디지털산업정보학회 논문지 제16권 제3호-2020년 9월
http://dx.doi.org/10.17662/ksdim.2020.16.3.025

토픽 레이블링을 위한 토픽 키워드 산출 방법

김 은 회* · 서 유 화**

A Method of Calculating Topic Keywords for Topic Labeling

Kim Eunhoe · Suh Yuhwa

〈Abstract〉

Topics calculated using LDA topic modeling have to be labeled separately. When labeling a topic, we look at the words that represent the topic, and label the topic. Therefore, it is important to first make a good set of words that represent the topic. This paper proposes a method of calculating a set of words representing a topic using TextRank, which extracts the keywords of a document. The proposed method uses Relevance to select words related to the topic with discrimination. It extracts topic keywords using the TextRank algorithm and connects keywords with a high frequency of simultaneous occurrence to express the topic with a higher coverage.

Key Words : LDA, Topic Model, Topic Label, Relevance, TextRank

Ⅰ. 서론

토픽 모델링은 구조화되지 않은 방대한 문서집단에서 자동으로 주제(topic)를 찾아내는 알고리즘이다. LDA(Latent Dirichlet Allocation)는 토픽 모델링 기법 중 가장 많이 사용되는 모델로서 특히, 뉴스, 블로그, 소셜미디어 등에서 시시각각 쏟아지는 방대한 양의 데이터를 분석하여 의미 있는 정보를 찾아내기 위해 주로 사용된다[1, 2]. LDA 토픽 모델링은 잠재 디리클레(Latent Dirichlet) 확률 모델에 기반을 두며, 토픽 별 단어들의 분포와 문서 별 토픽들의 분포를 모두 추정할 수 있다[3].

그런데 LDA 토픽 모델링을 사용하여 산출한 토픽들은 토픽의 레이블이 자동으로 붙여지지 않고, 토픽을 번호로서 구분하고 있다. 토픽의 레이블은 사람이 직접 토픽을 대표하는 단어들을 검토하여 붙여야 한다. 그러나 사람이 직접 토픽의 레이블을 붙이는 작업은 해당 도메인에 대한 전문적인 지식이 있어야 하고, 시간이 많이 소모될 뿐만 아니라, 매우 주관적인 토픽 레이블을 선정할 수 있다는 단점들이 있다. 따라서 토픽의 레이블을 자동으로 생성하기 위한 연구가 진행되어 왔다. 그런데 자동으로 토픽의 레이블을 생성하는 연구들 또한 토픽을 대표하는 단어들을 먼저 뽑고, 이를 기반으로 토픽의 레이블을 붙인다. 따라서 토픽에 적합한 레이블을 붙이기 위해서는 먼저 토픽을 대표하는 단어들의 집합을 잘 만드는 것이 중요하다.

토픽을 대표하는 단어들의 집합을 만들기 위해

* 서일대학교 소프트웨어공학과 조교수(제1저자)
** 숭실대학교 베어드교양대학 조교수(교신저자)

LDA 토픽 모델이 추정한 토픽 별 단어의 확률분포, PMI, 유사도 측정, Word2vec, doc2vec 등 다양한 방법들이 사용되어 왔고, Wikipedia, Google, DBpedia 등과 같은 외부 데이터 소스를 활용하는 방법들이 연구되어 왔다[4-8]. 그러나 전문가가 토픽의 대표 단어들을 선별하는 것 이상의 더 좋은 방법은 아직까지 없기 때문에 연구가 계속 필요한 분야이다.

본 논문은 토픽 레이블링을 위해 TextRank를 사용하여 변별성과 포괄성을 갖춘 토픽을 대표하는 키워드들을 산출하는 방법을 제안한다. TextRank는 문서의 키워드 또는 요약문을 추출할 때 사용하는 그래프 기반의 랭킹 모델이다[9]. 먼저 변별성(discrimination)이 있는 차별화된(discriminative) 토픽 관련 단어들을 선정하기 위해 Relevance[8]를 사용한다. Relevance를 사용해 선정한 토픽 관련 상위 단어들을 기반으로 TextRank 알고리즘을 이용해 토픽 키워드를 추출하고, 동시 출현 빈도가 높은 키워드들을 연결하여 토픽을 좀 더 포괄적으로 표현할 수 있는 문구를 만들어 토픽을 대표하는 단어 집합을 생성한다.

본 논문은 2장에서 기존 선행 연구를 살펴보고, TextRank 알고리즘을 설명한다. 3장에서는 TextRank를 사용한 토픽 키워드 산출 방법을 설명하고, 4장에서는 실험 결과를 논하고 평가한다. 5장에서는 결론과 향후 연구를 제시한다.

II. 관련연구

2.1 선행 연구

LDA 토픽 모델링을 통해 산출된 토픽에 적절한 토픽 레이블을 선정하기 위한 연구들은 모두 토픽 모델링을 통하여 추출한 토픽 별 상위 단어들(Top-ranking Terms)을 기반으로 한다.

[4]는 First-Order-Relevance를 제안하였다. 토픽을 구성하는 상위 단어들을 포함하는 문서들을 분석하여 단어의 확률 분포를 계산하고, 정보검색에서 많이 사용하는 MMR(Maximal Marginal Relevance)를 사용하여 포괄적인 토픽 레이블링 방법을 제안하였다. 또한 차별화된 토픽 레이블을 선정하기 위하여 해당 토픽에 대한 관련성이 높고 다른 토픽에 대해서는 관련성이 낮은 토픽 관련 단어들을 선정하는 방법을 제시하였다. [5]는 토픽 별 상위 10개의 단어들을 대상으로 대표성이 있는 단어를 찾는 방법을 제안하였다. 문서에 동시 출현하는 PMI(Pointwise Mutual Information) 값을 구하여 평균이 제일 높은 단어를 토픽의 대표 단어로 선정하거나, WordNet을 사용하여 상위 단어들 사이의 유사도를 측정하여 가장 평균이 높은 단어를 토픽의 대표 단어로 선정하는 등의 방법을 제안하였다. [4]와 [5] 등과 같은 연구들은 적합한 토픽의 레이블은 LDA 토픽 모델링에 사용된 내부 문서들에 포함되어 있다는 생각을 가지고 내부 문서들에서 토픽의 대표 단어들을 찾는 방법이다.

반면, 토픽의 레이블링을 위해 외부 데이터 소스(source)를 참조하는 연구들이 있다. [6]는 토픽의 상위 단어에 없는 적합한 단어를 찾기 위해 Wikipedia 제목 검색, Google 검색을 사용하여 토픽 대표 단어를 찾는 방법을 제안하였다. [7]는 DBpedia 제목을 활용하여 토픽 그래프를 생성하고 그래프의 가장 중심이 되는 노드를 해당 토픽의 레이블로 선정하는 방법을 제안하였다. [8]은 [6]에서 제안한 방법을 확장하여 자연어 처리기술인 word2vec, doc2vec 모델을 만들어 적합한 토픽 대표 단어들을 추출하는 연구를 하였다.

