K-평균 클러스터링을 이용한 국민청원 내용과 동의 수 분석

Analysis of National petition contents and number of consent using K-Means Clustering

저자 (Authors)
이다인, 김유섭
Da-In Lee, Yu-Seop Kim

출처 (Source)
한국정보과학회 학술발표논문집 , 2020.7, 1429-1431 (3 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09874801

APA Style
이다인, 김유섭 (2020). K-평균 클러스터링을 이용한 국민청원 내용과 동의 수 분석. 한국정보과학회 학술발표논문집 , 1429-1431.

K-평균 클러스터링을 이용한 국민청원 내용과 동의 수 분석

이다인[O1] 김유섭[2]
[1]한림대학교 융합소프트웨어학과
[2]한림대학교 소프트웨어융합대학
iigghh1004@naver.com, yskim01@hallym.ac.kr

Analysis of National petition contents and number of consent using K-Means Clustering

Da-In Lee[O1] Yu-Seop Kim[2]
[1]Convergence Software, Hallym University
[2]College of Software, Hallym University

요 약

본 논문에서는 K-평균 클러스터링을 활용하여 청원내용과 동의 수의 분포를 분석한다. 청원을 crawling 한 후, 각 청원을 twitter 형태소 분석기 (okt)로 토큰화 (tokenize) 한다. 그리고 scikit-learn의 TF-IDF를 활용하여 특성 단어의 빈도수에 따라 벡터화하고 문서에 대한 중요도를 수치화한다. 또한, 가중치에 따라 정렬된 내용의 특징 벡터를 PCA, T-SNE 알고리즘을 이용해 차원 축소하여 시각화하고 K-평균 연산으로 나눠진 cluster 별 데이터들의 키워드를 추출한다. 마지막으로 K-평균 클러스터링 그래프를 통해 cluster 별 동의 수의 평균, 표준편차, 최댓값, 최솟값을 구했으며, 이를 통해 청원내용과 동의 수 분포를 시각적으로 표현함으로써 국민청원의 실태를 살펴본다.

1. 서 론

[1]지난 2017년 8월 17일, 정부는 국민과 소통하는 정책의 하나로 청와대 국민청원 사이트운영을 개시하였다. 접근성도 좋고 개선하길 원하는 사회적인 문제들을 자유롭게 올릴 수 있는 장점이 있었지만, 낮은 청원등록 기준과 동의 댓글의 익명성 때문에 중복 주제의 게시물이 만연하고 여론조작이 가능했다. 이에 2019년 3월 31일 이후로 100명의 사전동의를 받아야 공개되도록 하며 중복 주제의 게시물에 대한 과도한 업로드를 방지하도록 개편되었다. 그러나, 청원에 대한 찬성과 반대가 따로 없어 국민의 의견을 표출하는 데에 불편함이 있었다. 최종적으로 30일 동안 20만 이상의 동의 수를 얻으면 정부의 답변을 받을 수 있지만, 법 제정과 같은 뚜렷한 해결방안을 받을 가능성은 또한 희박하다.

본 논문에서는 이와 같은 국민청원의 사이트에 임의로 나누어져 있는 17개의 분야 (정치개혁, 외교/통일/국방, 일자리, 미래, 성장동력, 농산어촌, 보건복지, 육아/교육, 안전/환경, 저출산/고령화대책, 행정, 반려동물, 교통/건축/국토, 경제민주화, 인권/성평등, 문화/예술/체육/인문, 기타)를 모두 web crawling 한 데이터를 이용하여 청원내용과 동의 수를 분석한다.

우선, 게시물의 내용을 숫자 형태로 벡터화하기 위해 한국어 형태소 분석기 (okt)를 통하여 전 처리를 한다. 그 후, TF-IDF를 통해 단어의 빈도수와 문서에서의 중요도에 따라 가중치를 처리하여 단어들을 벡터로 만들어주어 내용을 정리한다. 그리고 가중치에 따라 정렬된 내용의 특징 벡터를 PCA, T-SNE 알고리즘을 이용해 차원 축소하여 시각화하고 K-Means 연산으로 나눠진 cluster 별 데이터들의 키워드를 추출한다.

