The Korean Journal of Applied Statistics
2021, Vol. 34, No. 5, 795–806

DOI: https://doi.org/10.5351/KJAS.2021.34.5.795
Print ISSN1225-066X / Online ISSN 2383-5818

Topic change monitoring study based on Blue House national petition using a control chart

Heeyeon Lee[a], Jieun Choi[b], Sungim Lee[1,b], Won Son[b]

[a]Department of Biostatistics of the Catholic Research Coordinating Center, Catholic University;
[b]Department of Statistics, Dankook University

Abstract

Recently, as text data through online channels have become vast, there is a growing interest in research that summarizes and analyzes them. One of the fundamental analyses of text data is to extract potential topics. Although the researcher may read all the data and summarize the contents one by one, it is not easy to deal with large amounts of data. Blei and Lafferty (2007) and Blei *et al.* (2003) proposed topic modeling methods for extracting topics using a statistical model. Since the text data is generally collected over time, it is worthwhile to monitor the topic's changes. In this study, we propose a topic index based on the results of the topic model. In addition, a control chart, a representative tool for statistical process management, is applied to monitor the topic index over time. As a practical example, we use text data collected from Blue House National Petition boards between March 5, 2018, and March 5, 2020.

Keywords: text data, LDA model, topic monitoring, EWMA chart

1. 서론

최근에는 소셜미디어를 비롯한 다양한 온라인 매체를 통하여 대용량 텍스트 자료가 쌓이고 이에 대한 분석 요구가 많아지고 있는 추세이다. 그러나, 이들 텍스트 자료에 대한 정리와 분석은 간단하지 않다. 예를 들어, 국민청원자료를 살펴보면 2017년 8월에 국민청원이 개설된 이래 2020년 7월 기준으로 약 44만 건의 청원이 있었다. 매년 국민청원으로 올라오는 청원 개수와 주제의 다양성은 증가하고 있지만, 실제로 이러한 내용을 정리하여 어떤 내용이 자주 청원 되는지 혹은 어떤 내용이 새로운 이슈로 떠오르는지를 정리하는 것은 쉽지 않은 일이다. 시간에 비례하여 청원의 양이 넘쳐나기 때문에 청원에 대한 내용을 일일이 읽고 파악하기에는 한계가 있기 때문이다.

텍스트 자료가 담고 있는 내용을 파악하는 데 있어서는 텍스트 자료의 전처리를 통해 정리한 문서와 단어 행렬(document term matrix, DTM)을 기반으로 하는 토픽모형이 자주 활용된다(Blei 등, 2003; Blei와 Lafferty, 2007). 이 모형은 자료의 시점과는 무관하게 수집된 자료 전체를 잠재적인 여러 토픽의 모임으로 가정하고, 토픽의 개수 K가 주어질 때, 각 토픽을 이루는 단어 목록을 제시해 준다. 이러한 토픽모형의 결과는 우리가 일일이 확인할 수 없는 텍스트 자료의 주제를 이해하고 요약하는 데 유용하게 활용된다. 그런데 텍스트 자료가 시간에 따라 수집된다면, 우리는 자료의 토픽 뿐 아니라 자료가 수집된 시점에 따라 각 토픽이 어떤 강도로 등장하고 있는지를 모니터링하는 데에도 많은 관심이 있다. 예를 들어, 특정 토픽이 시점에 관계없이 일정하게

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1003257).
[1] Corresponding author: Department of Statistics, Dankook University, 152, Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do 16890, Korea. E-mail: silee@dankook.ac.kr

Published 31 October 2021 / journal homepage: http://kjas.or.kr

나타나는 것인지 아니면 특정 시점에 비정상적으로 많이 등장하고 있는지 등에 따라 시점별 자료의 정보를 올바르게 해석할 수 있게 된다.

어떤 관심 있는 관측값의 크기 변화를 모니터링하는 경우에 우리는 통계적 공정관리를 위한 대표적인 도구인 관리도를 활용할 수 있는데, 이때 관리도를 통해 관측값이 정상상태인지 이상상태인지 검정할 수 있다 (Montgomery, 2020). 이에 우리는 가장 널리 활용되는 잠재적 디리슈레 할당(latent Dirichlet allocation, LDA) 모형을 사용하고, 각 토픽의 단어목록으로부터 토픽지수를 제안하고 이를 기반으로 가중이동평균(exponentially weitghted moving average, EWMA) 관리도를 활용하는 방법을 제안하고자 한다.

2. 토픽모형

토픽모형이란 문서를 생성하는 확률적 모형으로, 문서 별 단어의 관측도수로 구성된 [문서×단어]행렬로부터, 각 문서는 잠재된 K개 토픽에 대한 혼합 확률분포로 표현된다고 가정한다 (Blei와 Lafferty, 2007; Blei 등, 2003). 말뭉치에서 관측한 서로 다른 단어를 $\{1,\ldots,V\}$라고 할 때, 토픽모형은 관측된 단어를 V차원의 벡터 w로 표현하는데 $v \in \{1,\ldots,V\}$번째 단어에 대하여 w벡터의 v번째 성분은 1이고 다른 성분들은 모두 0이 된다. 문서 d는 N_d개의 단어열로 $\boldsymbol{w}_d = \{w_{d,1},\ldots,w_{d,N_d}\}$가 되고, M개의 문서를 포함하는 말뭉치는 $D = \{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_M\}$로 나타낸다. 이렇게 표현된 텍스트 데이터에 대하여 LDA 모형은 다음과 같은 문서생성 과정을 가정한다.