자동으로 토픽의 레이블을 선정하기 위한 많은 연구들이 이루어졌지만, 아직까지 해당 도메인의 전문가가 토픽의 대표 단어들을 선별하는 것 이상의 좋은 레이블링 방법은 없다. 또한, 해당 토픽의 레이블이 적합한지를 평가하는 것 또한 전문가의 정성적인 평가에 의존할 수밖에 없으므로 토픽의 적합한 레이블링을 위해 토픽의 대표 단어 집합을 찾는 다양한 많

은 연구가 더 필요하다.

본 논문에서 제안하는 방법은 LDA 토픽 모델링을 통해 산출한 토픽의 상위 단어들을 기반으로 하며, [4]와 [5] 같이 토픽의 대표 단어들을 LDA 토픽 모델링에 사용된 내부문서에서 찾는 방법이다. TextRank 알고리즘을 사용하여 적합한 토픽 키워드들을 제공하며 분별력이 있는 차별화된 포괄성 있는 토픽 레이블을 선정할 수 있게 한다.

2.2 TextRank

TextRank란 구글 검색 알고리즘으로 알려진 PageRank를 텍스트에 적용한 그래프 기반의 랭킹 모델이다. PageRank는 중요도가 높은 웹 페이지는 다른 웹 페이지들로부터 링크를 받는다는 점에서 착안된 알고리즘이다. PageRank 알고리즘은 페이지를 정점(Node)로, 페이지와 페이지를 연결하는 링크를 간선(Edge)으로 하는 그래프를 만들고, 정점들 사이의 연결상태를 사용하여 중요한 정점을 계산한다[9].

TextRank는 PageRank를 텍스트에 적용한 알고리즘이다. 텍스트에서 정점이 될 만한 텍스트 단위(unit)을 추출하여 정점으로 만들고, 이 텍스트 단위끼리의 연결상태를 간선으로 하여 그래프를 만들어서 텍스트 단위의 순위를 계산하는 알고리즘이다. 텍스트 단위가 단어일 때, 문장내에서 단어들이 동시에 출현하는 상태를 간선으로 표현하고, 한 문장내의 동시 출현 빈도를 계산하여 단어의 TextRank 값을 계산할 수 있다. 한 문장내에서 동시 출현하는 빈도가 높을수록 TextRank의 값이 더 커지므로 중요도가 높은 단어를 알아 낼 수 있다. 그러므로 Text Rank 알고리즘은 보통 텍스트 단위를 단어로 하여 중요한 키워드 추출에 사용한다.

< 수식 1 > $TR(V_i) = (1-d) + d * \sum_{V_j \in In(V_j)} \frac{Wji}{\sum_{V_k \in Out(V_j)} W_{jk}} TR(V_j)$

<수식 1>는 TextRank 계산식이다. TR(Vi)는 단어 Vi에 대한 TextRank 값이다. Wi,j는 단어 Vi와 Vj 사이의 가중치를 나타낸다. In(Vi)는 단어 Vi와 연결되어 있는 모든 단어들의 집합이고, Out(Vj)는 단어 Vj와 연결되어 있는 모든 단어들의 집합이다. d는 제동계수(damping factor)이다. 제동 계수란 PageRank에서 웹 서핑을 하는 사람이 현재 페이지에 만족을 못하고 다른 페이지로 가는 링크를 클릭할 확률을 나타낸다. 통상적으로 d는 0.85가 사용되며, TextRank에서도 0.85를 사용한다. 단어를 표현하는 모든 정점들에 대하여 TR 값을 계산하면, 각 단어들 사이의 중요도를 알 수 있고 TR 값이 높은 단어를 키워드로 추출한다 [9].

TextRank는 텍스트 단위를 문장으로 하여, 문장을 정점으로 만들고, 문장들 사이의 유사도를 간선으로 표현하는 그래프를 만들어 TextRank 값을 계산하면 각 문장의 중요도를 알아낼 수 있다. 이렇게 계산된 TextRank 값이 높은 문장들을 추출하면 텍스트 요약(summary)문을 추출할 수 있다.

본 논문에서는 LDA 토픽 모델링과 Relevance를 통해 추출한 각 토픽의 상위 단어들을 포함하고 있는 문장들을 추출하고, 이러한 문장들에서 토픽을 대표할 수 있는 핵심 키워드를 찾기 위해 TextRank 알고리즘을 사용한다.

Ⅲ. TextRank를 이용한 토픽 키워드 산출

3.1 토픽의 키워드를 산출할 때 고려해야 할 사항

토픽의 레이블은 이해하기 쉽고, 의미적으로 토픽에 적절하고, 변별성(discrimination)과 포괄성이 있어야 한다[4]. 이들 4가지 요소들 중 가장 이슈가 되는

것은 변별성과 포괄성이다. 변별성이 있는 레이블은 산출된 토픽들을 서로 잘 분별할 수 있는 차별화된 레이블을 말한다. 포괄성이 있는 레이블은 한 토픽을 의미적으로 잘 포괄할 수 있는 표현력을 가진 레이블이다[4, 10]. 토픽 레이블은 토픽을 대표하는 단어들의 집합으로부터 산출되므로, 토픽을 대표하는 단어 집합 곧, 토픽의 키워드들을 산출할 때도 변별성과 포괄성을 반드시 고려해야 한다.

LDA 토픽 모델링을 통해 산출된 각 토픽의 상위 단어들(top-ranking terms)은 해당 토픽에 대한 단어의 출현 빈도수를 기준으로 선정된다. 이렇게 산출된 토픽의 상위 단어들은 여러 토픽의 상위 단어에 중복되어 나타날 수 있는데, 한 토픽에 출현 빈도가 높은 단어가 다른 토픽에도 출현 빈도가 높을 수 있기 때문이다. 따라서 빈도수만을 기준으로 토픽을 대표하는 상위 단어를 선택하면 토픽에 대한 분별성이 낮아진다. 따라서 본 논문에서는 변별성이 있는 차별화된 토픽의 상위 단어 집합을 선정하기 위하여 [10]에서 제안한 Relevance를 사용한다. Relevance란 해당 토픽에 대한 분별성을 높이고, 다른 토픽과는 차별화된 상위 단어들을 찾기 위한 방법으로 제안되었으며, <수식 2>를 사용하여 계산한다[10].

$$< \text{수식 } 2 > \quad \text{Relevance}(t, w) = \lambda \cdot P(w|t) + (1-\lambda) \cdot \frac{P(w|t)}{P(w)}$$

