양하게 존재할수록 바람직하다. 토픽모델의 성능은 응집도(Coherence)와 다양도(Diversity) 지표로 평가할 수 있으며 그 값이 클수록 좋다.

응집도는 해당 토픽에 속하는 단어 간의 의미적 연관성이 얼마나 높은지를 평가하는 지표이다. NPMI(Normalized Point-wise Mutual Information)은 응집도를 계산하는 방법 중 하나이다(Gerlof Bouma, 2009). NPMI는 기존의 PMI(Point-wise Mutual Information) 지표를 [-1 ~1] 의 범위로 정규화한 지표이다. NPMI값이 1에 가까울수록 토픽의 응집도는 높다. 식(2)는 PMI를 계산하는 식이다. $P(w_i)$는 단어 w_i가 해당 토픽에 등장할 확률이고, $P(w_j)$는 단어 w_j가 해당 토픽에 등장할 확률이며, $P(w_i,w_j)$는 단어 w_i와 w_j가 해당 토픽에 같이 등장할 확률을 의미한다. $PMI(w_i,w_j)$의 값이 높다면 단어 w_i와 w_j의 연관성이 높다는 의미가 된다.

$$\mathrm{PMI}(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}$$

----- 식(2)

다양도는 추출된 토픽 간의 유사성을 측정하는 지표이다. 추출된 토픽 간에는 유사하지 않을수록 좋다. 토픽 간의 유사도는 토픽 간 비교쌍을 구성하여 토픽 별 핵심단어들 사이의 코

사인(cosine) 거리를 계산하고, 모든 비교 쌍에 대한 코사인 거리의 평균값으로 측정할 수 있다. 코사인 거리는 코사인 유사도로부터 도출된다. 식(3)은 벡터 X와 Y사이의 코사인 유사도를 계산하는 식이고, 식(4)는 코사인 거리를 계산하는 수식이다. $\|X\|_2$는 X의 벡터값들의 제곱합에 대한 제곱근을 의미한다.

$$\text{유사도}(X, Y) = \frac{X \cdot Y}{\|X\|_2 \cdot \|Y\|_2} \quad \text{----- 식(3)}$$

$$\text{코사인거리}(X,Y) = 1 - \text{유사도}(X,Y) \quad \text{----- 식(4)}$$

이때, 단어들의 임베딩 값을 기준으로 코사인 거리를 계산할 수 있다(Silvia Terragni et al., 2021). 비슷한 방식으로 토픽 별 핵심문장을 이용하여 코사인 거리를 계산할 수 있으며, 다양도 지표로 사용할 수 있다. 다양도 지표는 [0,1]의 범위를 가지며, 1에 가까울수록 토픽 간에 유사하지 않은 것이며 중복성이 낮음을 의미한다.

2.2 텍스트 요약

텍스트 요약(Text Summarization)은 상대적으로 큰 텍스트 원문을 핵심 내용만 간추려서 상대적으로 작은 텍스트 요약문으로 변환하는 것을 말한다. 텍스트 요약은 크게 추출적 요약(Extractive summarization)과 추상적 요약(Abstractive summarization)으로 구분할 수 있다. 추출적 요약은 원문에서 중요한 핵심 문장 또는 단어구를 몇 개 뽑아서 이들로 요약문을 구성하는 방법이다. 추상적 요약은 핵심 문맥(context)을 고려하여 원문에 없더라도 새로운 문장을 생성하여 원문을 요약하는 방법이다. 추상적 요약을 인공신경망으로 훈련하기 위해서는 원문뿐만 아니라 '실제 요약문'이라는 레이블 데이터가 있어야 한다. 반면, 추출적 요약은 레이블 데이터가 필요 없어서 다양한 텍스트에 적용할 수 있다.

추출적 요약의 대표적인 알고리즘으로 텍스트랭크가 있다. 텍스트랭크 알고리즘은 페이지랭크(PageRank) 알고리즘을 텍스트 데이터에 적용할 수 있도록 변형한 것이다(Rada Mihalcea and Paul Tarau., 2004). 텍스트랭크 알고리즘은 텍스트에서 중요한 단어 또는 문장의 중요도 순위를 계산하여 중요도가 높은 단어 또는 문장들로 이루어진 텍스트 요약 결과를 생성한다. 텍스트의 각 단어 또는 문장을 그래프 구조의 정점(Vertex)으로 놓고, 연관성이 있는 단어 간 또는 문장 간에는 간선(Edge)을 연결한다. 정점 간 연관성 정도에 따라 초기 가중치가 설정된다. 정점 V_i의 중요도 점수 $WS(V_i)$는 식(5)와 같이 계산된다.

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

----- 식(5)

식(5)에서 $In(V_i)$는 정점 V_i를 가르키는 정점의 집합을 의미하며, $Out(V_i)$는 정점 V_i가 가르키는 정점의 집합을 의미한다. 방향 그래프인 경우에는 $In(V_i)$와 $Out(V_i)$가 다르지만, 무방향 그래프인 경우에는 동일하다. w_{ji}는 정점 V_j와 V_i 사이에 연결된 간선의 가중치를 의미한다. d는 페이지랭크 알고리즘에서는 그래프의 간선을 통하여 다른 정점으로 이동할 확률을 의미

한다. 즉, 텍스트랭크에서는 간선으로 연결된 정점 간에 연관성이 있을 확률이 될 것이며, (1 - d)는 정점 간의 간선이 없더라도 연관성이 존재할 확률을 의미하는 것으로 볼 수 있다. 정점 간의 초기 가중치 계산은 정점이 단어인 경우와 문장인 경우가 다르다. 정점이 단어인 경우는 어떤 문장에 동시 출현하는 단어 간에 간선이 연결되며 정점 간 가중치는 그 정점들이 동시 출현하는 문장의 수에 비례한다. 정점이 문장인 경우는 문장 간 유사도에 비례하여 가중치가 설정된다. 문장 S_i는 단어들 $w_1^i, w_2^i, \ldots, w_{N_i}^i$로 구성되고, S_j는 단어들 $w_1^j, w_2^j, \ldots, w_{N_j}^j$로 이루어질 때, 문장 S_i와 문장 S_j간 유사도 Similarity(S_i, S_j)는 식(6)과 같이 계산된다.

