K-means 클러스터링과 토픽 모델링을 기반으로 한 국민청원 사이트의 카테고리 재구성

우윤희*, 김현희*
*동덕여자대학교 정보통계학과
e-mail : 1215yhui@gamil.com,
heekim@dongduk.ac.kr

Reconstruction of Categories on the National Petition Site Using K-Means clustering and Topic Modeling

Yun Hui Woo*, Hyon Hee Kim*
*Dept. of Statistics and Information Science, Dongduk Women's University

요 약

국민 청원 사이트가 뛰어난 접근성과 신속성으로 인하여 국민들로부터 많은 관심을 받고 있다. 현재 국민청원 사이트의 카테고리 분류는 '미래', '성장동력' 등을 포함한 16개의 카테고리 및 기타로 구성되어 있으나 그 기준이 모호하여 많은 청원글들이 기타 카테고리로 분류되고 있는 상황이다. 이는 청원글의 내용을 명확히 반영하지 않고 미리 정의된 카테고리 구조를 사용하고 있는데서 기인한다고 할 수 있다. 본 논문에서는 보다 구체적으로 정의된 카테고리를 정의하고자 추천 순으로 1,500개의 청원글을 수집하였고, 수집된 청원글의 내용을 바탕으로 카테고리 구조를 추출하였다. 먼저, k-평균 알고리즘을 적용하여 청원글을 군집하여 대분류를 정의하였고, 보다 구체적인 세부 분류를 정의하기 위하여 토픽모델링을 실시하였다. 본 논문에서 제시하는 계층적 카테고리 구조는 청원글의 내용을 바탕으로 대분류와 세부분류로 구성된 것이므로 새로운 청원글을 등록하거나 분류하는데 적절한 것으로 보인다.

1. 서론

청와대 국민청원은 2017년 8월 17일 현 정부 출범 100일을 맞이하여 신설되었으며, 30일 동안 20만 명 이상의 국민들이 추천한 청원에 대해 정부 및 청와대 관계자가 응답하는 시스템이다. 국민청원은 음주운전 처벌 강화 '윤창호법', 심신미약 감경 의무 폐지 '김성수법' 등 국민청원으로 필요한 법을 재정하는 순기능을 가져다주기도 했으나, 무분별한 글 게시와 중복되는 글, 비방의 글도 여과 없이 올라오며 개편 필요성이 대두되었다. 이에 국민청원은 3월 31일부로 100명의 사전 동의를 받아야 청원글이 공개되도록, 폭력적, 선정적, 집단 혐오적 표현과 명예훼손 내용이 들어있는 청원글들은 삭제되도록 개편되었다.

추천 수에 따라 정부의 답변여부가 결정된다는 점에서 국민청원 사이트의 글 정렬, 분류 방식, 카테고리 설정은 매우 중요한 역할을 할 것이다. 본 논문은 그러나 개편된 사이트가 여전히 청원글의 카테고리 면에서 문제점을 가지고 있다는 부분에 주목한다. 국민청원의 카테고리는 '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/언론' 16가지와 기타로 구성된다. 그 중에는 '미래'와 '성장동력'처럼 상위 개념의 카테고리와 추상적으로 해석될 여지가 있는 카테고리도 포함되어 효과적인 내용 분류를 어렵게 만든다.

본 논문은 청원글의 내용에 기반한 카테고리 생성으로 의미가 불분명한 카테고리를 없애고 핵심부분만 남길 뿐만 아니라 계층적 카테고리 생성으로 효율적인 분류를 만들어내고자 한다. 마감된 청원(2019 3.31 13:00기준)을 추천 순으로 정렬하고, 그 중 상위 1500개의 글에 k-평균을 활용하여 상위 군집을 찾는다. 상위 군집별 핵심단어를 탐색하고 하위 카테고리가 필요한 경우, 토픽모델링을 추가로 실시하여 계층적 카테고리를 생성한다. 그 결과 '난민, 체육, 외교, 고용/노동, 금융/경제, 사이버범죄, 의료/보건, 보육/교육, 동물, 성/성평등, 폭행, 지역사회, 정치, 기타 ' 14개의 대분류를 생성하고, '보육/교육'의 하위분류로는 '보육'과 '교육', '성/성평등'의 하위로는 '성매매', '성

평등', '낙태법', '지역사회'는 '건설/교통'과 '주거'로 구성된다.