Relevance(t,w)는 토픽 t에서 단어 w의 Relevance 값이고, $\lambda(0 \le \lambda \le 1)$는 가중치를 나타내는 파라미터이다. $p(w|t)$는 토픽 t에서 해당 단어가 발생할 수 있는 확률이고, P(w)는 코퍼스에서 단어 w가 발생할 확률이다. 코퍼스란 전체 문서를 전처리한 말뭉치로 LDA 모델링의 입력으로 사용된다. λ값이 1이면, relevance 값은 $p(w|t)$이 된다. 즉 한 토픽에 자주 등장하는 단어일수록 높은 relevance 값을 가지게 되므로, 한 토픽에 등장하는 단어의 빈도수만을 가지고 토픽에 기여하는 상위 단어들을 찾을 수 있다. 반면, λ가 0이면 relevance의 값은 $p(w|t)/p(w)$이다. 한 토픽에서 해당 단어 w가 발생할 수 있는 확률을 코퍼스 전체에서 단어 w가 발생할 확률로 나눈 값이다. 즉, 한 토픽에 자주 등장하는 단어일지라도 다른 토픽에도 자주 등장하는 단어라면 relevance 값은 낮아지게 되고, 다른 토픽에 등장하지 않고 한 토픽에만 등장하는 단어일수록 relevance 값은 높다. 따라서 토픽에 기여하는 상위 단어들을 추출할 때, 적절한 λ 값을 설정하여 relevance 값을 구하면, 한 토픽에 출현하는 빈도가 높으면서 변별성이 있는 다른 토픽과는 차별화된 상위 단어들을 추출할 수 있다. [10]에서는 다양한 실험을 통하여 최적의 λ의 값 0.6을 제안했다. 본 논문에서도 λ의 값으로 0.6을 사용하고, Relevance의 값을 계산하여 토픽 별로 토픽을 대표하는 상위 단어들을 20개씩 선정하여 사용한다. 토픽별로 Relevance을 사용하여 산출한 상위 단어 20개를 'Relevance Top-20' 이라고 부른다.

포괄성 있는 토픽 대표 단어들을 생성하기 위한 방법은 다음 3.2절에서 기술한다.

3.2 TextRank을 이용한 토픽의 키워드 추출

토픽 별로 Relevance Top-20을 선정한 후, 이를 기반으로 토픽을 잘 표현할 수 있는 포괄성이 있는 대표 단어들을 산출해야 한다. 먼저, TextRank을 사용하여 토픽의 핵심 키워드를 찾아 토픽을 대표하는 단어 집합을 생성한다. 이를 위해 각 토픽별로 토픽에 기여하는 문서들을 모두 찾은 후, 각 문서에서 Relevance Top-20을 포함하는 문장들을 모두 찾는다. 찾아낸 모든 문장들은 LDA 토픽 모델링과 같은 전처리 과정을 통해 불용어를 제거하고 각 문장을 명사와 대명사의 형태소를 가진 n-gram 토큰으로 구성한 후, TextRank 알고리즘의 입력으로 사용한다. TextRank 알고리즘은 한 문장내에서 n-gram 토큰을 하나의 단어로 취급하여 정점으로 삼고, 단어들이 동시에 출현

하는 상태를 간선으로 표현한다. 그리고 한 문장내에서 단어들이 동시 출현하는 빈도를 가중치로 하여, <수식 1>에 의해 해당 단어의 TextRank 값을 계산하고, TextRank 값이 높은 단어들을 토픽의 키워드들로 추출한다.

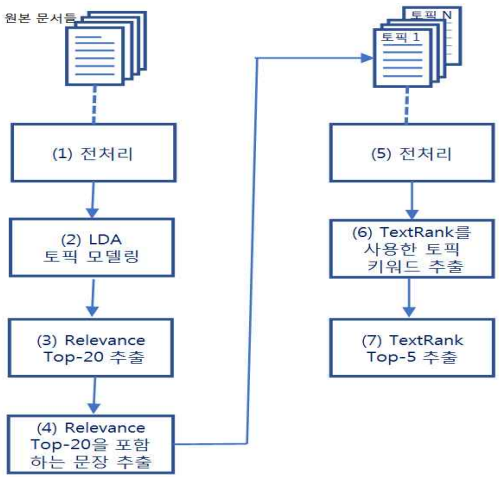
좀 더 포괄적으로 토픽을 잘 표현하기 위해서는 하나의 단어보다 문구(phrase)를 사용하는 것이 더 적합하다. 따라서 TextRank 알고리즘으로 찾은 토픽 키워드들은 n-gram 단어들이므로, 이 단어들이 앞뒤로 동시에 출현하는 단어의 빈도를 구하고, 높은 빈도를 가진 단어들을 앞뒤로 연결하여 문구를 만들어 토픽의 키워드에 추가한다. 이러한 과정을 거쳐 생성한 토픽의 키워드들 중에 TextRank 값이 큰 5개를 토픽을 대표하는 핵심 키워드로 선정한다. TextRank를 사용해 각 토픽별로 산출한 토픽을 대표하는 5개의 키워드를 'TextRank Top-5' 라고 부른다.

한 토픽의 TextRank Top-5에 있는 토픽 키워드는 다른 토픽의 TextRank Top-5에 중복되어 나타날 수 있다. 이러한 경우는 TextRank 값이 더 높은 값을 가지는 토픽을 우선시 하여 토픽 키워드를 선정하고, 보다 낮은 TextRank 값을 가지는 토픽의 키워드는 제외시키고, 해당 토픽의 다음 순위를 가지는 키워드를 TextRank Top-5에 선정하여 토픽의 변별성을 높인다.

3.3 TextRank을 사용한 토픽 키워드 생성 과정

토픽 레이블링을 위한 토픽을 대표하는 단어 즉 토픽 키워드를 생성하기 위한 과정은 <그림 1>과 같다.

(1) LDA토픽 모델링을 위하여 문서들을 전처리 한다.
LDA토픽 모델링을 위해 한글형태소 분석기를 사용하여 명사, 대명사형 단어들을 추출하고, 문서에서 불용어로 판단되는 단어들을 삭제한다. 연속적인 여러 개의 단어로 구성된 의미 있는 문구(phrase)가 의미를 가지며 반복적으로 문서에 나타나는 경우도 많기 때문에 n-gram으로 토큰화하여 코퍼스를 준비한다.



<그림 1> 토픽 키워드를 생성하기 위한 처리 절차

(2) LDA 토픽 모델링을 수행한다.
LDA는 직접 관찰할 수 있는 코퍼스 내의 각 문서들의 단어 분포로부터 주제의 단어 분포를 예측하는 사후 추론 방법이기 때문에, 먼저 토픽의 개수를 정해줘야 토픽 모델을 만들 수 있다. 토픽의 개수는 해당 토픽 모델이 잘 만들어졌는지 평가하는 의미론적인 일관성(coherence)에 영향을 미치게 된다. LDA 모델의 Coherence 값은 토픽 모델이 한 주제 안에서 얼마나 의미적으로 유사한 단어들이 모였는지를 판단할 수 있는 값으로써 coherence 값이 높을수록 일관성이 높다고 판단한다. 토픽 모델이 의미론적으로 일관성이 높게 모델링이 되어야 각 토픽의 레이블 또한 잘 선정할 수 있으므로, coherence가 높게 나오는 토픽의 개수를 찾아 해당 LDA 모델을 토픽 레이블링에 사용한다.

(3) Relevance 값을 계산하여 각 토픽의 Relevance Top-20 단어 추출한다.

