텍스트 마이닝을 이용한 건설기술 연구개발 동향 분석

정성윤 KICT 미래스마트연구본부 연구위원

들어가며

연구보고서는 특정 연구의 배경, 범위, 목적, 내용, 방법, 절차, 결과 등 연구과제에 관한 모든 내용을 기록한 문서를 말한다. 보고서 내용에는 연구과제 수행 시점뿐만 아니라 앞으로 관심을 가지는 암묵지 성격의 정보를 내포하고 있다. 이러한 정보를 유추한다면 시간의 흐름에 따라 연구 주제어가 어떻게 변해 왔고, 앞으로 어떻게 변할 것인지를 추정할 수 있다. 이 글에서는 텍스트 마이닝(Text Mining)을 사용하여 국내 건설기술 연구개발 동향을 분석하고자 한다.

한국건설기술연구원(이하 KICT)은 건설기술 연구개발을 대표하는 연구기관이라 할 수 있다.

KICT에서 발간한 연구보고서에 내재한 정보를 분석한다면, 국내에서의 건설기술 연구개발 동향을 파악할 수 있다고 판단하였다. 따라서 KICT 전자도서관에서 제공되는 보고서의 서지 정보를 수집했다. 먼저, 서지 정보 중 보고서 제목을 분석 대상으로 하여 불용어 제거, 형태소 분석 등 전처리 과정을 통해 출현 빈도가 높은 단어를 추출하였다. 토픽 모형화(Topic Modeling)를 통해 특정 주제에서 연관되는 단어를 분석하였으며, 네트워크 이론(Network Theory)을 적용하여 건설기술 연구개발 네트워크에서 단어 중요도를 분석하였다. 끝으로, 중요도가 높은 단어들이 시간이 흐르면서 어떻게 변화하였는지를 살펴보았다.

텍스트 마이닝 소개

텍스트 마이닝은 통계학적 계산과정을 통해, 비정형적인 텍스트에 내재한 일정한 형태를 보이는 유의미한 정보를 추출하는 기법을 말한다. 텍스트 마이닝을 적용하기 위해서는 먼저, 텍스트에 포함된 단어를 추출한 후에 의미가 낮은 특수문자, 숫자, 한 음절 등 단어를 분석 대상에서 제외한다. 또한, 형태소를 분석하여 고유한 의미를 갖는 명사를 추출한다. 이때 한글을 분석 대상으로 한 경우라면 영문자 단어는 분석에서 제외한다. 명사 단어 중 '연구', '자료', '보고서', '최종' 등과 같이 단독으로는 의미 전달이 낮은 단어도 분석에서 제외한다. 이러한 전처리 과정을 거친 후, 단어의 출현 빈도수를 분석한다.

다음으로, 단어 간의 출현순서 등을 반영한 행렬과 확률 분포를 토대로 유사한 의미를 갖는 연관된 단어를 하나의 그룹으로 묶은 토픽 모형화를 분석한다. 토픽 모형화는 문서, 뉴스, 블로그 등과 같은 비정형 텍스트에서 특정 주제 분포와 주제에 해당하는 단어의 가중치와 유사도를 분석한다. 계속해서, 단어 간의 연관관계를 시각적으로 표현한 네트워크를 생성한다. 네트워크에서 특정 단어가 이웃한 단어에 얼마나 영향을 미치는지, 이웃 단어로부터 얼마나 영향을 받는지에 대한 네트워크 중심성(Network Centrality)을 측정한다. 네트워크 중심성은 하나의 단어를 노드(Node)로 표시하고, 노드 간의 관계를 연결선(Edge)으로 나타낸 네트워크 형태의 그래프에서 노드의 중심적 위치를 정량적 계수(Coefficient)로 계산한 수치를 말한다. 이때 네트워크의 위치는 중요도(Importance) 또는 영향도(Power) 등 의미로 사용된다. 네트워크 중심성은 텍스트 유형과 분석 목적에 따라 여러 모형의 중심성을 사용할 수 있으나, 연결(Degree), 근접(Closeness), 매개(Between), 고유벡터(Eigenvector) 등 중심성 모형을 대표적으로 사용한다(정성윤, 김남곤, 2020).

기초 자료의 특성 분석

건설기술 연구개발 동향을 분석하기 위해 KICT 전자도서관[1]에서 제공하는 1980년부터 2022년까지의 연구, 조사, 출장, 회의 등 보고서를 비롯하여 KICT에서 발간한 지침서, 연보 등 7,472건의 소장자료를 수집하였다. 동향 분석과 관련성 낮은 출장, 회의 등의 보고서, 지침서, 연보와 중복된 연구보고서, 연구보고서를 분철한 자료, 출판연도가 명시되지 않은 자료 등을 제외한 4,242건의 보고서를 분석 대상의 기초 자료로 선정하였다. 표 1은 분석에 사용한 기간별 기초 자료의 건수를 나타낸 것이다.