본 논문은 1장 서론에 이어 2장 데이터 크롤링과 정제, 자연어 처리과정(konlpy), 피처 벡터화(Tf-Idf)를 시행하고, 3장 k-평균, 토픽모델링(LDA)으로 군집을 생성하고 핵심단어로 군집 이름 결정, 4장 재구성한 카테고리를 기존의 카테고리와 비교, 평가하는 과정으로 이루어진다.

2. 데이터 수집 및 전처리

2.1 데이터 수집

크롤링에는 파이썬의 BeautifulSoup라이브러리와 함께 Selenium라이브러리, 도구로는 chrome 브라우저를 사용하였다. 분야종합에서 만료된 청원을 추천 순으로 정렬한 뒤, 상위 1500개의 청원글에서 제목, 내용, 카테고리, 게시일시를 뽑아냈다(3월 31일 13:00 기준).

2.2 Konlpy로 토큰화

글을 토큰화하기 위해서는 파이썬을 이용한 형태소 분석기 Konlpy를 사용한다. Konlpy에는 5가지의 형태소 분석기(Kkma, Hannanum, Okt, Komoran, Mecab)가 포함되어있다. 이 5가지 형태소 중 주제와 핵심 단어를 찾는 데 필요한 '일반(보통) 명사'와 '고유명사'를 가장 잘 추출하는 형태소 분석기를 선택할 것이다. 같은 명사를 추출하더라도 어떤 주제와 글이냐에 따라 분석기의 성능이 달라지기 때문에, 성능 평가에는 청원글의 제목을 모은 리스트를 활용한다. 글의 제목이야말로 전체 청원글의 내용 전반을 파악하기에 가장 빠른 방법이기 때문이다.

< 표1 > 파이썬 Konlpy의 분석기별 명사 POS

본 논문에서는 윈도우환경을 사용하므로 윈도우 환경을 지원하지 않는 Mecab, <표1> 에서 알 수 있듯 Konlpy 내에서 일반(보통)명사와 고유명사를 분리할 수 있는 Pos를 가지고 있지 않은 Okt(과거 Twitter)와 Hannanum을 제외한 2가지 형태소 분석기(Kkma, Komoran)에 제목 리스트를 넣어 일반(보통)명사와 고유명사를 출력, 비교하였다.

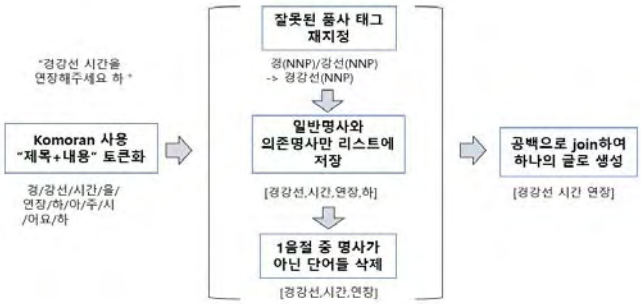
Kkma		Komoran		Mecab		Okt (Twitter)		Hannanum	
NNG	보통명사	NNG	일반명사	NNG	일반명사	NOUN	명사	N	명사
NNP	고유명사	NNP	고유명사	NNP	고유명사				
NNB	일반 의존명사	NNB	의존명사	NNB	의존명사				
NNM	단위 의존명사								
NP	대명사	NP	대명사	NP	대명사				
NR	수사	NR	수사	NR	수사				