(4) 각 토픽별로 토픽에 기여하는 문서들을 대상으로 Relevance Top-20 단어를 포함하는 문장들을 추출한다.
(5) 각 토픽별로 추출한 문장들을 TextRank를 위하여 전처리 한다.
LDA 토픽 모델링을 위한 전처 과정과 같이 형태소 분석과 불용어를 제거하여 n-gram으로 토큰화된 단어 리스트를 생성한다.
(6) TextRank를 알고리즘을 사용하여 각 토픽을 대표하는 단어로 토픽 키워드들을 추출한다.
(7) 토픽 키워드들 중 연속적으로 동시에 출현하는 키워드들을 문구로 만들어 TextRank Top-5를 추출한다.

Ⅳ. 실험 및 평가

4.1 실험 결과

본 논문에서는 제안한 TextRank을 사용한 토픽 키워드 산출 방법의 실험을 위해 네이버 뉴스 사이트에서 2020년 4월 1일부터 4월 30일까지 한 달 간의 IT/과학 관련 뉴스, 총 20,034개의 문서를 수집하여 사용하였다. LDA 토픽 모델을 위해서는 Mallet[11]을 사용하였고, 토픽 모델의 시각화를 하기 위해서는 LDA 모델의 학습 결과를 시각적으로 표현하는 LDAvis[10]의 Python wrapper인 genism[12]의 pyLDAvis 모듈을 사용하였다. 한글 전처리 과정에서 사용하는 형태소 분석기는 Komoran[13]을 사용하였다. 실험은 Intel(R) Core(TM) i7-7500U CPU@2.70GHz 2.90GHz, 8GB RAM, Microsoft Window 10 Home 운영체제로 구성된 머쉰에서 수행하였으며, 프로그램에 사용한 언어는 Python 3.8.3이다.

실험에 사용한 토픽의 개수는 12개이다. 다양한 토픽 개수를 설정해 토픽 모델링을 실행하고 coherence 값을 측정하여 상대적으로 높은 coherence값을 보인 토픽 개수를 선택하였다. 실험결과 토픽의 개수가 12개 때, coherence 값은 0.5704로 비교적 높은 일관성을 보였다.

<그림 2>는 토픽 개수가 12개일 때, 토픽 모델링의 결과를 pyLDAvis를 사용하여 시각화한 것이다. <그림 2> 왼쪽의 다이어그램은 토픽 벡터를 2차원으로 축소하여 토픽간의 관계를 보여준다. 비슷한 위치에 존재하는 토픽들은 비슷한 문맥을 지니고 있고, 겹치는 부분이 최소화되고 사분면의 좌표상에 골고루 배치되어 있을수록 좋은 모델이다. 토픽 개수가 12개인 토픽 모델은 겹치는 부분이 많지 않고 사분면에 골고루 배치되어 있으므로 비교적 좋은 토픽 모델이다. <그림 2>의 오른쪽 막대 그래프는 각 토픽의 상위 단어들을 보여주는데, 현재 <그림 2>에서는 토픽 1에 상위 30개의 단어들을 보여주고 있다.

<표 1>은 LDA 토픽 모델링을 통해 산출한 12개의 각 토픽의 상위 20개의 단어들과 각 단어가 해당 토픽에 얼마나 기여하는지를 반영하는 가중치를 나타낸 것이다. LDA 토픽 모델은 해당 단어의 토픽 기여도를 해당 토픽에 출현한 단어의 빈도수에 의해 결정하므로 <표 1>에 나타난 가중치는 $p(w|t)$이다. LDA 토픽 모델링을 통해 산출한 토픽의 상위 20개 단어들을 'Top-20' 이라고 부른다.

분별성이 있는 차별화된 상위 단어들을 산출하기 위해 <수식 2>에 의해 relevance를 계산하고, relevance 값이 높은 순으로 Relevance top-20 단어들을 산출하였다. <표 2>는 토픽 별 Relevance Top-20 단어들과 각각의 relevance 값이다. <표 1>과 비교할 때 토픽 별로 단어의 순위가 다소 변경이 되었다. <표 1>에서 '토픽 1'의 상위 단어들은 '온라인, 원격, 제공, 진행, 교육…' 순이었지만, Relevance를 기준으로 다시 산출한 결과 <표 2>를 보면 '온라인, 원격, 온라인_개학, 학생, 교육…' 순이다. '제공, 진행' 등의 단어들은 다른 토픽에도 중복되는 단어들이므로

Relevance 값을 기준으로 상위 단어를 다시 산출하면 순위가 떨어진다.

토픽을 대표하는 키워드를 추출하기 위하여 LDA 토픽 모델에 사용한 20,034개의 전체 뉴스 문서에서 Relevance Top-20 단어를 포함하는 문장들을 토픽별로 추출하고 전처리 한다. 3.2절에서 제안한 방법에 따라 <수식 1>의 TextRank 알고리즘으로 TextRank 값을 계산하고, 토픽을 포괄적으로 잘 표현할 수 있도록 앞, 뒤로 연속적으로 출현하는 n-gram 단어들을 합쳐 문구를 만들었다. <표 3>은 이러한 과정으로 만들어진 토픽 별 상위 10개의 키워드로 각 항목마다 (앞 단어, 뒤 단어, TextRank 값)의 쌍으로 구성되어 있다. 문구를 이루지 않는 키워드는 '뒤 단어' 항목이 없다.

<표 3>의 TextRank Top-10중에서 최종적으로 상위 5개의 키워드만 선택한다. 그런데 TextRank Top-10을 살펴보면 토픽 키워드들이 여러 토픽에 중복되어 나타나고 있다. 이렇게 중복된 키워드는 빨간색 폰트로 표시하였다. 예를 들어 '토픽 3'의 1순위 키워드인 '기업(0.0368)'은 '토픽 5'의 '기업(0.0417)'과 중복된다. 중복성을 제거하여 분별성이 있는 차별화된 토픽 레이블링을 위해 '토픽 5'의 '기업'의 TextRank 값이 0.0417로 더 큰 값을 가지므로 그대로 TextRank Top-5에 유지하고, '토픽 3'의 '기업'은 '토픽 3'의 TextRank Top-5에서 제외시킨다. '토픽 3'의 TextRank Top-5에는 6순위의 '정부'라는 키워드를 선정한다. <표 4> 중복성을 제거한 후 최종적으로 산출된 TextRank Top-5 나타내고 있다.