기초 자료에는 2011년 이후에 초록 서지 정보가 없어, 보고서 제목을 분석 데이터로 사용하였다. 그림 1은 분석 데이터에서 출현 빈도가 많은 상위 200개의 단어를 시각적으로 표현한 것이다.

그림 1에서 보듯이 1980년~1999년에는 시스템, 설계, 관리, 조사, 계획 등 단어가 분석 데이터에서 사용 빈도가 높았다. 2000년~2014년에는 시스템, 조사, 관리, 설계 등 단어가 높았으며, 2015년~2022년에는 시스템, 도로, 관리, 운영, 구축 등의 단어들이 높은 사용 빈도를 보였다. 전체 기간에는 시스템, 관리, 도로, 조사, 설계 등 단어의 출현 빈도가 높았다. 이들 단어는 연구개발에서 관심이 높았다고 유추할 수 있다.

1. KICT 전자도서관 내 KICT 보고서 검색, https://kict.codil.or.kr/srchdetail.do.

토픽 모형화 분석

분석 데이터에 포함된 단어 간에 일정한 형태를 보이는 유의미한 맥락을 파악하기 위해 토픽 모형화를 분석하였다. 토픽 모형화 분석을 위해 토픽의 개수를 결정해야 하는데, 토픽의 개수에 따라 토픽에 부합하는 주제어가 달라질 수 있다. 토픽 개수를 결정하기 위해 주제 일관성 점수(Topic Coherence Score)와 혼란도(Perplexity)를 적용할 수 있으나, 이 글에서는 9개 KICT 연구본부의 수를 토픽의 개수로 결정하였다. 표 2는 텍스트 마이닝 분석 도구인 'NetMiner 4'를 사용하여 토픽에 따라 유사한 의미를 갖는 상위 10개의 주제어를 나타낸 것이다. 토픽 모형화 분석을 통해 1980년~1999년에는 도로, 수자원, 하천, 주택에 관한 연구개발이 많았던 것으로 유추할 수 있다. 2000년~2014년에는 환경과 정보 시스템이,

구분	'89 이전	'90~'99	'00~'04	'05~'09	'10~'14	'15~'19	'20~'22	합계
건수	208	812	315	785	918	841	363	4,242

표 1 기초 자료 건수

(1980~1999 기간) (2000~2014 기간)

(2015~2022 기간) (전체기간)

그림 1 출현 빈도가 높은 상위 200개의 단어

구분	1980~1999	2000~2014	2015~2022
토픽 1	설계, 교량, 콘크리트, 도로, 공법, 시공, 조사, 유지관리, 관리, 보수	조사, 포장, 활용, 설계, 지침, 기준, 콘크리트, 구조, 품질, 합성	시스템, 도로, 운영, 관리, 개선, 유지관리, 시설, 설계, 구축, 업무
토픽 2	조사, 관리, 하천, 수자원, 하수, 환경, 현황, 실태, 수질, 개선	하천, 설계, 실험, 조사, 기준, 기능, 자연, 생태, 도시, 정비	정보, 도로, 교통, 서비스, 환경, 기획, 고도, 공간, 소음, 구축
토픽 3	설계, 에너지, 지하, 공간, 요소, 건물, 기준, 해석, 생활, 활용	구축, 정보, 관리, 시스템, 교통, 운영, 체계, 업무, 국도, 지식	성능, 활용, 구조물, 포장, 콘크리트, 공법, 아스팔트, 구조, 시공, 재료
토픽 4	정보, 시스템, 관리, 체계, 계획, 도로, 구축, 교통, 장기, 수립	성능, 공법, 구조물, 콘크리트, 설계, 보강, 내화, 건축물, 시공, 교량	관리, 시스템, 하천, 기획, 시설물, 제어, 활용, 구조물, 공정, 물질
토픽 5	공사, 기준, 안전, 구조, 진단, 건축, 표준, 정수, 설비, 시방서	도로, 사면, 대책, 해석, 안정, 공사, 절토, 설계, 용역, 교량	조사, 대응, 수문, 홍수, 예측, 재난, 하천, 플랫폼, 지역, 활용
토픽 6	제도, 개선, 설계, 용역, 폐기물, 매립, 시공, 방지, 표준, 대책	시스템, 관리, 안전, 도로, 터널, 통합, 재해, 예방, 공사, 최적	시스템, 에너지, 구축, 주택, 건축물, 제도, 건물, 성능, 관리, 인증
토픽 7	홍수, 프로그램, 경보, 모형, 유역, 조사, 시험, 시스템, 수문, 특성	조사, 성능, 현황, 시스템, 일본, 에너지, 제도, 주택, 개선, 현장	안정, 검토, 도로, 위험, 공사, 대책, 터널, 비탈면, 지반, 대응
토픽 8	시설, 하천, 계획, 정비, 터널, 사회, 능력, 홍수, 방어, 참고	계획, 조사, 특성, 수립, 수자원, 개선, 장기, 측정, 유량, 용역	활용, 수립, 시스템, 안전, 전략, 인프라, 순환, 계획, 환경, 콘크리트
토픽 9	주택, 시스템, 공동, 성능, 설계, 초고층, 최적, 아파트, 건물, 계획	환경, 영향, 오염, 물질, 저감, 소재, 매립지, 실내, 수질, 주변	친환경, 시스템, 교량, 기준, 지원, 설계, 모니터링, 현지, 생태, 상용