< 표2 > Kkma, Komoran의 고유·의존명사 분류 오류

분석기	고유명사 인식 오류	오타 인식 불가	외래어나 신조어
Kkma	평창 : 평 / 창 이수역 : 이수 / 역 표창원 :표 / 창원 유정호: 유정 / 호 등 NNP를 NNG로 분리	ㄱㅐ : 'ㄱ'(NNG) 'ㅐ'(UN)	버닝썬: '버닝'(NNG) '썰'(VV) 'ㄴ'(ETD) 드루킹: '드'(XPN)'루'(NNG) '킹'(NNG) 등
Komoran	유정호: 유 / 정호 경강선: 경 / 강선 등 NNP를 NNG로 분리		버닝썬: '버닝썬'(NA,인식불가) 드루킹: '드'(NNP)'루킹'(NNG) 등

그 결과, <표2>과 같이 Komoran의 성능이 더 좋다는 것을 확인할 수 있었다. 특히 Kkma는 Komoran보다 사람 이름 분석에 취약하였다.

Komoran으로 제목과 내용을 합한 글에 일반명사와 고유명사를 뽑았다. 그 중 Komoran이 잘못 분류한 고유명사와 외래어, 신조어들을 올바르게 분류한 후, 비교적 잘못 분류된 부분이 많은 1음절 단어들은 명사를 제외하고 모두 제거하였다. 이렇게 전처리 한 단어들은 각 글별로 탭으로 연결하여 리스트화한다.



< 그림1 > 전처리 과정

2.3 TF-IDF를 통한 벡터화

TF-IDF란 BOW(Bag Of Words)의 피처 벡터화 중 하나로, 문서에서 특정 단어가 얼마나 중요한 의미를 가지는지를 수치화하는 방법이다. 단순히 단어의 Count 기반에서 한 단계 더 나아가, 중요도가 높은 단어의 가중치를 크게, 빈도수는 많지만 중요도는 낮은 단어의 가중치를 작게 보정해준다. TF(Term Frequency)는 단어의 출현 빈도를 나타내며, Count 기법과 유사하다. 한 문서가 D개의 문장과 N개의 단어로 구성되어 있다고 가정할 때, TF는 특정 단어가 n번 출현할 확률을 뜻한다. IDF(Inverse Document Frequency)는 역문서 빈도로 단어의 희소성을 뜻한다. 전체 문장개수 D를 특정 단어를 포함한 문장 개수 d로 나눈 값에 로그를 적용한 값이다. TF-IDF는 이런 출현빈도인 TF와 역문서 빈도를 뜻하는 IDF를 곱한 값이다. 본 논문에서는 TF-IDF의 구현을 위해 scikit learn에서 제공하는 TfidfVectorizer모듈을 사용한다.

TF-IDF를 통해 2.2의 리스트들을 벡터화한다. 그 결과, (1500 ,16051) 크기의 CSR 행렬이 반환된다.

3. 카테고리 생성

3.1 k-평균으로 군집화

TF-IDF로 생성한 벡터에 k-평균 알고리즘을 사용하여 유사한 글끼리 그룹화한다. k-평균 알고리즘은 k개의 군집 중심의 위치를 랜덤으로 선정한 뒤, 그 중심들을 기준으로 군집을 구성하고, 군집별 평균 위치로 중심 k를 이동시킨다. 이 과정은 k가 더 이상 이동하지 않을 때까지 계속된다. k-평균 알고리즘은 단순하지만 성능이 비교적 좋으며, 구현이 쉬우나, 군집의 개수를 직접 설정해야 실시할 수 있다. 본 논문에서는 k-평균 알고리즘의 구현을 위해 scikit-learn 라이브러리를 활용한다. 또한 군집개수 설정을 위해, 군집개수를 8, 10, 12, 14개로 늘려가며 군집이 몇 개일 때 청원글이 가장 잘 분류되는지 알아낸다. 그 결과, 군집의 개수가 10일 때 청원글이 가장 효과적으로 분류된다.