(1) 말뭉치에 있는 $k\,(k = 1,\cdots,K)$번째 토픽에서 각 단어가 차지하는 분포 $\boldsymbol{\beta}_k = (\beta_{k1},\cdots,\beta_{kV})$를 추출한다.

$$\boldsymbol{\beta}_k \sim \text{Dir}(\boldsymbol{\delta}). \tag{2.1}$$

(2) 문서 d에 대해서 K개의 토픽을 구성하는 분포 $\boldsymbol{\theta}_d = (\theta_{d1},\cdots,\theta_{dK})$를 추출한다.

$$\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha}).$$

(3) 문서 d에 있는 서로 다른 N_d개의 단어 중 각 단어 $w_{d,n}$, $(n = 1,\ldots,N_d)$에 대하여,

a. 어떤 토픽 $z_{d,n}$에 해당할지 결정한다.

$$z_{d,n} \sim \text{Multinomial}(1,\boldsymbol{\theta}_d).$$

b. $\mathbf{z}_{d,n}$토픽이 주어졌을 때, 토픽에 해당하는 단어 $\mathbf{w}_{d,n}$를 선택한다.

$$w_{d,n} \sim \text{Multinomial}(1,\boldsymbol{\beta}_{z_{d,n}}).$$

위 생성모형에서 관측된 변수는 오직 단어 뿐이며, $\boldsymbol{\beta}_k\,(k = 1,\ldots,K)$와 $\boldsymbol{\theta}_d\,(d = 1,\ldots,M)$는 모두 잠재적인 변수로 생각할 수 있다. 이때 디리슐레 분포에서의 $\boldsymbol{\delta}$와 $\boldsymbol{\alpha}$는 하이퍼 모수가 된다. 이들 변수와 모수를 추론하기 위해서는 식 (2.2)에 있는 말뭉치가 생성될 가능성이 높은 추정치를 선택한다.

$$P(D|\boldsymbol{\alpha},\boldsymbol{\delta}) = \prod_{d=1}^{M}\int P(\boldsymbol{\theta}_d|\boldsymbol{\alpha})\left(\prod_{n=1}^{N_d}\sum_{z_{d,n}} p(z_{d,n}|\boldsymbol{\theta}_d)p(w_{d,n}|z_{d,n},\boldsymbol{\beta})\right)\mathrm{d}\boldsymbol{\theta}_d. \tag{2.2}$$

위 확률을 통하여 토픽모형을 정확히 추론하는 것은 어렵기 때문에 근사적인 방법이 사용되는 데, 자세한 것은 Blei 등 (2003)과 Blei와 Jordan (2003) 등을 참조하기로 한다. 본 논문에서는 R의 tm 라이브러리에 있는 LDA 함수를 통하여 추정한 결과를 사용한다.

앞에서 소개한 LDA 모형은 관측된 단어만을 기반으로 해서 토픽 모형을 추정하는데, 토픽의 개수 K가 미리 결정 되어야 한다. 따라서, LDA 모형과 같은 토픽모형에서는 적절한 토픽의 수 K를 선택하는 것이 중요한 문제이다. K를 너무 작게 설정하는 경우 지나치게 광범위한 주제가 생성되고, 또 너무 크게 설정하는

경우 지나치게 세분화된 주제가 설정되기 때문이다. 적절한 K를 선택하는 방법으로 Blei 등 (2003), Arun 등 (2010), Cao 등 (2009), Griffiths와 Steyvers (2004) 등의 연구가 있다.

- Blei 등 (2003)이 제안한 퍼플렉시티(perpelxisity)는 문서 $d(=1,\ldots,M)$의 단어열 $\boldsymbol{w}_d$에 관한 가능도함수 $p(\boldsymbol{w}_d)$의 기하평균에 의해 결정되며, 가능도 함숫값이 클수록 모형이 적절하다는 것을 의미하므로, 토픽의 개수 K는 퍼플렉시티를 최소화하는 점으로 결정할 수 있다.

$$\text{perplexisity}(D) = \exp\left[-\frac{\sum_{d=1}^{M}\log p(\boldsymbol{w}_d)}{\sum_{d=1}^{M} N_d}\right].$$

- Griffiths와 Steyvers (2004)은 디리슈레 분포의 모수 α와 δ를 고정시켜두고 토픽의 개수 K를 변화시켜 가며 최적의 K를 구하는 방법을 제안하였는데, 이 방법은 먼저 깁스 표집기(Gibbs sampler)를 이용하여 추출된 토픽 $\mathbf{z}$에 대해

$$p(\boldsymbol{w}|\mathbf{z},K) = \left(\frac{\Gamma(V\delta)}{\Gamma(\delta)^V}\right)^K \prod_{j=1}^{K}\frac{\prod_w \Gamma(n_j^{(w)}+\delta)}{\Gamma(n_j^{(\cdot)}+V\delta)},$$

  값을 구하고 이 값들의 기하평균으로 가능도함수를 계산하여 가능도함수가 최대가 되게 하는 K 값을 찾는다. $n_j^{(w)}$는 단어 w가 토픽 j에 할당된 횟수를 의미한다.

- Cao 등 (2009)은 LDA를 토픽과 단어의 군집분석 과정으로 보고 군집분석에서 사용하는 것과 비슷한 K 선택절차를 제안하였다. 즉, 군집분석에서 군집들이 잘 구분되어 있는 경우 군집 간의 거리가 커지고 반대로 K가 적절하게 선택되지 않은 경우 군집들이 잘 구분되지 않는다는 점을 LDA에 적용하여 다음과 같은 K 선택 절차를 제시하였다. 즉, 군집 간의 거리를 코사인 유사도(cosine similarity),

$$corr(C_i, C_j) = \frac{\sum_{v=1}^{V} C_{iv}C_{jv}}{\sqrt{\sum_{v=1}^{V} C_{iv}^2}\times\sqrt{\sum_{v=1}^{V} C_{jv}^2}}$$

  를 측정하고, 토픽 수가 K개인 LDA 모형의 군집 간의 평균적인 유사성을

$$\frac{\sum_{i=1}^{K}\sum_{j=i+1}^{K} corr(C_i, C_j)}{K(K-1)/2},$$

  로 측정한다. 여기에서 C_i와 C_j는 각각 i와 j번째 토픽을 대표하는 단어벡터에 해당된다. 군집 간의 평균 코사인 유사도 값이 작을수록 군집들 사이의 거리가 멀고, 반대로 코사인 유사도 값이 클수록 군집들이 서로 잘 구분되지 않는다고 볼 수 있다. 따라서, 평균 코사인 유사도 값이 최소가 되도록 K를 선택한다.