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \ \& \ w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)}$$

----- 식(6)

텍스트랭크 알고리즘에서도 정점 간 가중치를 계산할 때 임베딩 기술을 적용할 수 있다. 본 논문에서는 문장 기반 텍스트랭크 알고리즘에 임베딩 기법을 적용한다. 텍스트의 각 문장들을 임베딩 벡터로 변환하여 임베딩 벡터 간 코사인 유사도를 계산한다.

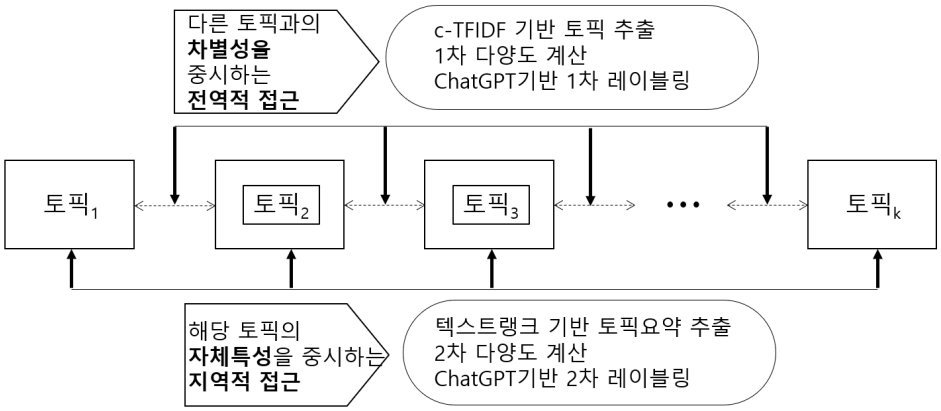
Ⅲ. 연구설계

버토픽은 임베딩 기술을 활용한 토픽모델링 방법으로서, 기존의 방법들보다는 좋은 성능을 보이지만. 몇 가지 한계가 있다. 첫째, 각 토픽에 할당된 핵심단어 및 문장들은 다른 토픽과의 차별성 에 가중치를 주는 전역적 방식이어서, 토픽표현 과정에서는 해당 토픽의 실질적 중요단어 및 문장은 선택되지 못할 가능성이 있다. 따라서, 다른 토픽과 독립적으로 토픽 자체의 내용만이 반영된 지역적 토픽표현이 도출될 필요가 있다. 즉, 각 토픽에 대한 토픽표현 과정에서 다른 토픽과 독립적으로 추출된 지역적 특성 기반의 2차 레이블이 설정되면 더 명확한 토픽표현을 도출할 수 있을 것이다. 둘째, 토픽모델의 최적 토픽 갯수를 결정할 때 사용되는 응집도와 다양도 값은 토픽모델 간 그 값의 편차가 클수록 최적 토픽모델을 선택하기 위한 의사결정이 효과적일 것이다. 즉, 토픽모델에 따라 그 값의 편차가 큰 추가적인 지표를 개발할 필요가 있다. 셋째, 각 토픽에 할당된 핵심단어 및 문장으로 구성되는 토픽 레이블은 분석자의 수작업 과정을 통하여 도출되기 때문에, 단어와 문장이 많아지면 레이블 설정의 정확도가 떨어질 수 있다. 핵심단어 및 문장이 많을 경우, ChatGPT를 활용하여 레이블을 설정하면 보다 효율적일 수 있다.

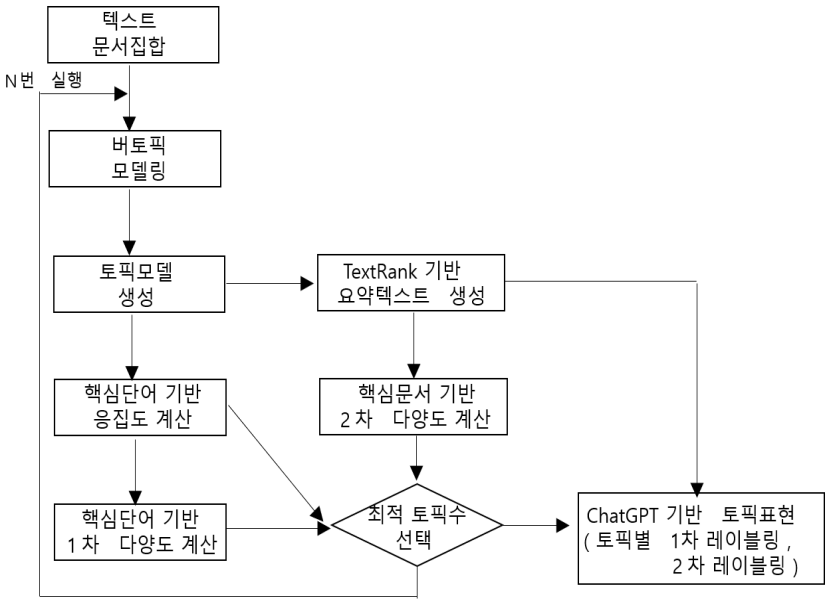
이러한 기존 버토픽의 한계를 극복하기 위하여 본 논문에서는 몇가지 개선 방안을 제안한다. 첫째, 버토픽 기반 토픽모델링 과정에 텍스트랭크 알고리즘을 융합하여 각 토픽에 대한 지역적 2차 레이블을 설정한다. 각 토픽의 2차 레이블은 텍스트랭크 알고리즘에 기반한 추출적 요약을 통하여 도출한다. 둘째, 각 토픽의 추출적 요약 결과를 바탕으로 2차 다양도 지표를 계산함으로써, 최적 토픽 개수 선택의 정확도를 높인다. 셋째, 각 토픽의 레이블을 도출할 때, ChatGPT를 활용함으로써 효율성을 높인다.