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음 (20180002160031001).

마지막으로, 청원내용과 동의 수의 분포를 K-평균 클러스터링으로 분석하여 cluster 별로 데이터들의 동의 수에 대한 평균, 표준편차, 최댓값, 최솟값을 구한다. 이처럼 국민청원 데이터를 시각적으로 표현하고 분석함으로써, 그때 당시의 이슈들을 다시 파악할 수 있고, 한층 더 적극적으로 우리나라에 관심을 가지는 계기가 된다.

2. 관련 연구

[1]에서는 K-평균 클러스터링에서의 초기 중심선정 방법을 제안하여 비교실험을 통해 할당-재계산 횟수를 줄이고 전체 clustering 시간을 줄이고자 한다. 삼각형의 높이를 이용하는 방법을 제안하여 결론적으로 소요시간은 38.4% 감소하였다. 초기 클러스터 중심을 임의로 선정하게 되면 그 결과의 편차가 심하므로 이와 같은 연구의 방법은 유용하다.

국민청원의 게시물을 다양한 방법을 통하여 분석해본 연구들도 있다 [2, 3]. 먼저, [2]는 구조적 토픽모형 (Structural Topic Model)을 활용하여 개별 문서들을 특정 주제로 분류하고, Word2Vec을 통해 문서들에 감정 점수를 매긴 다음 음이항 회귀 분석 (Negative Binomial Regression)을 통해 게시물의 주제와 감정이 동의 수에 미치는 통계적 효과를 추정한다. 다음으로 [3]은 딥러닝 기반 LSTM을 활용하여 20만의 청원 동의를 얻는 청원을 예측하기 위한 모델을 제안하는 연구이다. 그 결과, 본문과 함께 품사의 비율을 변수로 추가한 모델의 f1-score가 0.9 이상으로 높은 예측률이 나왔다.

앞선 연구들을 참고하여 데이터를 어떻게 수집할지부터, 청원 게시물을 분석하는 효과적인 방법을 탐구한다. 본 연구에서는 국민청원 청원내용의 특징 벡터를 PCA, T-SNE 알고리즘을 통하여 차원축소하고 그 분포를 시각화한다. 그리고 K-Means 연산으로 내용 주제 단어를 추출한다. 또한, 청원내용 벡터와 동의 수를 K-평균 클러스터링을 이용해 분석한다.

3. 실험 방법

3.1 데이터수집 및 전처리

국민청원 데이터는 개시일 2017.08.17.부터 2019.02.04.까지의 웹 크롤링 (web crawling) 된 데이터를 모았다 [4]. 초기에 설정한 청원의 수는 395,547개이지만, 비정상적인 내용으로 동의 수가 10개도 되지 않거나 중복된 형식의 전처리할 수 없는 내용이 있어 동의 수 100개 이상 받은 정상적인 청원들을 기준으로 데이터를 정리하였다.

다음으로 청원내용 데이터에 대해 Konlpy [5]의 okt 함수를 사용해 형태소 분리, 즉 각각의 문장에 대하여 토큰화 (tokenize) 한다. 또한, 특수문자를 제외하기 위해 정규 표현 식으로 text를 처리하고 자주 나오거나 불필요한 불용어인 '있습니다', '합니다', '에서', '위한' 등을 따로 삭제한다. 이러한 전처리 방법을 모든 청원내용에 적용했다.

3.2 TF-IDF의 벡터화

전처리 된 청원내용 데이터를 scikit-learn [6]의 Term Frequency-Inverse Document Frequency (TF-IDF)를 사용하여 문서에서 특정 단어가 얼마나 중요한 의미를 담고 있는지 수치화한다. TF-IDF란, 문서에서의 중요도가 높은 단어의 가중치를 크게, 빈도수만 많고 중요도는 낮은 단어의 가중치는 작게 처리하는 방법이다. 대부분의 문서 집합에서 일반적으로 사용하며 흔한 키워드는 없애고 더욱 구체적인 키워드를 추출하기 위해 가중치를 조절해주는 역할을 하는 것이다.