- Arun 등 (2010)은 LDA를 토픽-단어행렬 M_1과 문서-토픽행렬 M_2로 분해하는 것으로 보고 대칭형 쿨백-라이블러 발산(symmetric Kullback-Leibler divergence)을 이용하여 M_1 행렬의 특잇값 분포 C_{M_1}과 말뭉치 내의 토픽 분포 C_{M_2}의 유사성을 측정하였다. 그들은 경험적으로 토픽의 수 K가 적절히 선택되고 문서의 주제들이 명확히 구분될 때 C_{M_1}과 C_{M_2}가 비슷하게 나타나 대칭형 쿨백-라이블러 발산 값,

$$\text{Proposed measure}(M_1, M_2) = KL\left(C_{M_1}\|C_{M_2}\right) + KL\left(C_{M_2}\|C_{M_1}\right)$$

  이 최소가 된다고 주장하였다. Arun 등 (2010)이 제안한 방법은 자주 사용되고 있지만, 이 측도는 저자들도 언급한 바와 같이 엄밀한 이론적 증명과는 다소 거리가 있는 것으로 알려져 있다.

본 논문의 실제 자료분석에서 이들 방법을 모두 적용하기로 한다.

3. 시간에 따른 토픽 변화 탐지

LDA 모형을 통해 전체 문서를 K개의 토픽으로 나누고, 각 토픽을 구성하는 단어를 통해 문서의 내용을 이해하고 요약하는 것이 가능하다. 그러나 이러한 분석결과는 텍스트 자료가 한 시점에서 수집된 내용인 경우에는 자료에 대한 요약으로 충분할 수 있지만, 뉴스 기사나 국민청원 자료 등과 같이 시간을 두고 수집한 내용에 대해서는 충분하지 않아 시점 간 토픽의 변화를 모니터링할 필요가 있다. 일반적으로 이전 연구에서는 관심있는 주제가 시간에 따라 어떻게 변화하는지 살펴보기 위해, 주제와 밀접한 소수의 단어만을 설정하고 시점에 따른 단어의 빈도수를 모니터링 한 것이 일반적이다. 본 연구에서는 토픽모형의 결과를 바탕으로 자료수집 시점 별로 토픽의 변화를 보기 위한 새로운 토픽지수를 제안하고, 그 지표를 통해 토픽 변화를 모니터링하는 문제를 생각해 보고자 한다.

본 연구에서는 토픽모형의 결과로부터 식 (2.1)의 추정결과 즉, 특정단어가 각 토픽에 선택될 확률을 가중치로 하여 토픽지수를 생성하고 이를 모니터링 하고자 한다. 즉, 토픽모형 분석결과 $\hat{\beta}_{k,j}$가 0이 아닌 k번째 토픽의 단어를 $\{word_{k,1}, \ldots, word_{k,n_k}\}$라 하고, 각 단어가 토픽에 사용될 확률을 $\{\hat{\beta}_{k,1}, \ldots, \hat{\beta}_{k,n_k}\}$라 할 때, 새로운 문서 d에 대한 k번째 토픽지수 T_k는 다음과 같이 제안한다.

$$\mathrm{T}_{k,d} = \frac{\sum_{i=1}^{n_k} \hat{\beta}_{k,i} I(\text{word}_{\,k,i} \in d)}{\sum_{i=1}^{n_k} \hat{\beta}_{k,i}}. \tag{3.1}$$

이때, 토픽지수 T_k는 k번째 토픽과 관련한 단어의 상대적인 출현확률에 따른 가중치 $\hat{\boldsymbol{\beta}}_k$를 고려한 것이라고 볼 수 있다.

이제 토픽 별로 식 (3.1)의 토픽지수를 관측값으로 하여 관리도를 작성하기 위해, 본 연구에서는 지수가중이동평균(exponentially weighted moving average, EWMA) 관리도 (Roberts, 1959)를 사용하기로 한다. 각 토픽지수는 각 문서마다 작성할 수 있는데, 문서 d가 시점 i에 따라 수집되므로 $T_{k,d}$는 $T_{k,i}$로 쓰고 EWMA 통계량을 다음과 같이 나타내기로 한다.

$$Z_{k,i} = \lambda T_{k,i} + (1-\lambda) Z_{k,i-1}, \quad k = 1, \ldots, K, \tag{3.2}$$

이고 $\lambda \in (0, 1]$는 상수를 나타낸다. 이때 $Z_{k,0} = \mu_{k,0}$로 이 값은 $T_{k,i}$의 분포가 정상적으로 분포할 때의 기대값 $\mathrm{E}(T_k)$를 나타낸다. 따라서, 토픽 별 각 통계량의 관리상한선(upper control limit, UCL)과 관리하한선(lower control limit, LCL)은 다음과 같다.

$$\mathrm{UCL}_{\,k} = \mu_{k,0} + L\sigma_{k,0}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)[1-(1-\lambda)^{2i}]},$$
$$\mathrm{LCL}_{\,k} = \mu_{k,0} - L\sigma_{k,0}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)[1-(1-\lambda)^{2i}]}.$$

여기서 $\sigma^2_{k,0}$은 $T_{k,i}$의 분포가 정상적으로 분포할 때의 분산 $\mathrm{Var}(T_k)$을 나타낸다. 이때 관리한계선에서 $[1-(1-\lambda)^{2i}]$의 항은 i 가 클수록 1이 되므로, 여기서는 관리한계선을 다음과 같이 간단하게 사용하기로 한다.

