NEW TECHNOLOGY + 03

텍스트 마이닝을 활용한 토목분야 연구 토픽 분석

텍스트 마이닝은 자연어 처리 기술을 활용하여 텍스트 데이터를 추출하고 데이터가 가진 특징, 의미를 찾을 수 있도록 하는 기법을 뜻한다. 이에 따라 연구동향 분석뿐만 아니라 검색엔진, 민원 시스템 등 다양한 분야에서 활용되고 있다.

박준용 / 서울기술연구원 도시인프라연구실 전임연구원

서론

토목에서 다루는 연구 분야는 구조, 콘크리트, 지반, 도시계획 등 상당히 다양하다. 여기에 세부 분야별로 AI, IoT, 빅데이터 등 4차 산업혁명과 관련된 연구 분야까지 융합되어 연구의 방향이 더욱 다양해지고 복잡해지고 있다. 이 때문에 토목뿐만 아니라 다른 분야에서도 최신 연구 동향을 파악하기 위한 연구도 활발히 수행되고 있다[1)~4)].

연구 동향을 파악하기 위해서는 전문가 자문 등을 포함하는 정성적 방법과 정량적 방법인 계량서지학방법론이 있고 그중에서도 토픽 모델링 방법이 주로 활용되고 있다[1)]. 이 연구에서는 2017~2021년도 대한토목학회국문논문집에 수록된 논문들을 대상으로 텍스트마이닝 및 토픽모델링 기법을 활용하여 토목분야 최신 연구토픽 분석을 수행하였다.

텍스트 마이닝과 토픽 모델링

텍스트 마이닝은 자연어 처리 기술을 활용하여 텍스트 데이터를 추출하고 데이터가 가진 특징, 의미를 찾을 수 있도록 하는 기법을 나타낸다[2)]. 주로 인터넷에서 원하는 정보를 크롤링 혹은 스크래핑하여 축적하여 사용하며, 축적된 데이터는 불용어 제거 등을 포함하는 전처리 과정을 거쳐 토픽모델링, 감정분석 등의 분석 기법을 통해 분석 결과를 도출한다.

텍스트 데이터가 가지는 주제를 추출하기 위해서는 단순히 단어의 빈도만으로는 분석이 어렵기 때문에, 문서 집합에서 텍스트를 분석하여 주제를 추출해낼 수 있는 LDA(Latent Diraichlet Allocation) 기법이 토픽모델링에서 활발히 사용되고 있다. 특히 단어를 문맥에 따라 분석할 수 있으며 토픽별로 추출된 단어들이 독립성을 갖는다는 측면에서 장점을 가진다[3)]. 이러한 특징으로 연구동향 분석뿐만 아니라 검색엔진, 민원 시스템 등 다양한 분야에서 활용되고 있다.

토목 분야 연구 토픽 분석

토목학회는 우리나라 토목분야 최대 학회이므로 대표성을 띤다고 판단하여 대한토목학회국문논문집에 수록된 최근 5개년 논문을 활용하여 분석하였다. 총 436개의 논문 제목을 추출하였으며, 분석의 편의성을 위하여 영문만으로 작성된 영문제목을 사용하였다.

텍스트 마이닝 단계에서는 Python NLTK 라이브러리를 활용하여 특수문자 제거, 소문자 변환, 단어 토큰화, 불용어 처리를 수행하였다. 불용어 처리에서는 기본적으로 전치사, 접속사와 같은 무의미한 품사를 제거하였고, 추가적으로 연구주제 분석에 활용되지 않는 research, study, method

그림 1. 최적 토픽 수 선정을 위한 일관성지수 산정

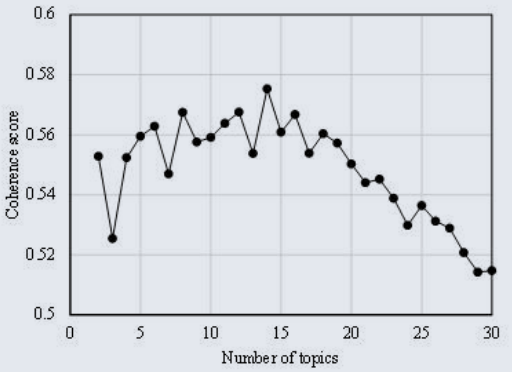
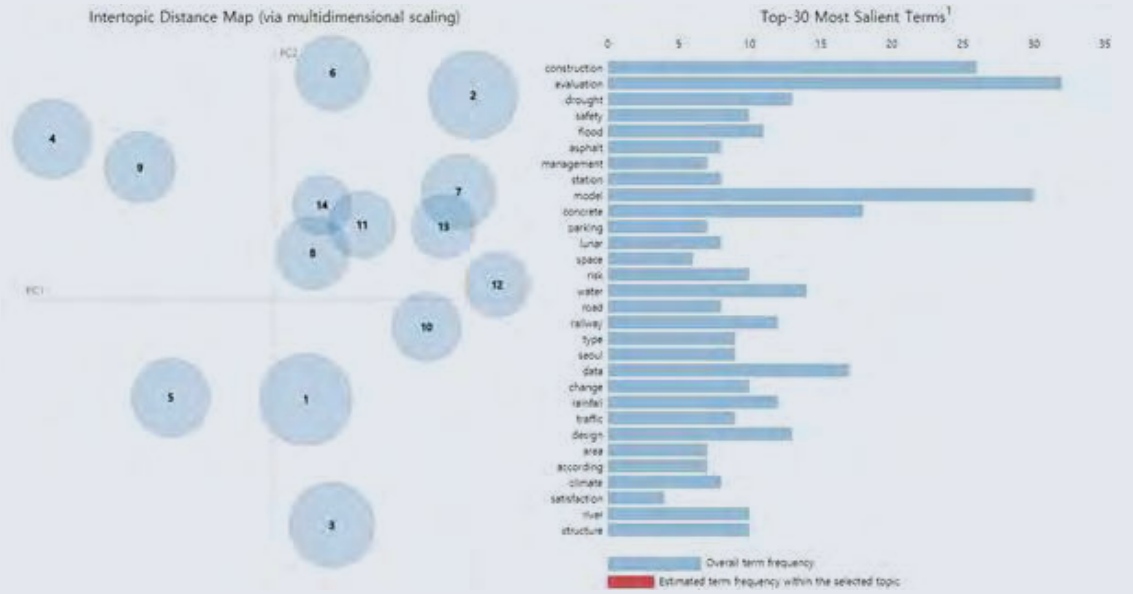