<그림 1>은 버토픽 기반 토픽모델링의 성능을 향상시킬 목적으로, 텍스트랭크와 ChatGPT를 융합하기 위한 개념도를 나타내고 있다. <그림 1>에서 1차 다양도는 기존의 다양도 지표를 의미하고 있으며, 2차 다양도는 토픽모델 간 편차가 더 클 수 있는 추가적인 다양도 지표를 의미한다. 마찬가지로, 1차 레이블링은 기존의 토픽표현 방식으로 도출되는 토픽 레이블 생성을 의미하며, 2차 레이블링은텍스트랭크 기반의 텍스트요약에 의한 토픽 레이블 생성을 의미한다.

<그림 2>는 버토픽을 통한 토픽모델링 과정에서 텍스트랭크와 ChatGPT를 융합하는 과정을 알고리즘 형태로 나타내고 있다. 토픽모델링



<그림 1> 버토픽 기반 텍스트랭크 및 ChatGPT 융합 개념



<그림 2> 버토픽 기반 텍스트랭크 및 ChatGPT 융합 알고리즘

대상인 텍스트 문서집합은 버토픽 기법으로 처리되며 그 결과로 토픽모델이 생성된다. 토픽모델은 텍스트 문서집합의 각 문서가 어느 토픽에 소속되는지와 각 토픽의 핵심단어 및 문장을 포함한다. 각 토픽의 핵심단어들을 바탕으로 응집도와 1차 다양도가 계산된다. 1차 다양도는 토픽별 핵심단어들의 임베딩값을 이용하여 토픽 간 모든 비교 쌍에 대한 코사인 거리의 평균값으로 최종 도출된다. 다음으로는 각 토픽의 문서들에 텍스트랭크 알고리즘이 적용되어 토픽별 요약텍스트가 생성된다. 2차 다양도는 토픽별 요약텍스트의 각 문서들에 대하여 토픽 간 모든 비교 쌍의 코사인 거리 평균값으로 최종 계산된다. 토픽모델은 N 번의 실행과정을 통하여 N개의 토픽모델이 생성되며, 응집도와 1차 및 2차 다양도가 가장 큰 토픽모델이 최적의 모델로 선택된다. 최적 토픽모델은 각 토픽별 핵심단어 및 문장들을 포함한다. 각 토픽별 핵심단어들은 버토픽 기반으로 생성되지만, 각 토픽별 핵심문장들은 버토픽 기반으로 생성된 것과 텍스트랭크 기반으로 생성된 것으로 구분된다. 각 토픽별 핵심단어 및 문장들을 대상으로 ChatGPT를 활용하면서 토픽표현을 생성한다. 토픽표현은 각 토픽 당 2개의 레이블을 포함한다. 하나는 1차 레이블로서, 토픽의 핵심 단어 및 문장 기반 레이블이다. 또 다른 하나는 2차 레이블로서, 토픽의 요약텍스트를 기반으로 도출된 레이블이다.

Ⅳ. 사례분석 및 평가

본 논문에서는 버토픽 기반의 토픽모델링 과정에 텍스트랭크 및 ChatGPT를 융합하여 실제 관광후기 텍스트 데이터를 분석하였다. 제주의 주요 관광지인 한라산, 성산일출봉, 우도, 만장굴 등 4곳에 대한 관광후기 데이터를 활용하였다. 2013년부터 2020년 사이에 작성된 관광후기로서, 각 관광지 별 319건 씩 총 1276건의 관광 리뷰 텍스트 데이터이다. <그림 3>은 관광후기 텍스트 데이터의 일부 예를 나타내고 있다.

<그림 3>의 데이터에서 'Category' 열은 '한라산', '성산일출봉', '만장굴', '우도' 등 4개의 범주값을 갖는다. 토픽모델링을 통하여 토픽들을 추출하였을 때, 논리적으로는 4개의 토픽들

Index	Text	Category
0	27개월 둥이들과 방문. 함께 걸어가는데 많이 힘들진 않았습니다. 중국…	성산일출봉
1	성산일출봉은 언제가도 정말 좋은곳이죠.. 넓은 초원에 말도 있고, 정상…	성산일출봉
2	제주에 관광명소이며 세계자연유산인 성산일출봉..정상에서 바라보는 수많은 오름과 아름다운 바다는 말로는 다 표현할수가없다.	성산일출봉
3	산 높이가 약 180m인 산, 푸른 바다 옆에 위치하여 있으며,많은 분화구 중 바다 속에서 수증 폭발한 화산제 이다.	성산일출봉
4	언제가도 일출보기에는 최적의 장소예요 근처에 갈치국집들이 많은데 어딜들어가서먹어도 기본적으로 맛이있는거같아요	성산일출봉
5	제주도 가면 꼭 가봐야하는 성산일출봉!! 진짜 올라갈때는 힘들었지만.…	성산일출봉
6	성산 일출봉에 와서야 중국 관광객이 많다는걸 새삼 느꼈어요. 한국인들…	성산일출봉
7	최고의 풍경을 자랑하는 화산분화구 성산일출봉~ 30분 정도로 정상을 정…	성산일출봉