$$\mathrm{UCL}_{\,k} = \mu_{k,0} + L\sigma_{k,0}\sqrt{\frac{\lambda}{2-\lambda}}, \tag{3.3}$$

$$\mathrm{LCL}_{\,k} = \mu_{k,0} - L\sigma_{k,0}\sqrt{\frac{\lambda}{2-\lambda}}. \tag{3.4}$$

이때 (λ, L)은 EWMA 관리도의 관리모수로서 탐지하고자 하는 평균 변화의 크기에 따라 다르게 설정될 수 있다. 관리도에서 관리한계선을 넘을 때까지 추출한 표본수의 수(본 연구에서는 문서의 수)를 런길이(run length)

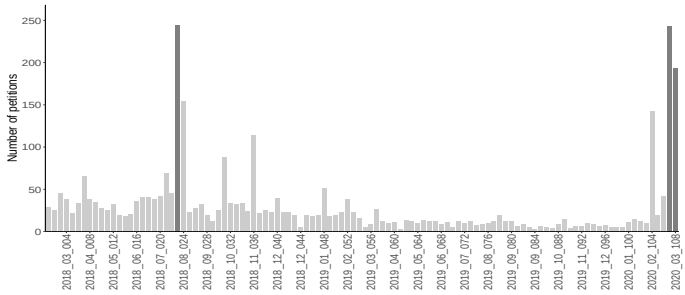

Figure 1: *The weekly number of petitions on health and welfare categories from March 5, 2018, to March 5, 2020.*

라 하는데, 이것은 확률변수로서 평균 런길이(average run length, ARL)를 통해 관리도의 성능을 평가할 수 있다. 관리상태일때 평균런길이를 ARL_0, 이상상태일때 평균런길이를 ARL_1이라고 나타내면, 관리상태일때는 ARL_0가 길수록 이상상태일때는 ARL_1이 짧을수록 좋은 관리도가 된다. EWMA 관리도에서 λ와 ARL_0가 설경되면 L이 결정될 수 있다. 이와 관련한 좀 더 자세한 사항은 Knoth (2007)을 참고하기 바란다. $\mu_{k,0}$와 $\sigma^2_{k,0}$는 m개의 과거 문서로부터 계산한 토픽지수 $T_{k,1}, \ldots, T_{k,m}$의 표본평균 $\bar{T}_{k.}$과 표본분산 $\sum_{t=1}^{m}(T_{k,t} - \bar{T}_{k.})^2/(m-1)$으로 추정한다. 이렇게 작성된 관리도를 1단계 관리도(phase I control chart)라 부르는데, 과거에 수집된 자료로부터 토픽의 상대적인 변화를 짐작해 볼 수 있는 대안을 제공할 수 있을 것이다.

4. 사례 분석

4.1. 자료수집 및 전처리 소개

본 연구에서 사용한 텍스트 자료는 2018년 3월 5일부터 2020년 3월 5일까지 국민청원 보건복지 범주에 올라온 청원을 분석 대상으로 정하였다. 파이썬의 Selenium 모듈과 Chrome 브라우저를 사용하여 국민청원 홈페이지를 크롤링(crawling)하여 총 3,197개의 청원을 수집하였고, 청원 토픽을 모니터링 하기 위해 주별 단위로 108개의 문서를 생성하였다. Figure 1은 주별로 청원 동의 수가 50명 이상인 청원의 개수를 히스토그램으로 나타낸 것이며, 이 기간 동안 평균적으로 매주 약 30개의 청원이 있었음을 알 수 있으며 표준편차는 약 41개로 크게 나타났다.

본 연구에서는 텍스트 자료를 전처리하기 위해 파이썬의 KoNLPy 패키지의 Twitter 형태소 분석기를 사용하여 명사를 추출하였고, R의 KoNLP 패키지의 bigram()함수를 사용하여 의미 있는 복합명사, 예를 들어 '희귀질환'과 '외상센터' 같은 단어를 추출하였다. 또한, '건강보험'과 '의료보험'처럼 같은 의미를 지닌 단어의 경우 하나의 단어로 통일하였고, 불용어와 같이 큰 의미가 없는 명사들을 제거하였는데, 국민 청원의 특성상 특히 정부를 지칭하는 단어 등이 많이 등장한다. 이는 토픽을 추출하는 과정과는 무관하다고 판단하여, '청원','대통령'과 같은 단어는 불용어로 처리하고 제거하였다.

보건복지 범주와 관련한 국민들의 요구사항을 토픽모형을 통해 알아보기 위하여 먼저 자료는 같은 주에 올라온 청원을 한 문서로 간주하였다. 108주 동안 수집된 문서에 대해서 말뭉치(corpus)를 생성한 후, 사전처리를 통해 '문서 단어 행렬'(document-term matrix, DTM)을 구축하였다. 즉, 문서별 단어에 대한 빈도수(term frequency, TF)를 계산하여 나타낸 2차원 행렬이다. 지금까지의 전처리를 통해 총 18,208개의 단어가 추출되었는데, 토픽모형 적용에 있어 좀 더 의미 있는 단어를 선택하기 위해 빈도수가 가장 높은 단어를 선택할 수도 있지만, 이와 같은 단순 빈도수에 기반한 접근방식은 문서의 주제를 나타내는데 있어 모든 문서에 매번 나타나는 단어, 즉 불용어에 가까운 단어들이 분석에 많은 영향을 미치게 할 수 있다. 이에 각 단어가 문