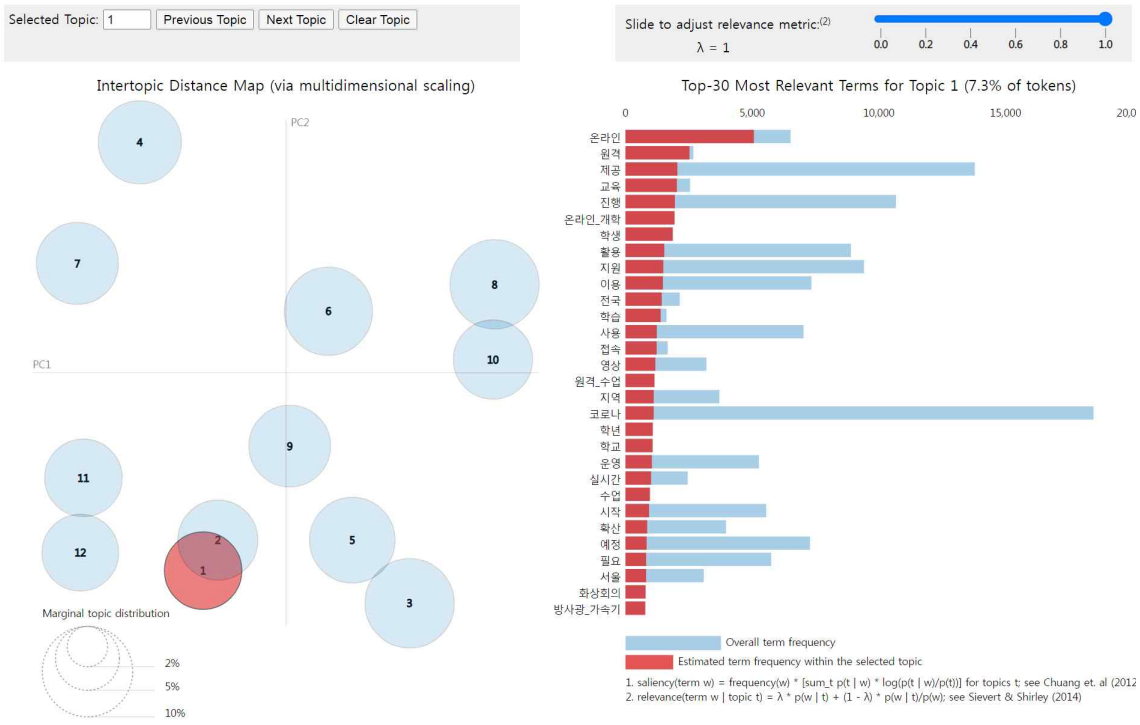
4.2 평가

본 논문에서 제안한 TextRank를 사용한 토픽 키워드 산출 방법을 평가하기 위하여 <표 1>, <표 2>, <표 4>에서 산출한 Top-5, Relevance Top-5, TextRank Top-5를 각각 비교하여 산출한 토픽을 대표하는 단어 집합이 분별성과 포괄성을 제공하여 토픽을 이해하기 쉬우며 토픽을 잘 표현하는지를 검토하였다. <표 5>는 토픽 6, 9, 12를 예를 들어 세 가지 방법이 산출한 Top-5를 비교한 예이다.

<표 5>에서 '토픽 6'의 Top-5 키워드는 '분기, 코로나, 증가, 기록, 지난해' 이다. Relevance Top-5는 '분기, 증가, 기록, 코로나, 영업' 이고, TextRank Top-5는 '분기, 영업 이익, 코로나 영향, 전년 동기, 동기 대비 증가' 이다. TextRank Top-5 키워드가 다른 방법에 비해 핵심어 추출을 통해 토픽의 키워드를 문구로 표현하므로 포괄적이며 타 토픽과 분별성이 더 있어서 5개의 토픽 단어들을 통해 토픽을 더 잘 이해할 수 있다. 반면 <표 1>에서 '토픽 1'의 LDA 모델링의 산출로 선택한 Top-20 단어들과 <표 2>에서 '토픽 1'의 Relevance Top-20의 단어들은 아무리 많은 단어가 제공되더라도 어떤 단어를 어떻게 결합해서 토픽을 이해해야 하는지 도메인에 전문적 지식이 없는 사람이라면 상위단어를 통해 토픽을 레이블링 하기가 매우 어렵다.

전문가도 토픽의 레이블을 선정할 때는 토픽을 대표하는 상위 키워드를 보고 추정한다. 자동화 과정을 통해서 선택한 토픽을 대표하는 단어집합이 토픽을 레이블링 하기에 적합한지를 평가하는 정량적인 공인된 방법은 아직까지 부재하며 전문가들의 검토를 통하여 적합성 여부를 판단하는 것이 일반적이다. 전문가 3인에게 <표 1>, <표 2>, <표 4>에서 제시한 세 가지 방법의 Top-5와 Top-10의 검토를 받은 결과, 본 논문에서 제안한 TextRank 키워드 추출방식을 통해 산출한 Top-5, Top-10이 가장 분별성이 있고 토픽의 의미를 잘 추정할 수 있는 포괄성이 있어 이해하기 쉽다는 평가를 받았다.

토픽 레이블링의 전제조건은 토픽 모델이 의미적으로 일관성 있게 잘 모델링 되어 있어야 한다는 것이다. <표 4>에서 '토픽 1'의 '방사광 가속기'는 '토픽 1'의 다른 키워드와 비교할 때 일관성이 떨어지는데, 이는 12개의 토픽으로 전체문서를 주제 분류를 하면