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \qquad (1)$$

이 수식 (1)에 따라 해당 문서에서 특정 단어가 나온 횟수 값과 그리고 이 단어가 다른 문서에는 잘 나오지 않는 드문 단어인지를 같이 고려한다. 이 값의 계산은 다음과 같이 두 값을 곱해준다. 이를 통해 각 단어의 빈도수뿐만 아니라 그 단어가 해당 문서에서 중요도가 높은 단어인지를 판단한다.

3.3 k-means clustering algorithm

clustering 이란 데이터에 대하여 특정 값에 따라 여러 개의 cluster로 분류하는 기법이다. 분할 clustering 중 k-mode 또는 k-median과 같은 종류의 알고리즘도 있지만, 본 연구는 비교적 일반적으로 사용하는 K-평균 클러스터링을 활용하여 구현한다. 이 기법은 cluster 내의 중심값이 평균값이 된다.

K-평균 클러스터링은 k개의 cluster 수 만큼 데이터를 분리하는 비지도 학습에 속한다. 알고리즘의 단계에는 세 단계가 있는데, 첫 번째는 초기 데이터를 무작위로 선택하거나, 데이터 정렬 후 k 부분으로 나누어 중심점을 선택하는 것이다. 그리고 두 번째 단계는 데이터 포인트와 모든 중심 사이의 거리를 계산하여 가장 가까운 중심을 가진 cluster에 지정한다. 이때, cluster를 할당할 때마다 L2 거리 계산법 (2)을 수행한다.

$$distance = (x2-x1)^2 + (y2-y1)^2 + (z2-z1)^2 \qquad (2)$$

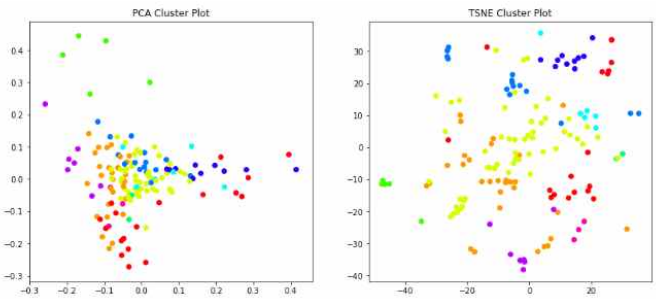
다음으로 각 데이터를 가장 가까운 중심에 할당하고, 업데이트된 cluster 중심 값은 해당 cluster의 모든 데이터 포인트의 평균값이다. 이에 정지 기준이 충족될 때까지 특정 cluster에 할당된 포인트와 중심을 같게 유지한 상태로 마지막 두 단계를 반복하여 준다. 정지 기준은 기존 중심과 새로운 중심의 차이 값이 임계값 보다 작을 때, 즉 cluster의 중심이 크게 움직이지 않을 때까지 할당하는 것이다.

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (\|x_i - \mu_j\|^2) \qquad (3)$$

이 수식 (3)을 보면 sample X를 한 세트 N개의 K개로 분리된 cluster C로 나누고, μ_j은 각 cluster sample X의 평균이며 일반적으로 클러스터 '중심점'을 나타낸다. 이처럼 k-means clustering algorithm은 관성 또는 cluster 내 제곱 기준을 최소화하여 중심을 선택하는 것을 목표로 한다.

3.4 청원내용 특징분포, 시각화

본격적으로 청원내용과 동의 수를 clustering으로 적용하기 전에 TF-IDF로 벡터화할 때 설정한 2,000개의 max feature 값을 차원축소 하여 시각화해 보았다. 특징값이 많으면 기계학습 모델이 잘 훈련되지 않거나 과적합을 일으키기 쉽고 훈련된 모델을 해석하여 유용한 정보를 얻기가 힘들다. 또한, 고차원 데이터는 시각화가 어려워 분석을 수행하기도 쉽지 않다.