군집개수가 10인 k-평균을 실시한 후, 군집별로 핵심 단어를 상위 10개씩 추출한다. 그 결과는 다음과 같다.

< 표3 > 10개 군집별 상위 10개 핵심단어

	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
1	난민	미세먼지	수사	병원	교사
2	선수	중국	처벌	간호사	어린이집
3	이슬람	시간	공매도	환자	보육
4	연맹	일본	불법	의료	학생
5	빙상	기업	사건	치료	교육
6	빙상연맹	회사	사이트	의사	원장
7	제주도	근무	판사	수술	시간
8	외국인	노동자	범죄	보험	아이들
9	한국	업체	법	아이	보육교사
10	문화	우리나라	주식	병	학교

	cluster_5	cluster_6	cluster_7	cluster_8	cluster_9
1	여성	동물	지역	가해자	학생
2	남성	견	주민	폭행	대통령
3	성매매	강아지	임대	피해자	학교
4	성	반려	주택	아이	국회의원
5	낙태	학대	공공	사건	정부
6	지원	유기	아파트	경찰	반대
7	여성가족	유기견	신도시	처벌	국가
8	평등	반려동물	교통	사람	교육
9	사회	보호소	공사	말	사람
10	인권	반려견	분양	생각	의원

<표3>을 보면, cluster_0은 '난민, 이슬람, 선수, 제주도'의 난민에 관한 분야와 '선수, 빙상, 빙상연맹'의 체육 분야로 구성된다. 난민과 체육이 전혀 다른 주제라고 생각됨에도 같은 군집에 속하는 것은 난민문제와 체육 분야가 '외국인', '문화', '한국', '나라'와 같이 공통적인 단어들을 많이 사용하기 때문이다.

cluster_1은 '미세먼지, 중국, 일본, 우리나라'와 같은 미세먼지 및 외교 문제와 '기업, 회사, 근무, 노동자, 업체'와 같은 고용, 노동의 주제로 파악된다. 외교와 고용문제는 깊은 연관이 있다. 외교문제가 악화될 때 우리는 관련된 우리 기업의 타격을 우려한다.

cluster_2는 '공매도, 주식, 사이트'의 단어들로 보아 금융, 화폐분야, 혹은 사이트에 관한 주제들로 파악된다. 금융 주제와 '사이트'라는 단어가 같은 군집에 속한 이유는 주식, 블록체인 등 금융의 많은 부분이 인터넷과 뗄 수 없는 관계이기 때문으로 파악된다.

cluster_3은 '병원, 간호사, 환자, 의료, 치료, 의사, 수술, 보험, 병'의 의료 관련 단어들로 구성된다. cluster_4는 '교사, 어린이집, 보육, 학생, 교육, 원장, 아이들, 보육교사, 학교'로 보아 보육과 교육 관련 주제이다. cluster_5는 '여성, 남성, 성매매, 성, 낙태, 지원, 여성가족, 평등, 사회, 인권'으로 성과 성평등에 관한 주제이다. cluster_6은 '동물, 견, 강아지, 반려, 학대, 유기, 유기견, 반려동물, 보호소, 반려견'으로 동물에 관한 주제이다.

cluster_7은 '지역, 주민, 임대, 주택, 공공, 아파트, 신도시, 교통, 공사, 분양'의 단어들로 주거, 건축, 교통에 관한 주제들, 즉 지역사회 문제들로 구성된다. cluster_8은 '가해자, 폭행, 피해자, 사건, 경찰, 처벌'의 단어로 보았을 때 폭행사건과 관련된 주제이다. cluster_9는 '학생, 학교, 교육'의 교육과 관련된 분야와, '대통령, 국회의원, 정부, 국가, 의원'과 같은 정치에 관한 분야로 파악된다.

k-평균 알고리즘과 군집별 핵심단어 추출을 통해 상위 카테고리에 올 수 있는 주제를 확인해보았다. 3.3에서는 LDA를 활용하여 k-평균으로 확실히 분류되지 않은 상위 카테고리를 재설정하고, 하위 카테고리를 결정한다.