<그림 2> pyLDAvis를 이용한 토픽 개수가 12개인 LDA 모델의 시각화

<표 1> LDA 모델이 산출한 토픽 별 상위 20개 단어: Top-20

Ranking	Topic 1		Topic 2		Topic 3		Topic 4		Topic 5		Topic 6		Topic 7		Topic 8		Topic 9		Topic 10		Topic 11		Topic 12	
1	온라인	0.0216	서비스	0.014	지원	0.0098	출시	0.0204	기업	0.015	분기	0.0229	게임	0.0285	코로나	0.0203	네이버	0.0073	개발	0.0144	코로나	0.0139	제공	0.0141
2	원격	0.0108	이용	0.0072	기업	0.0092	제품	0.0164	제공	0.0116	코로나	0.0204	진행	0.0105	연구	0.0087	관련	0.0065	기술	0.0138	진행	0.0127	서비스	0.0131
3	제공	0.0088	넷플릭스	0.0066	사업	0.008	스마트폰	0.0126	데이터	0.0113	증가	0.0142	이용자	0.01	개발	0.0079	확인	0.0065	블록_체인	0.0093	지원	0.0071	고객	0.0099
4	교육	0.0087	국내	0.006	정부	0.0071	아이폰	0.012	클라우드	0.0099	기록	0.011	모바일	0.0095	바이러스	0.0074	정보	0.0061	활용	0.0083	지역	0.007	콘텐츠	0.0087
5	진행	0.0083	카카오	0.006	기술	0.0062	탑재	0.0098	서비스	0.0098	지난해	0.0108	신규	0.0079	백신	0.0063	사용자	0.0053	블록	0.0069	제공	0.0062	sk텔레콤	0.0085
6	온라인_개학	0.0083	결제	0.0058	계획	0.0062	lg전자	0.009	업무	0.0089	올해	0.0105	추가	0.007	환자	0.0054	내용	0.0043	사용	0.0067	거리	0.0054	이용	0.0073
7	학생	0.008	통신	0.0052	추진	0.006	애플	0.0088	기술	0.0083	중국	0.0092	공개	0.0058	확인	0.0052	이용자	0.0042	적용	0.0056	참여	0.0049	배달	0.0067
8	활용	0.0065	제공	0.0052	투자	0.0059	가격	0.008	솔루션	0.0066	전망	0.0087	콘텐츠	0.0057	연구_팀	0.0052	구글	0.0038	가상	0.0051	배_송	0.0046	소상_공인	0.0047
9	지원	0.0064	확대	0.0051	산업	0.0053	삼성전자	0.0078	고객	0.0065	매출	0.0085	출시	0.0055	미국	0.005	해당	0.0038	가상_자산	0.0047	예정	0.0043	수수료	0.0045
10	이용	0.0063	가입자	0.005	선정	0.0053	모델	0.0076	기반	0.0062	영업	0.0084	이벤트	0.0055	감염	0.0047	개인정보	0.0036	반도체	0.0041	마스크	0.004	무료	0.0044
11	전국	0.0061	지난해	0.0049	과학기술_정보통신부	0.0051	적용	0.0071	보안	0.0062	영업_이익	0.0081	인기	0.0054	치료제	0.0046	최근	0.0035	이용	0.0037	전달	0.0039	운영	0.0042
12	학습	0.0059	고객	0.0047	과기_정통부	0.005	디자인	0.0069	활용	0.006	전년	0.008	업데이트	0.0051	진단	0.0043	기능	0.0031	기존	0.0037	캠페인	0.0039	국내	0.0041
13	사용	0.0053	계획	0.0043	컴퓨	0.0048	카메라	0.0066	글로벌	0.0056	시장	0.0074	캐릭터	0.0044	확진	0.0041	사용	0.0029	거래	0.0034	기부	0.0038	요금	0.0041
14	접속	0.0053	플랫폼	0.0042	분야	0.0045	공개	0.0061	지원	0.0056	예상	0.0073	예정	0.0044	세계	0.004	페이스북	0.0029	기반	0.0032	어려움	0.0037	주문	0.0038
15	영상	0.005	sk브로드밴드	0.0041	활용	0.004	세대	0.0055	구축	0.0055	달러	0.0073	확득	0.0043	임상	0.0038	지적	0.0028	필요	0.0031	구매	0.0037	소상	0.0037
16	원격_수업	0.0049	lg유플러스	0.004	국가	0.0039	벨벳	0.0055	디지털	0.005	영향	0.0069	글로벌	0.004	분석	0.0036	발생	0.0027	성능	0.003	온라인	0.0034	최대	0.0036
17	지역	0.0048	금융	0.0034	참여	0.0039	사용	0.0053	도입	0.0049	미국	0.0067	넥슨	0.004	결과	0.0035	승인	0.0027	로봇	0.003	대상	0.0033	광고	0.0034
18	코로나	0.0048	이용자	0.0032	과학기술	0.0038	화소	0.0051	국내	0.0045	삼성전자	0.006	시작	0.0038	의료	0.0034	공격	0.0027	계획	0.0027	고객	0.0032	오픈	0.0034
19	학년	0.0046	업계	0.0032	예정	0.0037	지원	0.005	회사	0.0042	감소	0.006	등장	0.0031	치료	0.0034	추가	0.0026	안전	0.0027	확산	0.0032	상품	0.0033
20	학교	0.0046	네이버	0.0032	혁신	0.0037	기능	0.0048	관리	0.0041	성장	0.0058	공식	0.003	효과	0.0033	검색	0.0025	에너지	0.0026	마음	0.0031	유튜브	0.0033

<표 2> Relevance를 기준으로 산출한 토픽 별 상위 20개의 단어: Relevance Top-20

Ranking	Topic 1		Topic 2		Topic 3		Topic 4		Topic 5		Topic 6		Topic 7		Topic 8		Topic 9		Topic 10		Topic 11		Topic 12	
1	온라인	0.0216	서비스	0.014	지원	0.0098	출시	0.0204	기업	0.015	분기	0.0229	게임	0.0285	코로나	0.0203	네이버	0.0073	개발	0.0144	코로나	0.0139	제공	0.0141
2	원격	0.0108	이용	0.0072	기업	0.0092	제품	0.0164	제공	0.0116	코로나	0.0204	진행	0.0105	연구	0.0087	관련	0.0065	기술	0.0138	진행	0.0127	서비스	0.0131
3	제공	0.0088	넷플릭스	0.0066	사업	0.008	스마트폰	0.0126	데이터	0.0113	증가	0.0142	이용자	0.01	개발	0.0079	확인	0.0065	블록_체인	0.0093	지원	0.0071	고객	0.0099
4	교육	0.0087	국내	0.006	정부	0.0071	아이폰	0.012	클라우드	0.0099	기록	0.011	모바일	0.0095	바이러스	0.0074	정보	0.0061	활용	0.0083	지역	0.007	콘텐츠	0.0087
5	진행	0.0083	카카오	0.006	기술	0.0062	탑재	0.0098	서비스	0.0098	지난해	0.0108	신규	0.0079	백신	0.0063	사용자	0.0053	블록	0.0069	제공	0.0062	sk텔레콤	0.0085
6	온라인_개학	0.0083	결제	0.0058	계획	0.0062	lg전자	0.009	업무	0.0089	올해	0.0105	추가	0.007	환자	0.0054	내용	0.0043	사용	0.0067	거리	0.0054	이용	0.0073
7	학생	0.008	통신	0.0052	추진	0.006	애플	0.0088	기술	0.0083	중국	0.0092	공개	0.0058	확인	0.0052	이용자	0.0042	적용	0.0056	참여	0.0049	배달	0.0067
8	활용	0.0065	제공	0.0052	투자	0.0059	가격	0.008	솔루션	0.0066	전망	0.0087	콘텐츠	0.0057	연구_팀	0.0052	구글	0.0038	가상	0.0051	배_송	0.0046	소상_공인	0.0047
9	지원	0.0064	확대	0.0051	산업	0.0053	삼성전자	0.0078	고객	0.0065	매출	0.0085	출시	0.0055	미국	0.005	해당	0.0038	가상_자산	0.0047	예정	0.0043	수수료	0.0045
10	이용	0.0063	가입자	0.005	선정	0.0053	모델	0.0076	기반	0.0062	영업	0.0084	이벤트	0.0055	감염	0.0047	개인정보	0.0036	반도체	0.0041	마스크	0.004	무료	0.0044
11	전국	0.0061	지난해	0.0049	과학기술_정보통신부	0.0051	적용	0.0071	보안	0.0062	영업_이익	0.0081	인기	0.0054	치료제	0.0046	최근	0.0035	이용	0.0037	전달	0.0039	운영	0.0042
12	학습	0.0059	고객	0.0047	과기_정통부	0.005	디자인	0.0069	활용	0.006	전년	0.008	업데이트	0.0051	진단	0.0043	기능	0.0031	기존	0.0037	캠페인	0.0039	국내	0.0041
13	사용	0.0053	계획	0.0043	컴퓨	0.0048	카메라	0.0066	글로벌	0.0056	시장	0.0074	캐릭터	0.0044	확진	0.0041	사용	0.0029	거래	0.0034	기부	0.0038	요금	0.0041
14	접속	0.0053	플랫폼	0.0042	분야	0.0045	공개	0.0061	지원	0.0056	예상	0.0073	예정	0.0044	세계	0.004	페이스북	0.0029	기반	0.0032	어려움	0.0037	주문	0.0038
15	영상	0.005	sk브로드밴드	0.0041	활용	0.004	세대	0.0055	구축	0.0055	달러	0.0073	획득	0.0043	임상	0.0038	지적	0.0028	필요	0.0031	구매	0.0037	소상	0.0037
16	원격_수업	0.0049	lg유플러스	0.004	국가	0.0039	벨벳	0.0055	디지털	0.005	영향	0.0069	글로벌	0.004	분석	0.0036	발생	0.0027	성능	0.003	온라인	0.0034	최대	0.0036
17	지역	0.0048	금융	0.0034	참여	0.0039	사용	0.0053	도입	0.0049	미국	0.0067	넥슨	0.004	결과	0.0035	승인	0.0027	로봇	0.003	대상	0.0033	광고	0.0034
18	코로나	0.0048	이용자	0.0032	과학기술	0.0038	화소	0.0051	국내	0.0045	삼성전자	0.006	시작	0.0038	의료	0.0034	공격	0.0027	계획	0.0027	고객	0.0032	오픈	0.0034
19	학년	0.0046	업계	0.0032	예정	0.0037	지원	0.005	회사	0.0042	감소	0.006	등장	0.0031	치료	0.0034	추가	0.0026	안전	0.0027	확산	0.0032	상품	0.0033
20	학교	0.0046	네이버	0.0032	혁신	0.0037	기능	0.0048	관리	0.0041	성장	0.0058	공식	0.003	효과	0.0033	검색	0.0025	에너지	0.0026	마음	0.0031	유튜브	0.0033