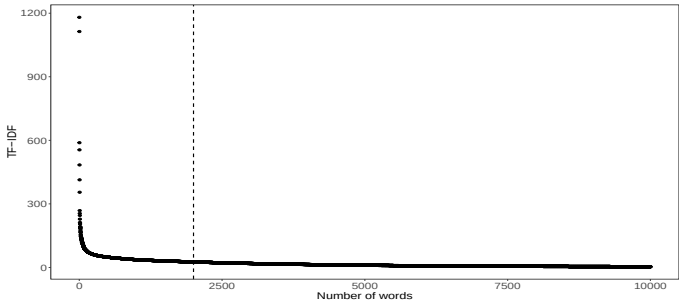

Figure 2: *The scree plot for TF-IDF with respect to the number of keywords.*

서의 중요도에 미치는 가중치를 고려한 '단어 빈도 역문서 빈도'(term frequency-inverse document frequency, TF-IDF)를 생성해 보았다. 역문서 빈도(IDF)란,

$$\text{IDF}(\text{word}, D) = \ln \frac{|D|}{1 + |\{d \in D : \text{word} \in d\}|}$$

로 특정 단어(word)가 전체 문서 $D = \{d_1, \ldots, d_{108}\}$에서 차지하는 상대적인 중요도를 계산하게 된다. $|D|$는 문서의 개수로 여기서는 108이 되고, $|\{d \in D : word \in d\}|$는 특정 단어가 등장한 문서의 개수를 나타낸다. 따라서 IDF는 모든 문서에 동일하게 등장한 단어의 경우 IDF 값이 작아지고, 오히려 문서에 드물게 등장한 단어에 대해서는 큰 IDF 값을 갖게 된다. Figure 2는 단어별로 IDF와 TF의 곱에 대해 내림차순으로 정리한 그림이다. 말뭉치에는 총 18,208개의 단어가 등장하였지만, Figure 2를 통해 많은 단어들이 상대적으로 그 중요도가 크지 않음을 알 수 있다. 본 연구에서는 총 2,000개의 단어를 주요 단어로 설정하여 [문서×단어] 행렬을 생성하고 이를 기반으로 LDA 모형을 적용하였다.

4.2. 토픽 생성

주어진 자료에 토픽 모형을 분석하기에 앞서 최적의 토픽 개수(K)를 찾기 위해 다양한 분석법을 적용해 보았다. 2절에서 언급한 것처럼 최적의 토픽 개수는 대표적으로 Blei 등 (2003), Arun 등 (2010), Cao 등 (2009), Griffiths와 Steyvers (2004) 이 제안한 것으로 Figure 3에서 각각 Perplexity, Arun2010, CaoJuan2009, Griffths2004로 나타난 결과를 사용한다. Figure 3에서 Perplexity, Arun2010, CaoJuan2009는 최솟값을 Griffth2004의 경우는 최댓값을 갖는 토픽의 갯수를 설정해야 하는데, 모두 토픽이 20개부터 30개 사이 이후에는 값의 변화가 둔감하고, 특히 이러한 패턴은 토픽의 수를 200까지 늘려서 계산한 경우도 마찬가지여서, 본 연구에서는 정성적인 판단과 함께 20개를 최적의 토픽 개수로 설정하였다. 아래 Table 1은 토픽 개수 $K = 20$으로 LDA 모형을 적용한 결과이다. 각 토픽을 구성하는 단어 중 등장 확률이 높은 순서대로 상위 10개의 단어들에 대하여 나열해 보았다. 위 결과로부터 토픽명을 지정해 보았는데, 토픽주제와 관련하여 Bang과 Moon (2019) 등에서 언급했듯이 토픽에 대한 명명은 연구자의 정성적 판단이 개입되는 부분으로, 본 연구에서는 토픽 확률이 높은 문서들에 대하여 토픽 생성확률이 높은 단어를 함께 비교해 보면서 Table 2와 같이 결정하였다.

4.3. 토픽 지수 모니터링

이 절에서는 Table 2에서 정리한 20개의 토픽 중 토픽 1, 11, 13, 그리고 19에 대해서만 모니터링 결과를 소개하기로 한다. 이 4개 토픽에 대한 $\hat{\beta}_k$ (k = 1, 11, 13, 19)를 내림차순으로 그려보면 Figure 4와 같다. 토픽 1, 11, 13, 그리고 19에 대하여 $\hat{\beta} \neq 0$인 단어 개수는 각각 1,028개, 1,462개, 1,108개, 1,242개가 된다. 108개 문서에