3.2 토픽모델링(LDA) 실시

k-평균을 실시한 것 중 상위 카테고리가 명확하고 세부 군집이 필요 없는 cluster_3의 의료, cluster_6의 동물, cluster_8의 폭행을 제외한 7개 군집에 토픽모델링을 적용하여 군집별로 10개의 단어를 뽑아낸다. 토픽모델링으로는 LDA(Latent Dirichlet Allocation)를 활용한다.

< 표5 > 토픽모델링을 통한 세부 군집화 결과

군집이름	핵심주제	LDA로 세부 군집
cluster_0	난민,체육	난민, 외국인, 대한민국, 이슬람, 예멘, 문화, 문제, 제주도, 정부, 나라
		선수, 한국, 무슬림, 나라, 사람, 사회, 올림픽, 이민자, 문화, 아랍
cluster_1	외교,노동	미세먼지, 중국, 우리나라, 일본, 문제, 사람, 한국, 게임, 정부
		시간, 근무, 업체, 회사, 노동자, 직원, 최저임금, 기업, 개선, 지역
cluster_2	금융,사이트	공매조, 시장, 주식, 금융, 불법, 정부, 투자자, 화폐, 블록체인, 가상
		처벌, 불법, 사건, 촬영, 수사, 범죄, 사이트, 피해자, 사림, 일베
cluster_4	보육,교육	교사, 학생, 교육, 학교, 학대, 아이, 학부모, 문제, 수업, 의무
		교사, 어린이집, 보육, 시간, 아이들, 아이, 원장, 교육, 보육, 교원
cluster_5	성,성평등	남성, 지원, 가족, 예산, 전용, 정책, 대한민국, 여성, 폐지, 의무
		남성, 낙태, 사회, 임신, 평등, 차별, 임금, 현실, 문제, 여자
		성매매, 남성, 처벌, 한국, 노르딕, 모델, 여성, 매수, 인식, 사람
cluster_7	지역사회	주민, 지역, 사업, 교통, 서울, 정부, 공사, 아파트, 문제, 건설
		임대, 공공, 분양, 주택, 10년, 아파트, 서민, 전환, 지역, 부동산
cluster_9	정책,정치	학교, 학생, 교육, 장애인, 시간, 지원, 경찰, 제도, 생각, 가족
		대통령, 정부, 사림, 국가, 생각, 의원, 청와대, 국회, 국회의원, 국채

cluster_0의 첫 번째 토픽은 난민, 두 번째는 체육과 난민에 관한 주제인데, k-평균과 마찬가지로 공통된 단어들이 많아 명확히 분리되지 않았다. cluster_1은 미세먼지와 인근 국가들에 대한 주제와 노동, 고용에 대한 주제이다. 미세먼지는 군집의 개수를 늘려도 외교, 국가문제와 완벽히 분리되지 않았는데, 이는 미

세먼지에 관한 글에 중국이라는 단어가 함께 쓰였기 때문이다.

cluster_2는 금융과 관련된 주제와, 불법촬영 및 유포, 사이버범죄와 관련된 주제로 구성된다. k-평균에서는 사이버 범죄에 관련된 단어가 명확히 드러나지 않았는데, 토픽모델링으로 불법 촬영, 일베 등의 구체적인 내용이 확인되었다. cluster_4는 학교/교육과 보육으로 명확히 분류되었다.

cluster_5에서 여성가족부와 성평등에 관한 제도와 정책에 대한 주제로 파악되며, 두 번째는 성평등과 낙태/임신, 임금불평등에 관한 주제로 파악된다. 군 집개수를 늘려도 임신과 임금이라는 단어는 분리되지 않는데, 이는 임신으로 인한 육아휴직, 휴가가 직장 내 임금 불평등과 관련이 있기 때문이다. 세 번째는 성매매와 관련된 주제이다.