<표 3> TextRank로 산출한 키워드 상위 10개 키워드: TextRank Top-10

Ranking	Topic 1		Topic 2		Topci 3		Topic4		Topic 5		Topic 6	
1	온라인	온라인_개학 0.0968	서비스	0.0469	기업	0.0368	출시	0.0513	서비스	제공 0.048	분기	0.0751
2	원격	원격_수업 0.0656	이용	0.0357	지원	0.0348	제품	0.0492	기업	0.0417	영업	영업_이익 0.0691
3	방사광	방사광_가속기 0.0454	넷플릭스	sk브로드밴드 0.0347	과학기술	과학기술_정보통신부 0.0334	아이폰	0.0472	클라우드	0.0353	코로나	영향 0.0678
4	교사	학생 0.0408	국내	0.0335	기술	0.0333	스마트폰	0.0417	데이터	0.0336	전년	전년_동기 0.0559
5	학습	학습_터 0.0385	카카오	카카오_페이 0.0315	사업	0.029	애플	0.0361	고객	0.0308	동기_대비	증가 0.0418
6	교육	0.0348	결제	결제_서비스 0.0305	정부	0.026	탑재	0.0354	기술	0.0306	올해	0.0333
7	제공	0.0327	지난해	0.0246	투자	0.0243	카메라	화소 0.0351	업무	0.0285	중국	0.033
8	수업	진행 0.0306	플랫폼	0.0228	추진	계획 0.0218	삼성전자	0.0349	보안	0.0277	지난해	0.0306
9	초등학교	학년 0.0272	확대	0.0205	산업	0.0205	lg전자	벨벳 0.0333	솔루션	0.0223	매출	0.03
10	영상	영상_회의 0.0263	금융	0.0203	과기	과기_정통부 0.0198	적용	0.0297	기반	0.0218	달러	기록 0.0295
Ranking	**Topic 7**		**Topic 8**		**Topic 9**		**Topic 10**		**Topic 11**		**Topic12**	
1	게임	0.0756	코로나	0.0683	코로나	관련 0.0449	블록	블록_체인 0.0924	코로나	0.0601	배달	민족 0.0655
2	이벤트	진행 0.042	연구	연구_팀 0.0453	네이버	0.0351	가상	가상_자산 0.0678	진행	0.042	소상	소상_공인 0.053
3	모바일	모바일_게임 0.0395	치료제	백신 0.0395	정보	제공 0.03	기술	개발 0.0513	지원	0.0336	서비스	제공 0.0495
4	이용자	0.0386	개발	0.0339	확인	0.0278	활용	0.0353	어려움	지역 0.0336	고객	0.0386
5	신규	0.0314	바이러스	0.0316	사용자	0.0269	사용	0.0303	구매	고객 0.0268	공공_배달	배달_앱 0.0362
6	추가	0.0289	진단	진단_키트 0.0297	내용	0.0232	적용	0.0253	캠페인	참여 0.024	콘텐츠	0.0319
7	콘텐츠	0.0254	환자	0.0282	해당	0.0231	성능	반도체 0.022	배_송	0.019	이용	0.0314
8	캐릭터	0.0251	미국	0.0271	선거	0.0217	비트코인	거래 0.0179	응원	마음 0.0187	배_민	수수료 0.0301
9	아이템	획득 0.0244	감염	0.0215	개인정보	0.0212	전기	생산 0.0131	전달	예정 0.0177	요금	요금_체계 0.0291
10	넥슨	인기 0.0242	임상	0.0215	공격	사용 0.0199	전력	0.0124	거리	0.0176	오픈	오픈_서비스 0.0275

서 의미를 고려하지 않고 출현하는 단어의 빈도수에 의존하는 LDA 토픽 모델링의 한계로 인해 생긴 문제점이다.

또한 본 논문에서는 최종적으로 TextRank Top-5를 선택할 때, 변별성을 높이기 위해서 토픽들 간에 서로 중복되어 선택되는 키워드는 TextRank 값을 비교하여 값이 작은 키워드를 해당 TextRank Top-5에서 제외를 시켰다. 예를 들면 <표 3>에서 '기업, 지원, 기술, 서비스 제공, 고객, 코로나' 같은 키워드들은 서로 중복되므로 중복을 제거하여 토픽 레이블의 분별성

<표 4> 최종 선정된 토픽 키워드: TextRank Top-5

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
온라인 개학	서비스	지원	출시	기업	분기
원격 수업	이용	과학기술 정보통신부	제품	클라우드	영업 이익
방사광 가속기	넥플릭스 sk브로드밴드	기술	아이폰	데이터	코로나 영향
교사 학생	국내	사업	스마트폰	업무	전년 동기
학습터	카카오_페이	정부	애플	보안	동기 대비 증가
Topic 7	**Topic 8**	**Topic 9**	**Topic 10**	**Topic 11**	**Topic 12**
게임	코로나	코로나 관련	블록 체인	진행	배달 민족
이벤트 진행	연구팀	네이버	가상 자산	어려움 지역	소상 공인
모바일 게임	치료제 백신	정보 제공	기술 개발	구매 고객	서비스 제공
이용자	개발	확인	활용	캠페인 참여	고객
신규	바이러스	사용자	사용	배송	공공 배달 앱