그림 2. 토픽 수 선정을 위한 일관성지수 산정



등의 일반적인 문구 또한 불용어 처리하였다. 불용어 처리 전 전체 단어 수는 5,984개이고, 불용어 처리 후 최종적으로 분석에 활용된 단어 수는 3,807개이다.

LDA 기법을 활용하여 토픽 모델링을 수행하기 위해서는 먼저 사용자가 텍스트 데이터가 갖는 토픽의 개수를 결정해야 한다. 토픽의 개수가 결정되면 LDA는 토픽마다 가장 적합한 단어를 할당하여 결과적으로 토픽별로 단어가 군집화된다. 이때 토픽의 개수를 임의로 가정하기도 하지만, 일관성(Coherence) 지수를 활용하여 설정된 토픽의 개수에 따라 도출된 각각의 토픽과 단어들의 일관성을 비교하여 최적의 토픽 개수를 결정하기도 한다. 이때 일관성 지수는 0에서 1까지의 값을 가지고 큰 값을 가질수록 토픽에 할당된 단어들의 일관성이 높다고 볼 수 있다. 토픽의 개수를 2부터 30까지 고려하여 LDA모델을 도출한 뒤 일관성 지수를 비교한 결과, <그림 1>과 같이 14개의 토픽에서 0.575의 가장 높은 일관성 지수를 보였다.

토목에서 구조공학, 수공학, 교통공학, 도시공학, 지반공학 등 10여 개의 세부분류가 가능한 것을 놓고 보았을 때 합리적인 토픽 수라고 보인다.

14개 토픽에 대한 모델링 결과는 <그림 2>와 같이 포함되는 단어들의 특성에 따라 군집화된 형태로 표현된다. 토픽 1은 전체 단어의 10.8%가 포함되며 construction, data, work와 같은 키워드들이 중심을 이루고 있어 시공단계에서 중장비, 작업자의 안전 및 관리 관련 연구가 많은 것으로 보인다. 토픽 2는 10%를 차지하며 flood, parking, learning, new, drone, detection 등의 단어로 구성된다. 단어의 구성으로만 보면 연관성이 떨어져 보이지만, 단어가 쓰였던 논문의 제목들을 살펴보면 영상데이터 분석, 빅데이터 분석을 활용한 융합분야로 보인다. 토픽 3은 9.3%를 차지하며 climate, change, drought, risk, air 등의 단어가 포함되어 기후변화와 이에 대한 리스크를 분석하기 위한 연구가 많이 수행되고 있음을 알 수 있다. 이외에도 도로시설물의 성능 평가 및 유지관리, 딥러닝등 최신기술과 융합된 교통 안전 관련 연구 등의 연구주제를 확인할 수 있다.

텍스트 마이닝 및 토픽 모델링 기법을 적용하여 토목학회 국문논문집에 수록된 논문들의 연구 주제를 분석해 본 결과, 단어들의 조합만으로 세부분야가 많은 토목분야의 연구주제들이 대략적으로 추정이 되었으며 연구 동향 분석에 도움이 될 것으로 판단된다. 다만, 단순히 군집화된 단어들로만 상세 연구 주제를 추정하기는 생각보다 쉽지 않으며 군집화된 단어들이 사용된 논문을 다시 살펴보는 등의 추가적인 결과 분석이 동반되어야 합리적인 동향 분석이 이루어질 것이다. 그리고 대표성을 갖는 학회지라고 하더라도 샘플의 수가 크지 않다 보니, 연도별 유행하는 연구주제 분석, 특정키워드의 연도별 지분율 추이 등 다각적 분석이 어려웠다. 다각적, 심층적 분석을 위해서는 충분한 샘플이 확보되도록 분석계획을 잡는 것 또한 중요하다.

참고문헌

[1] 이유빈, 이영호, 성정창, 애나 스타네스쿠, 지상훈, 황철수. (2020). 계량적 모델을 통한 지리학 연구의 최신동향 및 토픽 분석, 대한지리학회지, 55(6), 589-599.
[2] 박홍진. '인공지능', '기계학습', '딥 러닝' 분야의 국내 논문 동향 분석. 한국정보전자통신기술학회 논문지 13.4 (2020): 283-292.
[3] 신명선, 조경원. (2019). 텍스트마이닝을 활용한 한국언어치료학회지의 토픽 모델링 및 트렌드 분석:(2002~2018), 언어치료연구, 28(3), 81-91.
[4] 김성연. (2020). 텍스트마이닝 기법을 활용한 미국산업응용수학 학회지의 연구 현황 및 동향 분석, 한국콘텐츠학회논문지, 20(7), 212-222.