<그림 3> 관광후기 텍스트 데이터의 일부 예

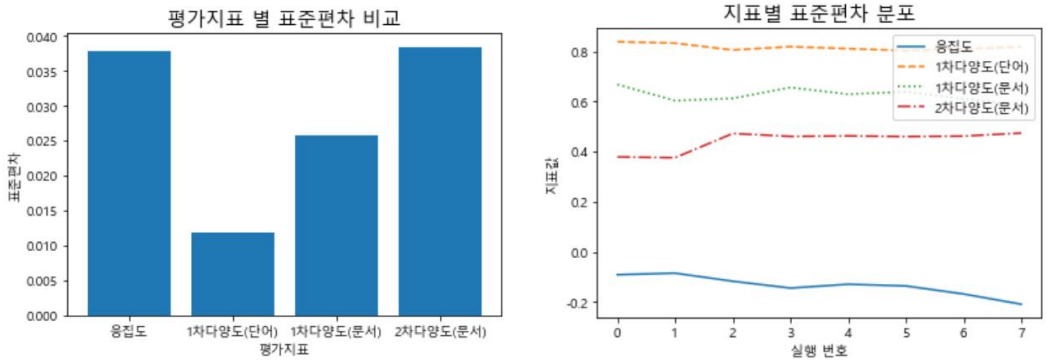
이 도출되면 합리적일 것이다. 본 논문에서는 토픽별로 할당될 핵심단어는 20개, 핵심문서는 10개로 설정하여 8번의 토픽모델링을 실행하였다. 처음 실행에는 추출될 토픽 수를 2로 설정하였으며, 이후 2씩 증가시켜 실행시켰다. 각 실행마다 토픽모델의 평균 응집도(토픽별 응집도의 평균값), 1차 다양도, 2차 다양도를 계산하였으며, 그 결과는 <표 1>과 같다. 각 토픽모델링의 실행마다 생성된 토픽 수는 수동으로 설정된 토픽 수보다 적은 경우가 있다. HDBSCAN 기반 군집화 방식은 이상치로 판단하는 문서들을 기타 군집으로 분류하기 때문에 수동 설정된 군집 개수보다 적은 수의 군집들을 생성하기도 한다. 1차 다양도는 버토픽으로 생성된 토픽모델의 각 토픽에 할당된 핵심단어와 핵심문서를 사용하였다. 1차 다양도는 핵심단어 기반 1차 다양도와 핵심문서 기반 1차 다양도로 구분하여 계산하였다. 2차 다양도는 각 토픽에 할당된 문서들을 대상으로 텍스트랭크 기반의 요약텍스트를 생성하여 계산하였다.

응집도의 범위는 -1과 1 사이의 값으로서, 1에 가까울수록 토픽 내 단어 및 문장들이 서로 연관되어 있음을 나타낸다. 응집도는 음수의 값을 가질 수도 있으며 1에 가까울수록 토픽모델링이 양호한 것으로 판단한다. 다양도는 0 ~ 1 사이의 범위 값을 가지며 1에 가까울수록 토픽 간 중복성이 적음을 의미한다. 즉, 다양도가 1에 가까울수록 토픽모델의 결과는 양호한 것으로 판단할 수 있다. <표 1>에서는 '실행번호0'에 대응하는 토픽모델의 응집도가 다른 실행의 토픽모델링 결과에 비하여 상대적으로 1에 더 가깝지만, 다양도는 그렇지 않다. '실행번호3'에 대응하는 토픽모델의 경우, 다양도는 양호하지만 응집도는 그렇지 못하다. 그러나, 응집도와 다양도를 균형적으로 고려하였을 때, '실행번호0'과 '실행번호3'의 토픽모델은 다른 토픽모델에 비하여 상대적으로 양호하다고 볼 수 있다.

응집도 및 다양도와 같은 모델 평가지표 값에 의하여 최적 토픽모델을 선택하고자 할 때, 지표값들의 편차가 클수록 효과적인 의사결정을 할 수 있다. <그림 4>는 <표 1>의 응집도, 1차 다양도(단어기반), 2차 다양도(문서 기반) 값들의 분포에 대한 표준편차를 나타내고 있다. 응집도와 2차 다양도의 표준편차는 1차 다양도의 표준편차보다 크게 나타나고 있다. 최적 토픽모델을 선택할 때 버토픽 기반의 1차 다양도보다 텍스트랭크 기반의 2차 다양도가 최적 토픽 수를 선택하는 기준으로는 더 효과적일 수 있음을 알 수 있다.

<표 1> 토픽모델링의 성능을 평가하기 위한 지표값

실행번호	토픽수	평균 응집도	1차 다양도 (단어기반)	1차 다양도 (문서기반)	2차 다양도	다양도 평균
0	**2**	**-0.037**	**0.794**	**0.534**	**0.280**	**0.536**
1	**4**	**-0.175**	**0.809**	**0.596**	**0.440**	**0.615**
2	2	-0.091	0.839	0.668	0.379	0.609
3	**8**	**-0.138**	**0.815**	**0.648**	**0.457**	**0.640**
4	2	-0.085	0.833	0.604	0.376	0.604
5	2	-0.085	0.833	0.604	0.376	0.604
6	14	-0.212	0.811	0.567	0.469	0.615
7	14	-0.184	0.806	0.593	0.469	0.622



<그림 4> 평가지표 별 표준편차 분포 및 크기 비교

본 논문에서는 최적 토픽모델을 선택할 때, 응집도와 다양도 평균(1차 다양도와 2차 다양도의 평균)을 활용하였다. 응집도가 가장 좋은 토픽모델은 '실행번호0'에 대응하는 모델이고, 다양도가 가장 좋은 모델은 '실행번호3'에 대응하는 토픽모델이다. 따라서, '실행번호0'와 '실행번호3'에 대응하는 토픽모델을 선택하여 ChatGPT 기반의 토픽표현 과정을 진행하였다. 실험용 텍스트 데이터의 텍스트 범주는 4개(한라산, 성산일출봉, 만장굴, 우도)이기 때문에 4개의 토픽들을 포함하는 토픽모델('실행번호1'에 대응)도 토픽표현 결과의 정확도 비교 대상으로 선정하였다. 따라서, '실행번호0', '실행번호1', '실행번호3'의 토픽모델에 대하여 ChatGPT를 활용하면서 토픽표현을 도출하기 위한 과정을 진행하였다.