[그림 1] PCA, T-SNE Cluster 시각화 결과

그러므로 높은 차원의 특징을 저차원의 특징으로 추출하여 시각화하는 방법을 사용한다. 우선, scikit-learn의 k-means 연산으로 설정된 청원내용에 대한 cluster를 사용하며, 주요 구성요소를 분석하는 PCA [7]와 분산 확률론적으로 고차원 데이터를 임베딩 하는데 적합한 T-SNE [8]을 활용한다. 이 둘은 탐색적 데이터 분석과 집단 간의 거리와 관련성을 시각화하는 것에 유리하다. 프로세스의 속도를 높이기 위해 T-SNE 알고리즘으로 1000개의 특징을 샘플링하고, PCA로 기존의 데이터를 유지한 채 150개까지 차원 축소하여 청원내용 특징 벡터를 산점도 그래프로 표현하며, 그 결과는 [그림 1]과 같다. 또한, k-means cluster 10개를 설정하여 cluster 별로 모든 차원에서의 평균값을 계산하고 정렬한 후, 빈도가 높은 상위 단어들을 10개씩 추출하였다.

[표 1] 각 cluster의 상위 키워드 추출

cluster (n)	상위 키워드
0	살인, 범죄, 사람, 폭행, 경찰, 가해자, 피해자, 사건, 처벌, 여성
1	이재명, 정부, 수사, 판사, 나라, 의원, 대한민국, 청원, 대통령, 국민
2	유럽, 무슬림, 수용, 예멘, 나라, 제주도, 외국인, 이슬람, 국민, 난민
3	협회, 축구, 스포츠, 감독, 경기, 올림픽, 연맹, 선수, 국민, 박탈
4	건강, 아이, 간호사, 의사, 수술, 의료, 보험,

	환자, 치료, 병원
5	세금, 사업, 건설, 신도시, 주택, 분양, 택배, 지역, 주민, 아파트
6	교육청, 대학, 수업, 학부모, 선생님, 교사, 교육, 아이, 학생, 학교
7	정치, 시급, 월급, 세금, 연봉, 국회의원, 인상, 의원, 국민, 국회
8	직업, 경영, 그룹, 대한항공, 기업, 갑질, 일가, 대한민국, 국적, 국민
9	어린이집, 선생님, 교사, 원장, 보육, 유치원, 휴게, 시간, 교육, 아이

위의 [표 1]은 청원내용을 TF-IDF 벡터화를 통해 가중치가 높은 10개의 단어를 추출한 것이다. 문서에서의 중요도를 파악할 수 있었기 때문에 cluster 별로 키워드가 잘 분류된다. 그러나 청원의 특성상 '국민', '아이', '나라' 와 같은 자주 나오는 단어들은 중복된다. 초기 cluster의 개수는 6개부터 점차 하나씩 높여가며 수행하였고, 그 결과 cluster 10개로 설정하는 것이 비교적 성능이 높았다.

4. 실험 결과

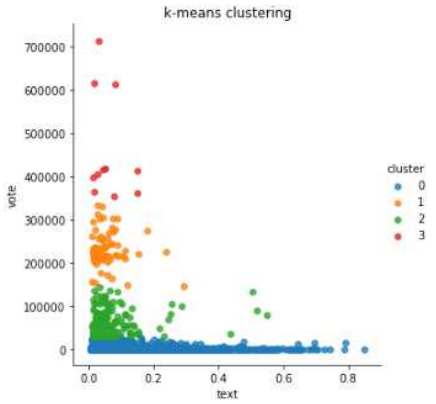
마지막으로, K-평균 클러스터링을 이용하여 청원내용과 투표수의 분포를 실험하였으며, 그 결과는 [그림 2]와 같다. 청원내용에 있어서 단어들의 의미가 2차원 벡터로 표현되었기 때문에, 투표수와의 분포를 표현하기 위해 차원의 수를 1차원으로 맞춰 주었다. cluster의 수는 10개부터 하나씩 줄여가며 실행한 결과 4개로 설정한 그래프가 가장 안정적으로 분류되었다.