cluster_7은 지역사회에 관한 내용으로, 개발/건설과 주거문제로 구성된다. 개발/건설은 신도시나 재개발 문제 등의 이슈로 나타난다. 주거문제는 내 집 마련이나 부동산문제들로 인한 것으로 보인다.

cluster_9는 학교와 정부에 관한 단어들로 구성되는데, cluster_9의 학교주제는 cluster_4의 학교/교육과 달리 교육 정책이나 제도에 관한 내용으로 파악된다. 두 번째 주제는 정치에 관련된 주제이다.

4. 결론

그 결과, <표6>의 카테고리는 <그림3> 와 같은 카테고리로 변경되었다. 새롭게 제시하는 카테고리의 특징을 크게 네 가지로 정리할 수 있다. 첫째, '미래', '성장동력'과 같이 분류가 명확하지 않은 카테고리들이 삭제되었다. 둘째, '사이버범죄'와 '폭행' 과 같은 상위 카테고리가 생성되었다. 이는 기존 카테고리로는 명확히 분류될 수 없는 내용까지 포함하는 것이다. 셋째, 하위 카테고리가 생성되었다. '보육/교육', '성/성평등', '지역사회'의 상위 카테고리에서 가장 핵심이 되는 하위 카테고리를 생성하여 청원 글들의 주요 내용을 한눈에 파악하고 세분화된 분류를 이끌어냈다.

넷째, 기존 카테고리에 존재하던 몇몇 주제들이 새로운 카테고리에는 존재하지 않거나, 혹은 통합되었다. '외교/통일/국방'은 '외교'로 바뀌었으며, '농산어촌', '안전환경', '저출산/고령화', '행정'이 삭제되었다. 이는 삭제된 카테고리들이 신규 카테고리에서도 남아있을 만큼 핵심 내용이 되지는 못했다는 것으로 해석된다.

두 카테고리의 공통점으로는 <그림3>의 카테고리가 <표6>과 마찬가지로 기타를 포함하고 있다는 것이다. 이는 기타가 청원내용을 기반으로 생성한 13개 카테고리 중 어느 곳에도 해당하지 않는 청원내용들을 포함하기 위함이며, 향후 연구에서 더 보완적인 카테고리를 생성하는 데에 활용되기 위함이다. 본 논문에서 제시하는 카테고리가 주기적으로 갱신되어 더욱 체계적이고 많은 부분들을 포함하게 된다면, 청원글 전체의 효율적 관리뿐 아니라, 카테고리 구조만 보고도 사회쟁점과 이슈의 흐름에 대해 파악할 수 있을 것이다.

< 표6 > 국민청원의 기존 카테고리

정치개혁	외교/통일/국방	일자리	미래
성장동력	농산어촌	보건복지	유아/교육
안전환경	저출산/고령화	행정	반려동물
교통/건축/국토	경제민주화	인권/성평등	문화/예술/체육/언론
기타			

< 그림3 > 새로운 카테고리

<<참고문헌>>

[1] David M.Blei, Andrew Y.Ng, Michael I.Jordan, "Latent Dirichlet Allocation",Journal of Machine Learning Research 3: 993-1022, 2003

[2] 박건숙, "빅데이터의 하위 주제어 의미 분석 연구, Vol.65, 한국언어학회, 2013

[3] 유홍연, 이승우,고영중, "실시간 이슈 분석을 위한 뉴스 군집화 및 다중 문서 요약", 제 30회 한글 및 한국어 정보처리 학술발표 논문집

[4] 고동우, 양정진,KoNLPy와 Word3Vec을 활용한 한국어 자연어 처리 및 분석, No.6, 한국정보과학회 학술발표논문집,2018

[5] 박은정, 조성준, KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지, 제 26회 한글 및 한국어 정보처리 학술대회 논문집, 2014