토픽표현을 위한 ChatGPT 프롬프팅은 제로샷 방식(박상언 외, 2023)으로 진행하였다. <그림 5> ~ <그림 7>은 토픽표현을 위한 ChatGPT 프롬프팅의 예를 나타내고 있다. <그림 5>는 토픽에서 추출된 20개 핵심단어 기반의 1차 토픽 레이블링을 위하여 제로샷 방식으로 ChatGPT에 질문한 내용과 답변을 나타내고 있다.

<그림 6>은 버토픽에 의한 토픽모델의 특정 토픽에 할당된 10개의 핵심 문서를 이용하여 1차 토픽 레이블을 도출하는 과정을 나타내고 있다. 제로샷 방식으로 ChatGPT 에 질문한 내용과 답변을 보여주고 있다.

<그림 7>은 텍스트랭크로 요약한 요약텍스트 기반으로 2차 토픽 레이블을 도출하는 과정을 나타내고 있다. 마찬가지로, 제로샷 방식의 ChatGPT 프롬프팅 및 그 결과를 보여주고 있다.

<질문> 제주의 관광 도메인을 고려하면서, 아래의 단어들을 50문자 이내의 한글문장으로 요약 해 주세요.
['용암', '튜브', '바닥', '내부', '만장굴', '설명', '형성', '조명', '하나', '길이', '바위', '터널', '이곳', '규모', '거리', '전체', '화산', '온도', '거대', '종유석']

<답변> 만장굴: 용암 튜브 형성, 길이와 규모 거대, 조명 아래 바위와 종유석 관찰.

<그림 5> 단어 기반 1차 레이블링을 위한 ChatGPT 프롬프트 예

<질문> 제주의 관광 도메인을 고려하면서, 다음의 텍스트를 50문자 이내로 요약해 주세요.

['우도! 특별한 수식어는 필요없습니다 보통 아침에 일찍 들어와 저녁에 나가야 일정이 좋다고들 합니다 그렇지만.... 제주를 여행할때 꽉 채우는것보다 잠시 머물러 보세요... 알찬시간여행은 효율적임에 틀림없지만 블러그에서는 자주 말하지않는 정적이 보입니다 막배로 들어와서 첫배로 나가는것도 아낌없이 효율적인 하루입니다 술은 언제든 먹을수 있으니 절주하고 사진은 언제든 남아있을테니 눈으로 보고 생각을 멈춰주세요....',...중략... '깨끗한 바다와 다정하게 생긴 모습이 참 친근한 느낌을 주는 섬으로 제주도와 얽힌 설화도 매우 재밌다. 제주도에서 떨어져나온 화산섬이라서 그런 듯 하다. 검은 해변, 검멜레 해수욕장은 많은 드라마에서 본 그대로 파란 바다와 어우러지는 절경이다. 다만, 배에서 내리자마자 겪게되는 스쿠터 호객행위에 넘어가 스쿠터를 이용했는데 스쿠터 대여 비용도 바가지이고(2대 65,000원) 너무 성수기인데다가 폭염으로 날씨도 너무 덥고,']

<답변> 우도 여행: 여유롭게 즐기세요. 막배-첫배 일정 추천, 스쿠터 대여 바가지 주의 필요.

<그림 6> 문서 기반 1차 레이블링을 위한 ChatGPT 프롬프트 예

<질문> 제주의 관광 도메인을 고려하면서, 다음의 텍스트를 50문자 이내로 요약해 주세요.

['갑작스런 출장겸 가게 되었는데 날씨가 큰 영향을 끼치는듯~ 전날 비가 와서 등산 당일날 너무 깨끗한 하늘에 바다까지 보이고 백록담도 선명하게 잘 보고왓당 가는길이 길지만 힘든 코스는 아니지만 길어서 힘든것~~', '일출봉은 아름다운 곳이다 계단을 오르는게 다소 가파를수있으나 올라가서 제주를 한눈에 내려다보는것은 매우 매력적이다 보고 또 봐도 질리지 않고 사계절내내보아도 다 다른매력이있다', '제주에서 빠질수 없는곳 한라산입니다 ,...중략... 하산하는 동선은 올라갈때보다 수월하게 만들어져있더군요', '일출봉으로 오르는 무수한 계단은 고통스럽지만 참고 끝까지 올라가면 그만한 보람이 있는 곳입니다. 아름다운 경치와 시원한 바람 때문에 제주도 갈 때마다 방문합니다.', '어머니 아버지와 함께 간만에 등반 했어요 너무너무 좋아요~~신라스테이에서 하루밤자고 한라산고고 제주도 살지만 자주못가는데 제주내에서도 공기가 너무좋았어요~~', '정상까지 올라간 적은 없지만 날씨 풀리면 한번 도전해보고싶어요 특히 눈덮인 한라산은 정말 최고의 뷰에요~~', '성산일출봉 여러번 가봤지만 제주도 가게 되면 꼭 들리는곳 아이가 있어 정상까지 다녀오진 못했지만 주변 풍경이랑 바다가 너무 좋았어요']

<답변> 한라산과 성산일출봉: 등반 힘들지만 아름다운 경치와 전망 추천.