표 2 토픽 모형화 결과

구분	1980~1999	2000~2014	2015~2022
연결 중심성	설계(0.073) 시스템(0.057) 관리(0.046) 정보(0.040) 공사(0.039) 도로(0.036) 조사(0.034) 주택(0.033) 계획(0.030) 시설(0.029)	시스템(0.064) 설계(0.057) 관리(0.048) 성능(0.048) 조사(0.043) 도로(0.038) 기준(0.032) 정보(0.032) 콘크리트(0.032) 안전(0.031)	시스템(0.069) 도로(0.056) 활용(0.052) 관리(0.049) 설계(0.042) 성능(0.037) 안전(0.034) 구조물(0.032) 에너지(0.030) 콘크리트(0.029)
근접 중심성	설계(0.343) 관리(0.329) 시스템(0.327) 주택(0.310) 공사(0.307) 조사(0.306) 계획(0.306) 공법(0.303) 개선(0.301) 시공(0.300)	설계(0.347) 관리(0.345) 시스템(0.342) 성능(0.342) 안전(0.335) 조사(0.328) 도로(0.322) 현황(0.321) 개선(0.319) 기준(0.318)	시스템(0.359) 도로(0.357) 설계(0.349) 관리(0.345) 활용(0.343) 성능(0.340) 안전(0.336) 인프라(0.332) 개선(0.332) 구조물(0.328)
매개 중심성	설계(0.175) 시스템(0.127) 관리(0.090) 주택(0.062) 공사(0.057) 도로(0.054) 조사(0.052) 환경(0.049) 공법(0.049) 정보(0.045)	시스템(0.112) 설계(0.104) 성능(0.090) 관리(0.083) 도로(0.076) 조사(0.076) 콘크리트(0.052) 구조물(0.049) 안전(0.044) 환경(0.042)	시스템(0.125) 도로(0.113) 활용(0.099) 설계(0.069) 관리(0.067) 성능(0.052) 구조물(0.048) 개선(0.045) 콘크리트(0.044) 저감(0.039)
고유벡터 중심성	설계(0.457) 관리(0.383) 시공(0.310) 시스템(0.294) 기준(0.208) 정보(0.201) 교량(0.197) 체계(0.189) 계획(0.161) 지침(0.144)	시스템(0.532) 관리(0.457) 운영(0.257) 구축(0.244) 정보(0.193) 안전(0.164) 체계(0.159) 통합(0.147) 설계(0.147) 조사(0.124)	시스템(0.586) 관리(0.477) 운영(0.367) 구축(0.224) 통합(0.173) 도로(0.165) 정보(0.154) 교통(0.114) 유지관리(0.105) 품질(0.097)

표 3 네트워크 중심성 결과

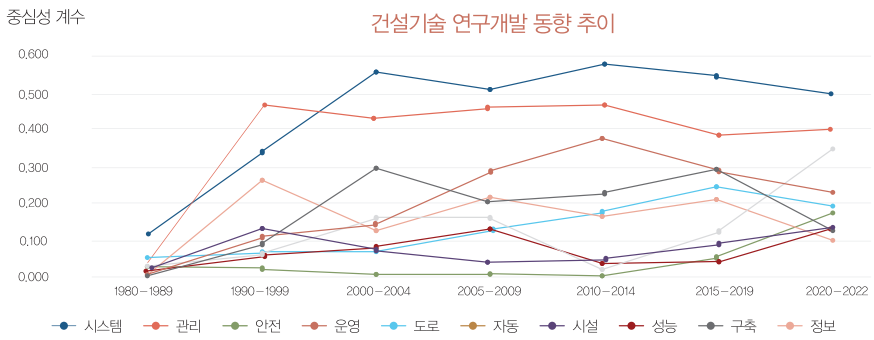

그림 2 시기별 연구개발 관련 단어의 중요도 변화 추이

2015년~2022년에는 에너지, 친환경, 안전 등의 주제에 관한 연구개발이 추가로 많았다.