우선, cluster 0의 분포를 보면 다양한 topic들의 내용인 만큼 투표수가 고르게 퍼져있다. 이 cluster의 투표수 평균은 반올림하여 1041개, 표준편차는 2791개, 최솟값은 101개, 최댓값은 2만8780개이다. cluster에 속한 청원내용은 넓게 퍼져있어서 거의 모든 분야의 게시물이 포함된다.

cluster 1의 투표수 분포는 비교적 잘 모여있고 평균은 반올림하여 23만0687개, 표준편차는 3만9797개, 최솟값은 14만6068개, 최댓값은 33만4173개이다. 청원내용은 정치개혁, 보건복지 분야의 빈도수가 많다.

cluster 2의 투표수 분포는 cluster 1보다 적은 투표수를 가진 cluster이며 반올림하여 평균 5만7459개, 표준편차 2만7441개, 최솟값 2만9612개, 최댓값 14만2715개로 분포되어 있다. 청원내용의 분야는 cluster 0과 같이 고른 편이며 육아/교육과 문화/예술/체육/언론 등이 있다.

cluster 3의 투표수 분포는 가장 높은 수치를 보이는 청원 게시물이며 평균 46만1670개, 표준편차 12만4563개, 최솟값 35만4935개, 최댓값 71만4875개로 이루어진다. 청원내용은 인권/성평등과 외교/통일/국방 분야가 많다.



[그림 2] 청원내용 (text)과 투표 수 (vote)

5. 결론 및 향후 연구계획

본 논문은 이 실험을 통하여 국민청원의 내용의 주제와 투표수에 대하여 분석했다. 먼저, 주제를 분석하기 위해서 형태소 분리한 후, TF-IDF를 사용하여 문서에서의 중요도에 따라 가중치를 주어 벡터화한다. 그리고 주제가 되는 중요한 단어들을 cluster 화하여 비슷한 내용의 cluster에 따라 중요한 단어를 추출했다. 또한, 청원내용의 특징 벡터를 PCA, T-SNE 알고리즘을 통해 데이터는 유지한 채로 차원 축소하여 간편하게 시각화하였다. 마지막으로, 청원내용과 동의 수의 분포를 K-평균 클러스터링 그래프를 활용하여 cluster 별로 동의 수에 대한 평균, 표준편차, 최댓값, 최솟값을 구하여 분포를 분석하였다.

국민청원의 내용과 동의 수에 대하여 시각화, 내용 추출 등의 실험을 하여 우리나라의 당시 이슈들을 파악할 수 있었다. 이처럼 우리나라 사회문제에 대해 분석해볼 수 있는 계기가 되었지만, 다수의 의견을 표출할 수 있는 국민청원 사이트가 좀 더 실용성 있고 효과적으로 발전되었으면 좋겠다는 생각이 들었다. 향후 더 많은 최신 데이터들을 모아 Topic Model을 활용하여 분석하고 기계학습을 통해 청원내용과 분야를 학습시켜 예측하는 모델을 만들어 보고자 한다.

참고문헌

[1] Lee, Shinwon. "Comparison of Initial Seeds Methods for K-Means Clustering." *Journal of Internet Computing and Services* 13.6 (2012): 1-8.

[2] Junmo Song, and Youngdeuk Park, What happens in the Blue House Online Petition? : An Analysis of the Blue House Online Petition Based on Natural Language Processing, *한국정치학회보*, 53.5 (2019): 61-77

[3] Woo Yun Hui, and Hyon Hee Kim, Topic Analysis of the National Petition Site and Prediction of Answerable Petitions Based on Deep Learning, *정보처리학회논문지. 소프트웨어 및 데이터 공학*, 9.2 (2020): 45-52

[4] Data https://github.com/akngs/petitions

[5] Konlpy http://konlpy.org/ko/latest/

[6] Scikit-learn https://scikit-learn.org/stable/

[7] Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." *Chemometrics and intelligent laboratory systems* 2.1-3 (1987): 37-52.

[8] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9.Nov (2008): 2579-2605.