<그림 7> 2차 레이블링을 위한 ChatGPT 프롬프트 예

<표 2>는 '실행번호0', '실행번호1', '실행번호3'에 대응하는 각 토픽모델에 대하여 ChatGPT를 활용한 토픽표현 결과를 나타내고 있다. 토픽표현은 1차 레이블과 2차 레이블로 구성된다. 1차 레이블은 버토픽으로 생성된 토픽모델의 핵심단어들과 문장들을 바탕으로 도출되었다. 2차 레이블은 텍스트랭크로 생성된 요약텍스트들을 바탕으로 도출되었다. ChatGPT는 많은 정보와 문맥을 포함하는 질문일수록 더 정확한 답변을 생성하기 때문에, 핵심단어보다는 핵심문장을 활용하여 도출된 토픽표현이 정확한 레이블이라 할 수 있다. 1차 레이블은 핵심단어에 의한 경우와 핵심 문서에 의한 경우가 거의 동일하였다. 이것은 버토픽에서 각 토픽의 핵심문장을 도출할 때, c-TFIDF 기반의 중요단어들을 포함하는 문장들이 핵심문장으로 선택되기 때문이다.

<표2>에서 2차 레이블의 결과는 1차 레이블의 결과와 차이가 있음을 알 수 있다. 이것은 버토픽과 텍스트랭크 간에 핵심문장을 추출하는 접근방식이 다르기 때문이다. 버토픽에서는 다른 토픽과의 차별성을 높이는 문장을 중요한

것으로 간주하는 반면, 텍스트랭크는 다른 토픽과 독립적으로 토픽 내 문장들 중에서 서로 연관성이 높은 문장에 높은 가중치를 부여하기 때문이다.

<표 2> ChatGPT 기반 토픽 표현 결과

실행 번호	토픽 번호	1차 레이블		2차 레이블
		단어기반	문서 기반	
0	1	제주 관광: 야경, 점포, 호객, 바가지 조심하며 즐기는 가족 여행.	우도 여행: 여유롭게 즐기세요. 막배-첫배 일정 추천, 스쿠터 대여 바가지 주의 필요.	한라산과 성산일출봉: 등반 힘들지만 아름다운 경치와 전망 추천.
	2	제주 관광: 외국인들도 즐기는 거대 볼거리와 드라이브 코스.	제주도 만장굴: 시원한 용암 터널, 편한 운동화 착용 추천, 택시 이용 편리.	우도: 아름다운 바다와 야경, 스쿠터 바가지 주의, 해물 맛집 추천.
1	1	제주 여행: 우도, 성산일출봉, 만장굴 등 최고의 관광 코스 추천.	성산일출봉: 일출 감상 필수, 편의시설 비싸니 물과 간식 준비 필요.	한라산과 성산일출봉: 힘들지만 아름다운 경치와 상쾌한 바람 추천.
	2	제주도 관광: 한라산 등산, 성산일출봉 경치, 우도 풍경 즐기기.	제주 관광: 한라산과 성산일출봉 필수, 아름다운 경관과 시원한 바람.	우도: 스쿠터, 전기 자전거로 투어하며 해변과 특산물 즐기기 추천.
	3	제주 우도 여행: 자전거, 스쿠터 투어와 땅콩 아이스크림 즐기기.	우도: 땅콩아이스크림 맛있고 해변 아름다움, 전기자전거 추천, 바람 강함.	제주 동굴 관광: 견고한 신발과 긴팔 준비, 시원한 내부 체험.
	4	제주 관광: 안전 위해 운동화, 손전등, 긴팔, 비옷 휴대 권장.	제주 관광: 견인력 좋은 신발, 손전등 필수, 재킷 불필요.	제주 관광: 우도 숙박 추천, 만장굴 시원함, 성산일출봉 경치, 한라산 등반 필수.
3	1	우도와 만장굴: 버스, 택시, 스쿠터, 자전거로 투어하며 일출과 식당 즐기기.	우도: 전기 스쿠터 위험, 일주버스 불편, 먹거리 비싸고 특색 없음.	제주 관광: 한라산과 성산일출봉 등반, 아름다운 경치와 시원한 바람 추천.
	2	한라산 등반: 성판악, 영실, 관음사 경로로 일출과 백록담 경치 즐기기.	한라산: 멋진 경치와 시원한 바람, 겨울엔 특히 아름다움.	제주 만장굴: 견고한 신발과 재킷 필수, 용암동굴의 경이로움 체험.
	3	만장굴: 용암 튜브 형성, 길이와 규모 거대, 조명 아래 바위와 종유석 관찰.	만장굴: 용암 튜브와 다양한 형성물, 1km 길이의 흥미로운 동굴.	우도: 스쿠터나 자전거로 섬 투어하며 아름다운 경치와 특산물 즐기기.
	4	우도: 자전거와 스쿠터 대여, 땅콩 아이스크림, 하루 투어 추천.	우도: 땅콩아이스크림 맛있고, 해변 아름다움, 자전거나 스쿠터 추천.	성산일출봉: 탁 트인 경치와 아름다운 사진 명소, 오르기 쉬움.
	5	우도와 한라산: 일출 경치, 스쿠터 투어, 매표소에서 신청 강추.	우도: 아름다운 자연, 친절한 주민, 비수기 방문 추천.	우도: 버스 투어 추천, 아름다운 풍경과 다양한 먹거리 경험 필수.
	6	우도 투어: 버스, 택시 이용해 페리로 이동, 하루 충분히 즐기기.	제주 동굴과 한라산: 버스로 접근 가능, 영어 표지판과 편리한 교통.	성산일출봉: 중국인 관광객 많음, 멋진 경치와 분화구, 주차 어려움.
	7	성산일출봉: 중국인 많고 주차 어려움, 경치 아름답고 문화유산으로 필수 방문.	성산일출봉: 전망 좋지만 중국인 관광객 많음, 필수 방문 코스.	제주 동굴 관광: 시원한 내부, 견고한 신발과 긴팔 준비 필수.
	8	제주 동굴 탐험: 견고한 신발과 긴팔 재킷 착용, 손전등 휴대 필수.	제주 동굴: 시원한 내부, 운동화와 비옷 필수, 박쥐 관찰 가능.	우도: 아름다운 해변, 스쿠터와 자전거 투어 추천, 땅콩 아이스크림 필수.