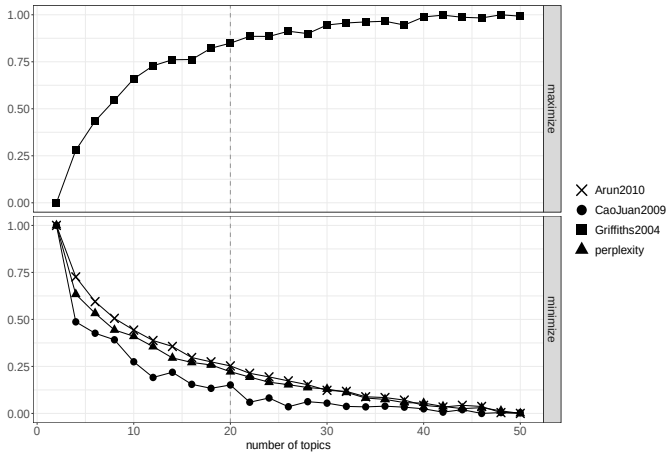

Figure 3: *Determination of the proper number of topics using several measures.*

Table 1: Top 10 words list by topics based on LDA results

Topic	Rank1	Rank 2	Rank 3	Rank4	Rank5	Rank6	Rank7	Rank8	Rank9	Rank10
Topic1	흡연	제품	물질	의료	식약처	해당	한국	일본	어르신	관리
Topic2	센터	입원	수술	의료	보험사	보험	요양원	지급	공단	노숙인
Topic3	외국인	수술	의사	건강보험	의료	혜택	아이	생명	가입	처벌
Topic4	세금	택배	아파트	다산	신도시	실버	해결	아이	의사	통증
Topic5	의사	수술	교수	증상	남편	아이	입원	확인	진행	간호사
Topic6	국민연금	연금	공무원	세금	폐지	군인	공무원연금	가입	개혁	수령
Topic7	아이	의사	장애	의료	검사	보험	사망	시술	수술	적용
Topic8	아이	접종	아기	비소	부모	검사	백신	헬기	의료	제발
Topic9	의료	아이	건강보험	외국인	장애	민원	운영	영리병원	적십자사	업체
Topic10	장애	아이	수술	아동	엄마	유공자	의사	보호	시설	요양원
Topic11	마스크	코로나	판매	확진자	감염	신천지	지역	격리	사태	대구
Topic12	간호사	간호조무사	장애	활동	간호	보조	서비스	업무	복지	여성
Topic13	누진세	아이	에어컨	폐지	폭염	사용	전기	기세	가정	의료보험
Topic14	아이	의사	수술	의료	사고	생명	시력	안과	의학	방법
Topic15	장애	검사	의료	복지	발달장애	입원	상담	의사	입양	건강보험
Topic16	치과	의사	여성	의료사고	진료	교정	피해자	의료	아이	수술
Topic17	마스크	코로나	신천지	확진자	대구	감염	지역	확산	중국	바이러스
Topic18	교사	아이	어린이집	장애	보육	보육교사	아동	휴게시간	원장	외국인
Topic19	중국	마스크	우한	입국	바이러스	폐렴	감염	아이	중국인	코로나
Topic20	아이	간호사	의료	의사	수술	진료	피해	선생님	기증	폭행

대하여 식 (3.1)의 토픽지수 T_k를 계산하고, 식 (3.2)을 사용해서 EWMA 통계량 $z_{k,i}, i = 1, \ldots, 108$을 생성한다. 이때, $\mu_{k,0}$와 $\sigma_{k,0}$는 관리상태 하에서 토픽지수의 평균과 표준편차를 의미하는데, 108개 문서에 대한 토픽지수 $T_{k,1}, \ldots, T_{k,108}$로부터

$$\hat{\mu}_{k,0} = \frac{\sum_{j=1}^{108} T_{k,d}}{108}, \quad \hat{\sigma}_{k,0} = \sqrt{\sum_{j=1}^{108} \frac{(T_{k,d} - \hat{\mu}_{k,0})^2}{107}},$$

와 같이 추정한다. 이러한 관리도는 1단계 관리도에 해당하며, 과거 시점 동안의 상대적인 토픽 변화를 모니터링할 수 있다. 식 (3.3)과 식 (3.4)로부터 토픽별 관리한계선을 추정하였다. 이때, λ는 일반적으로 $0.05 \leq$

Table 2: The representative sentence for each topic from LDA results

	Topics
Topic1	의료목적인 제품 사용에 대한 식약처의 허용 요청 관련 청원
Topic2	의료관련 센터와 보험사에서의 보험금 처리 방식
Topic3	외국인에 대한 의료혜택 제공
Topic4	다산신도시 아파트 단지의 실버 택배 제도에 대한 세금 지원
Topic5	가족의 수술과 입원 과정에서의 의사의 대처에 대한 청원
Topic6	국민연금, 공무원연금, 군인연금에 대한 개혁 및 폐지
Topic7	장애아동, 한부모가정 아동을 대상으로 한 의료관련 청원
Topic8	아기들의 백신 접종 과정에서 비소 검출
Topic9	외국인을 대상으로 운영하는 영리병원 설립
Topic10	장애아동 및 유공자를 위한 보호 시설 설립 요구
Topic11	코로나 마스크 판매 및 공급 체계 개선
Topic12	간호사와 간호조무사의 업무분장
Topic13	여름철 폭염으로 인한 가정 누진세 폐지 청원
Topic14	안과 의료사고와 관련한 보상 방법
Topic15	발달장애인의 의료 복지 관련 청원
Topic16	교정 치과 의료사고 피해자들의 보상 문제
Topic17	대구 신천지 관련 코로나 확산 문제
Topic18	어린이집 보육교사의 휴게시간과 관련한 문제
Topic19	우한 발 코로나 바이러스 감염에 따른 중국인 입출국 금지 요청
Topic20	인력 부족으로 의료진 사이에서 발생하는 업무에 대한 피해와 폭행

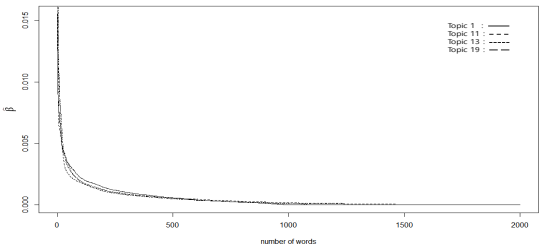

Figure 4: $\hat{\beta}$ *for the words in Topic 1, 11, 13, and 19.*

$\lambda \leq 0.25$ 사이에서 EWMA 관리도 성능이 효율적이라고 알려져 있어 (Lucas와 Saccucci ,1990), 관리한계선은 $\lambda = 0.2$로 정하고 관리상태 하에서의 평균 런길이 $ARL_0 = 370$로 하여 $L = 2.86$으로 설정하였다. Figure 5는 토픽 1과 토픽 13에 대한 것인데, 자료가 수집된 전체 기간 동안 각 주제의 출현 강도가 높은 구간이 자료 수집 중간에 나타난 경우이다. 토픽 1은 '의료목적인 제품 사용에 대한 식약처의 허용 요청' 관련 주제인데, 실제로 2018년도 10월과 11월에 대마오일에 대한 치료제 허용 요구에 대한 청원이 많았음을 확인할 수 있다. 이후 관련 청원이 크게 보이지 않고 오히려 평소보다 관련 토픽이 적게 나타났는데, 2020년 2월부터 대마오일 해외직구 허용 청원이 등장하며 토픽지수가 다시 증가한 것을 알 수 있다. 토픽 13은 '여름철 폭염으로 인한 가정 누진세 폐지 청원' 으로 2018년 7월에서 8월에 관련 토픽지수가 전체기간의 토픽지수 값에 비해 크게 나타나 이상상태로 판정되었는데, 2018년 여름은 우리나라 기상관측 이래 최악의 폭염이 기록된 시기로서, 당시 전기 누진세로 인한 청원이 다른 때보다 많았던 것을 짐작할 수 있다. 또한 같은 해 가을에는 난방비 누진