<표 5> LDA 모델, Relevance , TextRank가 산출한 Top-5 키워드 비교 예

Topic 6	
LDA 모델	분기, 코로나, 증가, 기록, 지난해
Relevance	분기, 증가, 기록, 코로나, 영업
TextRank	분기, 영업 이익, 코로나 영향, 전년 동기, 동기 대비 증가
Topic 9	
LDA 모델	네이버, 관련, 확인, 정보, 사용자
Relevance	네이버, 정보, 개인정보, 관련, 사용자
TextRank	코로나 관련, 네이버, 정보 제공, 확인, 사용자
Topic 12	
LDA 모델	제공, 서비스, 고객, 콘텐츠, sk텔레콤
Relevance	sk텔레콤, 배달, 서비스, 제공, 콘텐츠
TextRank	배달 민족, 소상 공인, 서비스 제공, 고객, 공공 배달 앱

을 높였다. 그런데 '토픽 8'의 '코로나'와 '토픽 11'의 '코로나'의 경우는 TextRank Top-5에서 중복되어 '토픽 11'의 '코로나'를 최종적으로 제외시켜 '토픽 11'의 TextRank Top-5는 '진행, 어려움 지역, 구매 고객, 캠페인 참여, 배송'이 선정되었다. 결과적으로 '토픽 11'의 TextRank Top-5는 '코로나'라는 4월 당시 이슈의 중심에 있던 키워드가 변별성을 높이기 위해 삭제되어 '토픽 11'를 의미적으로 잘 표현할 수 없는 문제가 발생하였다. 이러한 문제점은 본 논문에서 제안한 토픽 키워드 산출 방법의 한계로서 변별성을 높이기 위해 일괄적으로 중복된 키워드를 삭제하는 것이 아니라 토픽을 의미적으로 잘 표현할 수 있다면 키워드의 중복을 허용하도록 예외상황에 대한 추가적인 연구가 더 필요하다.

V. 결론 및 향후 연구

LDA 토픽 모델링을 사용하여 산출한 토픽들은 토픽의 레이블이 자동으로 붙여지지 않기 때문에, 토픽을 분별하고 잘 이해하기 위해 토픽의 레이블을 별도로 붙여야 하는 번거로움이 있다. 토픽의 레이블링을

할 때는 토픽을 대표하는 단어들을 살펴보고 레이블을 붙이게 되므로, 먼저 토픽을 대표하는 단어들의 집합을 잘 만드는 것이 중요하다.

본 논문은 문서의 키워드를 추출하는 TextRank를 사용하여 토픽을 대표하는 키워드들을 산출하는 방법을 제안하였다. 제안된 방법은 변별성이 있는 토픽 관련 단어들을 선정하기 위해 Relevance를 사용하며, TextRank 알고리즘을 사용해 토픽 키워드를 추출하고, 동시 출현 빈도가 높은 키워드들을 연결하여 토픽을 좀 더 포괄적으로 표현할 수 있게 하는 특징이 있다.

본 논문에서 제안한 토픽 키워드 산출 방법은 최종적으로 토픽을 대표하는 키워드를 선택할 때 변별성을 높이기 위해 토픽들 간에 중복된 키워드를 삭제하면서 토픽을 의미적으로 잘 표현하지 못하는 경우가 발생하는 한계가 있다. 따라서 향후, 이러한 예외상황을 처리하기 위한 추가적인 연구를 진행할 예정이다. 또한, TextRank를 사용하여 산출한 토픽 키워드들을 기반으로 토픽을 의미적으로 잘 표현할 수 있도록 자연어처리 기술을 활용한 토픽 레이블 선정 방법을 연구해 나갈 예정이다.

참고문헌

[1] 박종순, 김창식, “빅데이터 연구동향 분석: 토픽 모델링을 중심으로,” 디지털산업정보학회논문지, 제15권, 제1호, 2019, pp.1-7.

[2] 김창식, 김남규, 곽기영, “머신러닝 및 딥러닝 연구동향 분석: 토픽모델링을 중심으로,” 디지털산업정보학회논문지, 제15권, 제2호, 2019, pp.19-28.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, Vol. 3, Mar. 2003, pp. 993–1022..

[4] Q. Mei, X. Shen, and C.X. Zhai, “Automatic labeling of multinomial topic models.” In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2007, pp.490–499.

[5] Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin, “Best topic word selection for topic labelling,” In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING ’10), Association for Computational Linguistics, 2010, pp.605–613.

[6] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin, “Automatic labelling of topic models,” In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, Association for Computational Linguistics, 2011, pp.1536–1545.

[7] Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene, “Unsupervised graph-based topic labelling using dbpedia,” In Proceedings of the sixth ACM international conference on Web search and data mining (WSDM ’13), Association for Computing Machinery, 2013, pp.465–474.

[8] S. Bhatia, J. H. Lau, and T. Baldwin, “Automatic labelling of topics with neural embeddings,” in 26th COLING International Conference on Computational Linguistics, 2016, pp.953–963.

[9] Mihalcea, Rada and Tarau, Paul, “TextRank: Bringing Order into Text,” Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Association for

Computational Linguistics, Jul. 2004, pp.404-411.
[10] Carson Sievert and Kenneth E. Shirley, "LDAvis: A method for visualizing and interpreting topics," Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 2014, pp.63-70.
[11] Mallet, http://mallet.cs.umass.edu/
[12] Gensim, https://radimrehurek.com/gensim/
[13] Komoran, https://www.shineware.co.kr/products/komoran/

논문접수일 : 2020년 8월 26일
수 정 일 : 2020년 9월 7일
게재확정일 : 2020년 9월 15일

■ 저자소개 ■

김 은 회
(Kim, Eunhoe)

2013년 3월~현재
서일대학교 소프트웨어공학과
조교수
2009년 9월~2013년 2월
숭실대학교 지능형로봇연구소
전임연구원
2007년 9월~2009년 8월
숭실대학교 정보미디어기술연구소
전임연구원
2006년 8월 숭실대학교 컴퓨터학과(공학박사)
1998년 8월 숭실대학교 컴퓨터학과(공학석사)
1993년 2월 숭실대학교 전자계산학과(공학사)

관심분야 : 빅데이터, 텍스트마이닝, IoT, 분산처리, 병렬처리
E-mail : ehkim@seoil.ac.kr

서 유 화
(Suh, Yuhwa)

2019년 3월~현재
숭실대학교 베어드교양대학 조교수
2016년 3월~2019년 2월
서일대학교 정보통신공학과 조교수
2016년 2월 숭실대학교 컴퓨터학과(공학박사)
2007년 11월~2009년 10월
정보통신산업진흥원 연구원
2005년 8월 숭실대학교 컴퓨터학과(공학석사)
2003년 2월 숭실대학교 컴퓨터학부(공학사)

관심분야 : 그린네트워킹, 유무선네트워크, 빅데이터, 인공지능
E-mail : yhsuh@ssu.ac.kr