<표 3> 토픽별 범주 분포

실행번호	토픽번호	한라산	우도	만장굴	성산일출봉	문서 수 합계
0	1	284	270	299	271	1124
	2	35	49	20	48	152
						1276
1	1	274	71	255	250	850
	2	27	231	34	39	331
	3	0	0	13	0	13
	4	18	17	17	30	82
						1276
3	1	249	23	12	193	477
	2	2	1	210	2	215
	3	3	165	2	5	175
	4	25	53	31	59	168
	5	14	33	25	2	74
	6	1	3	0	27	31
	7	0	0	13	0	13
	8	25	41	26	31	123
						1276

<표3>은 버토픽이 도출한 각 토픽에 할당된 문서들의 원래 범주(Category)를 나타내고 있다. '실행번호0'에 대응하는 토픽모델의 '토픽1'에 할당된 문서는 '한라산' 범주의 문서가 284건, '우도' 범주의 문서가 270건, '만장굴' 299건 등을 의미한다. <표3>에서 제시된 실제 텍스트의 범주별 분포와 1차 및 2차 레이블링에 의한 토픽표현 결과를 비교하여 보면, 토픽 레이블의 정확성을 판단할 수 있다. 여기서 주목할 사항은 텍스트랭크 기반의 요약텍스트로 도출한 2차 레이블이 더 정확한 토픽표현이라는 것이다. <표3>에서 '실행번호1'의 '토픽3'의 경우, 실제 '만장굴' 범주였던 문서들이 '토픽3'에 할당되었다. <표2>에서 볼 수 있는 바와 같이, 1차 레이블은 '우도' 관련 내용으로 설정된 반면, 2차 레이블은 '만장굴' 관련 내용으로 설정되었다. 텍스트랭크 기반의 2차 레이블이 더 정확하게 토픽표현을 제시하고 있음을 알 수 있다. <표3>의 '실행번호3'에서 '토픽1'의 경우도 마찬가지이다. 대부분 '한라산' 범주와 '성산일출봉' 범주에 속하는 문서들이 이 토픽에 할당되었는데, <표2>에서 '실행번호3'에서 '토픽1'의 1차 레이블은 '우도' 관련 내용으로 설정되었으며, 2차 레이블은 '한라산'과 '성산일출봉' 관련 내용으로 설정되었다. 2차 레이블이 1차 레이블보다 더 정확한 토픽표현인 것을 확인할 수 있다. '실행번호3'에 대응하는 토픽들 2, 3, 6, 7의 경우도 2차 레이블에 의하여 정확한 토픽표현이 제시되고 있음을 알 수 있다. 이러한 결과는 토픽표현에서 텍스트랭크가 버토픽을 보완하여 좋은 성능의 토픽모델을 도출할 수 있음을 시사한다. 버토픽을 통하여 토픽모델을 생성하되, 도출된 각 토픽에 대한 토픽표현은 텍스트랭크를 활용하면 더욱 좋은 토픽모델을 생성할 수 있다.

Ⅴ. 결 론

디지털기기의 대중화 및 SNS의 활성화로 각 분야에서 대량의 텍스트 데이터가 생산되고 있다. 텍스트 데이터의 분석을 통하여 의미있는 인사이트를 도출할 수 있다. 토픽모델링은 대량의 텍스트 데이터로부터 주요 내용을 추출하는 군집화 분석 방법의 일종이다. 버토픽은 최근에 제안된 토픽모델링 기법으로서, 워드임베딩 기법을 적용하여 그 성능이 비교적 우수하다고 알려져 있다. 그러나, 버토픽의 토픽표현 기능은 c-TFIDF의 전역적 처리방식 때문에 토픽내용을 정확하게 반영하는데 한계가 있음을 확인할 수 있었다.

본 논문에서는 버토픽 기반의 토픽모델링 과정에서 토픽모델의 성능을 향상시키기 위하여 텍스트랭크 기법을 융합하는 방법을 제안하였다. 버토픽 기반으로 추출된 각 토픽의 문서집합을 텍스트랭크에 적용하여 토픽별 요약텍스트를 추출하고 이를 바탕으로 2차 다양도와 2차 레이블을 도출함으로써, 최적 토픽모델의 선택과 성능 개선이 가능할 수 있도록 하였다.

본 논문에서는 제안한 토픽모델링 방법을 활용하여 관광후기 텍스트 데이터를 분석하고 도출된 토픽모델의 성능을 평가하였다. 그 결과, 본 논문에서 제안한 2차 다양도는 최적 토픽수를 계산하는데 효과적으로 활용될 수 있음을 확인하였다. 특히, 토픽표현을 위한 2차 레이블은 각 토픽의 주요 내용을 정확히 표현할 수 있음을 확인하였다. 관광후기 분석사례에서 175건의 문서 중 165건이 ‘우도’에 대한 리뷰였는데도 불구하고 버토픽의 토픽표현은 ‘만장굴’로 레이블링하였지만, 텍스트랭크 기반의 토픽표현에서는 ‘우도’로 정확히 레이블링하였다. 버토픽 기반의 1차 레이블링 결과는 부정확한 토픽표현이 40%(12건의 토픽 중 6건)에 달했지만, 텍스트랭크 기반의 2차 레이블링을 통하여 왜곡된 토픽표현을 바로 잡을 수 있었다. 결과적으로 버토픽 기반 토픽모델링 과정에서 토픽표현 부분은 텍스트랭크를 적용하는 것이 바람직함을 알 수 있었다.