네트워크 중심성 분석

토픽 모형화는 주제어 간의 연관관계는 파악할 수 있으나 특정 주제어가 네트워크에서 얼마나 중요한지 또는 영향력을 갖는지는 알 수 없다. 이를 보완하기 위해 네트워크 이론에서 대표적으로 사용하는 연결, 매개, 근접, 고유벡터 등 모형별로 중심성 계수를 측정하여 단어의 중요도를 분석하였다. 표 3은 중심성 모형별로 중요도가 높은 상위 10개의 단어를 나타낸 것이다.

모든 중심성 모형에서 시스템, 설계, 도로, 관리, 성능, 안전, 조사 등의 단어가 건설기술 연구개발 네트워크에서 영향력이 높은 것으로 파악되었다. 특히, 2015년 이후에는 추가로 에너지, 안전, 저감, 품질, 개선 등에 대한 연구개발이 중요한 것으로 유추할 수 있었다.

연구개발 동향 분석

고유벡터 중심성은 특정 단어와 이웃 단어 간의 최단 거리를 연쇄적으로 누적한 중심성 계수를 가지고서 단어의 중요도를 측정한다. 고유벡터 중심성을 기초로 단어의 중요도가 시간의 흐름에 따라 어떻게 변하였는지를 살펴보았다. 이 글에서는 단어의 중요도 변화추이를 일관되게 분석하기 위해 중요도가 높은 단어 순으로 모든 기간에서 존재하는지를 확인하였다. 그림 2는 모든 기간에서 존재하는 상위 10개 단어의 중요도가 어떻게 변화하였는지를 시각화한 것이다.

그림 2에서 보면 '시스템'과 관련한 연구개발은 2000년대 초반까지 급격히 증가하였고 현재까지도 꾸준히 중요한 위치에 있다. '관리'와 관련한 연구개발은 1990년대에 급격히 증가하였다. '정보', '구축', '운영', '도로'와 관련한 연구개발은 2000년대에 관심이 많았으나 최근에는 중요도가 감소하였다. 연구개발에서 '자동'의 중요도는 2020년대에 급격히 증가하였다. 이는 스마트 건설에 관한 관심과 투자가 많아지면서 이에 따른 연구개발도 많아졌다고 해석할 수 있다.

맺음말

건설기술 연구개발을 본격적으로 시작한 1980년대 이래에 현재까지 KICT 전자도서관에서 제공하는 연구 및 조사보고서를 기초 자료로 하여 건설기술 연구개발과 관련한 단어가 시간이 흐름에 따라 어떻게 변화하였는지를 살펴보고자 하였다. 이를 위해 텍스트 마이닝 기법 중 토픽 모형화를 통한 주제별 연관단어와 건설기술 연구개발 네트워크에서 중심적 위치에 있는 단어 그리고 단어의 중요도가 시기별로 어떻게 변화하였는지를 시각적으로 표현하였다. 하지만 다음과 같이 제약과 추가연구가 필요하다. 첫째, 텍스트 마이닝은 비정형 텍스트를 대상으로 하나, 이 글에서는 보고서 제목이라는 단문 형태로 텍스트를 분석 대상으로 하였다. 따라서 일정한 형태를 보이는 유의미한 정보를 추출하는 데 한계가 있다. 이를 보완하기 위해서는 보고서의 초록을 분석 대상으로 사용할 필요가 있다. 둘째, 국내뿐만 아니라 미국과 같은 기술 선진국을 비롯하여 베트남, 중남미 등 개발도상국에서 최근에 관심이 높거나 중요 쟁점이 되는 건설기술이 무엇인지를 모니터링할 수 있는 추가연구가 필요할 것으로 사료된다. K

• 정성윤, 김남곤. (2022), "네트워크 중심성을 이용한 건설 사업 리스크 관련 키워드 동향 분석", 한국산학기술학회 논문지, 23(8), pp.190–201.