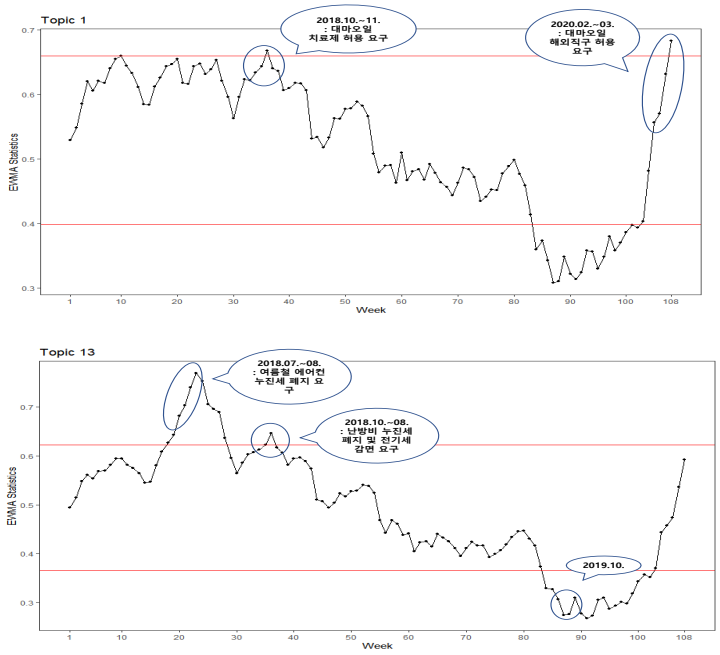

Figure 5: *EWMA chart for topic indices for Topic 1 and 13. The solid horizontal lines stand for control limits for EWMA chart.*

세 폐지에 관련 청원이 늘어나면서 토픽지수가 크게 높았음을 알 수 있다. 이후 토픽지수는 계속해서 낮아져 이 토픽과 관련한 청원이 많지 않았음을 알 수 있다. 특히 2019년 여름에도 누진세 관련한 이 토픽지수는 높지 않고 오히려 감소추세에 있는 것으로 나타났는데, 2019년도 여름에는 심한 더위가 기록되지 않았다는 점을 고려할 때 토픽지수가 잘 작동한다고 판단할 수 있다. 이후 이 토픽과 관련한 단어의 출현 지수가 커지고 있기는 하지만 관리한계선 내에 있기 때문에 정상상태로 판단한다. Figure 6의 토픽 11과 19는 모두 '코로나 19'와 관련된 토픽으로 각각 마스크와 외국인 입국 금지와 관련된 내용이다. 따라서 두 토픽은 모두 2020년 12월 말부터 지수가 급격히 증가하여 관리한계선을 넘는 것을 알 수 있는데, 흥미롭게도 두 토픽 모두 2018년에도 관리한계선을 넘은 주가 확인되었다. 문서를 살펴본 결과 토픽 11의 경우는 2018년 10월에도 미세먼지 대책 마련과 마스크 정기지급에 대한 요구가 있었으며, 토픽 19와 관련해서는 같은 시기에 중국발 미세먼지 대책 요구가 많던 시기로 각 토픽의 주요 단어가 그 시기 문서에 자주 등장했음을 확인하였다. 이러한 결과로부터 본 논문에서 제안한 토픽지수를 모니터링함으로써 시점별로 각 토픽에 대한 관심 정도를 모니터링하는 것이 가능함을 알 수 있다.

5. 결론

본 논문에서는 시점에 따른 토픽에 대한 변화양상을 모니터링하기 위하여, LDA 모형 분석 결과 각 토픽을 이루는 단어 목록으로부터 토픽지수를 생성하고, EWMA 관리도를 적용해 보았다. 과거에 수집된 전체 자료로부터 시점에 관계없이 토픽을 추출하고, 시간에 따른 토픽의 변화를 모니터링하기 위해 토픽지수를 제안하고, EWMA관리도를 통해 토픽지수의 이상상태 발생 유무를 모니터링하였다. 이는 1단계 모니터링(phase I monitoring)에 해당한다고 볼 수 있다. 이처럼 관리도를 적용함으로써 시간의 흐름에 따른 각 토픽의 변화를