본 논문에서는 텍스트랭크 기반의 요약텍스트를 생성할 때 임베딩 기술을 적용하였다. 버토픽과 텍스트랭크의 융합을 통한 토픽모델링 방법은 토픽 표현 부분에서는 버토픽을 능가하는 성능을 보여주었다.

본 논문에서 현존하는 토픽모델링 방법 중 가장 우수하다고 알려진 버토픽 기법의 한계점을 발견하고 이를 극복하기 위한 방법으로 버토픽과 텍스트랭크의 융합방안을 제안하였다. 본 논문에서 제안한 방안은 기존의 분석방법들을 적절하게 융합함으로써 문제 해결의 새로운 돌파구를 마련하였다는 측면에서 기술 개선 연구로서의 학문적 시사점이 있다. 본 논문에서 제안한 토픽모델링 방법이 다양한 도메인의 텍스트분석에 실무적으로 활용될 수 있기를 기대한다.

본 논문에서 제안한 토픽모델링 방법은 특정 사례에 적용하였을 때는 효과가 있는 것으로 판명되었지만, 일반화된 토픽모델링 기법으로 정착되기에는 한계가 있다. 추후, 다양한 도메인에서의 추가적인 검증이 필요하며, 많은 연구자들의 응용 연구가 지속될 수 있기를 기대한다.

참고문헌

김근형, “귀납적 사회과학연구 방법론을 위한 토픽모델링의 확장 및 사례분석,”, 정보시스템연구, 제31권, 제4호, 2022, pp.25-45.

김효곤, 유동희, “BERT를 활용한 미국 기업 공시에 대한 감성분석 및 시각화,”, 정보시스템연구, 제31권, 제3호, 2022, pp.67-87.

박상언, 강주영, “ChatGPT 및 거대언어모델의 추론 능력 향상을 위한 프롬프트 엔지니어링 방법론 및 연구현황 분석,” 지능정보연구, 제29권, 제4호, 2023, pp.287-306.

Cédric Févotte and Jérôme Idie, “Algorithms for nonnegative matrix factorization with the - divergence,” *Neural computation*, Vol.23, No.2, 2011, pp.2421-2456.

Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim, “On the surprising behavior of distance metrics in high dimensional space,” *In International conference on database theory*, 2001, pp.420-434.

David M Blei, Andrew Y Ng, and Michael I Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research 3*, 2003, pp.993-1022.

Devlin Jacob, Chang Ming-Wei, Lee Kenton and Toutanova, Kristina. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *arXiv:1810.04805v2 [cs.CL]*, 2018, pp.1-16.

Gerlof Bouma, “Normalized (pointwise) mutual information in collocation extraction,” *Proceedings of GSCL 30*, 2009, pp.31-40.

L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” *arXiv:1802.03426 e-prints*, 2018, pp.1-63.

L. McInnes and John Healy, “Accelerated hierarchical density based clustering.” *Data Mining Workshops (ICDMW) In IEEE International Conference*, 2017, pp.33-42.

Maarten Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” *arXiv:2203.05794 [cs.CL]*, 2023, pp.1-10.

Nils Reimers and Iryna Gurevych, Sentencebert, “Sentence embeddings using siamese bertnetworks,” *In Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2019, pp.3982-3991.

Rada Mihalcea and Paul Tarau, “TextRank: Bringing Order into Text,” *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004, pp.404-411.

Silvia Terragni, Elisabetta Fersini and Enza Messina, “Word Embedding-Based

Topic Similarity Measures," *Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems*, 2021, pp.33-45.

김 근 형 (Kim, Keun Hyung)

서강대학교 컴퓨터공학과에서 학사, 석사, 박사학위를 취득하였다. 현재 제주대학교 경영정보학과 교수로 재직하고 있으며, 주요 관심분야는 데이터베이스, 데이터마이닝, 빅데이터분석 등이다.

강 재 정 (Kang, Jae Jung)

제주대학교 경영학사와 한국외국어대학교 경영학석사, 고려대학교 경영학 박사학위를 취득하였다. 현재 제주대학교 경영학과 교수로 재직하고 있으며, 주요 관심분야는 경영정보시스템, 시스템다아내믹스 등이다.

<Abstract>

Improvement of topic modeling and case analysis through convergence of Bertopic and TextRank

Kim, Keun Hyung · Kang Jae Jung

Purpose

The purpose of this paper is to develop a method to improve topic representation by incorporating the TextRank technique in Bertopic-based topic modeling and additional indicators for determining the optimal number of topics.

Design/methodology/approach

In this paper, we propose a method to extract important documents from documents assigned to each topic of a topic model using the TextRank technique, and to calculate secondary diversity and generate topic representations based on the results. First, we integrate the TextRank algorithm into the Bertopic-based topic modeling process to set local secondary labels for each topic. The secondary labels of each topic are derived through extractive summarization based on the TextRank algorithm. Second, we improve the accuracy of selecting the optimal number of topics by calculating the secondary diversity index based on the extractive summary results of each topic. Third, we improve the efficiency by utilizing ChatGPT when deriving the labels of each topic.

Findings

As a result of performing case analysis and analysis evaluation using the proposed method, it was confirmed that topic representation based on TextRank results generated more accurate topic labels and that the secondary diversity index was a more effective index for determining the optimal number of topics.

Keyword: Text data, Topic modeling, Bertopic, Textrank, Topic expression, Diversity, Coherence, Convergence

* 이 논문은 2024년 7월 30일 접수, 2024년 8월 27일 1차 심사, 2024년 8월 31일 게재 확정되었습니다.