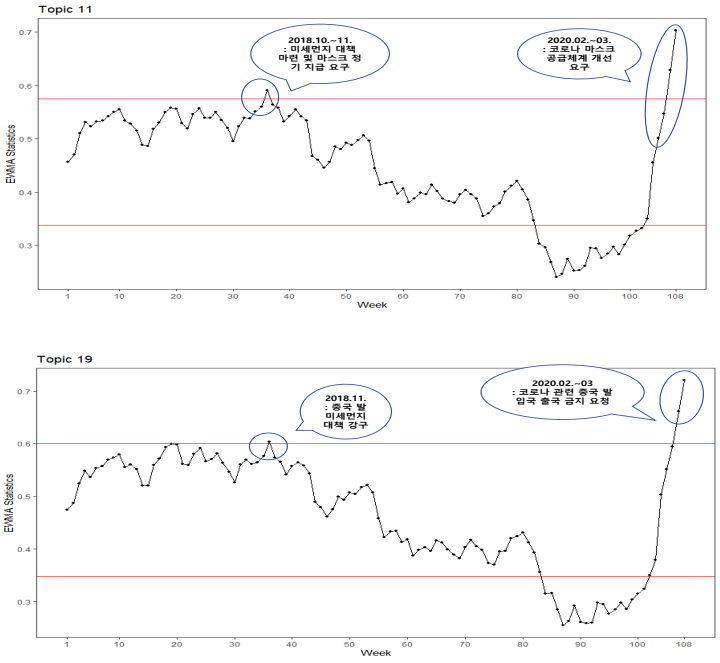

Figure 6: *EWMA chart for topic indices for Topic 11 and 19. The solid horizontal lines stand for control limits for EWMA chart.*

볼 수 있었으며, 그 변화가 실제 청원의 주제를 잘 반영할 수 있음을 살펴보았다. 실제로 이러한 모니터링은 텍스트 자료의 정보를 요약하고 그 변화를 정량적으로 이해 하는 데 사용될 수 있다고 판단된다.

또한, 이러한 토픽지수는 새롭게 등장하는 청원에도 적용될 수 있으며, 이를 통하여 관련 토픽들의 관심 정도를 실시간으로 파악할 수 있게 된다. 다만, 새로운 청원이 과거에 언급되지 않았던 새로운 청원이라면, 예를 들어, 갑작스러운 코로나 19 펜데믹과 같은 상황으로 과거에 없던 새로운 청원이 발생하는 경우라면, 과거 데이터로만 분석한 토픽은 그러한 내용을 담을 수 없음으로 새롭게 등장하는 토픽 또는 토픽의 수를 파악하기 위해 주기적인 LDA 모형 적합이 필요할 것으로 예상되며 추후 관련 연구가 필요할 것이다.

6. 참고문헌

References

Arun R, Suresh V, Madhavan CEV, and Murthy MN (2010). On finding the natural number of topics with latent dirichlet allocation: Some observation, *Pacific-Asia conference on Knowledge Discovery and Data Mining, Par I, LNAI (6118)*, 391–402.

Bang H and Moon H (2019). A study on the methodology to express the main topics of text in time series using text mining, *Journal of the Korean data and information science socieity*, **30**, 1259–1276.

Blei DM, Jordan MI (2003). Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 127–134.

Blei DM and Lafferty JD (2007). A correlated topic model of science, *The Annals of Applied Statistics*, **1**, 17–45.

Blei DM, Ng AY, and Jordan MI (2003). Latent Dirichlet allocation, *Journal of Machine Learning Research*, **3**, 993–1022.

Cao J, Xia T, Li J, and Zhang Y, and Tang S (2009). A density-based method for adaptive LDA model selection, *Neurocomputing*, **72**, 1775–1781.

Deveaud R, SanJuan E, and Bellot P (2014). Accurate and effective latent concept modeling for ad hoc information retrieval, *Document Numérique*, **17**, 61–84.

Griffiths TL and Steyvers M (2004). Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 5228–5235.

Knoth S (2007). Accurate ARL calculation for EWMA control charts monitoring simultaneously normal mean and variance, *Sequential Analysis*, **26**, 151–264.

Lucas JM and Saccucci MS (1990). Exponentially weighted moving average control schemes: properties and enhancement, *Technometrics*, **32**, 1–12.

Montgomery DG (2000). *Introduction to Statistical Quality Control*, John Wiley & Sons, New York.

Roberts SW (1959). Control chart tests based on geometric moving averages, *Technometrics*, **41**, 97–101.

Received April 30, 2021; Revised June 16, 2021; Accepted July 16, 2021

관리도를 활용한 국민청원 토픽 모니터링 연구

이희연[a], 최지은[b], 이성임[1,b], 손원[b]

[a]가톨릭중앙의료원 임상연구지원팀 통계지원실 [b]단국대학교 정보통계학과

요 약

최근 온라인 채널을 통한 텍스트 자료가 방대해 지면서 이를 요약하고 분석하는 연구에 관한 관심이 커지고 있는 추세이다. 먼저 텍스트 자료에 대한 기본적인 분석 중 하나는 어떤 주제나 내용을 포함하고 있는지 잠재된 토픽을 추출하는 것이다. 연구자가 일일이 모든 자료를 읽고 내용을 요약할 수도 있겠지만, 대용량 데이터를 다루는 경우에는 결코 쉽지 않기 때문에, 통계적 모형을 사용하여 토픽을 추출하는 토픽모형 방법들이 제안되어 왔다 (Blei와 Lafferty, 2007; Blei 등, 2003). 시간에 따라 수집된 텍스트 데이터로부터 토픽의 변화를 모니터링하기 위하여, 본 연구에서는 잠재적 디리슈레 할당(latent Dirichlet allocation) 모형을 통해 토픽을 분류하고 그 결과를 바탕으로 한 토픽 지수를 제안하였다. 또한, 이를 통계적 공정관리의 대표적 도구인 관리도에 적용하여 시간 경과에 따른 토픽의 변화를 모니터링하는 데 적용해 보았다. 실제 데이터로 2018년 3월 5일부터 2020년 3월 5일 사이에 청와대 국민청원 온라인 게시판에 접수된 텍스트 데이터를 사용하였으며, 토픽 지수를 모니터링함으로써 토픽에 대한 이상변화를 탐지할 수 있음을 살펴 보았다.

주요용어: 텍스트 데이터, LDA 모형, 토픽 모니터링, EWMA 관리도

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2019R1A2C1003257).

[1]교신저자: (16890) 경기도 용인시 수지구 죽전로 152, 단국대학교 정보통계학과. E-mail: silee@dankook.ac.kr