

건설기술 분야의 연구개발 동향 분석

정성윤*, 김남곤*

*한국건설기술연구원 미래융합연구본부
e-mail:syjeong@kict.re.kr

Analysis of R&D trends in construction technology -Focused on the R&D reports of KICT-

Seong-Yun Jeong*, Nam-Gon Kim*

*Department of Future Technology and Convergence Research,
Korea Institute of Civil Engineering and Building Technology

요약

본 논문은 현재까지 추진하였던 건설기술 분야의 연구 개발 사례를 분석하여 앞으로 어떤 기술이 연구개발에 유용하게 될 것인가를 예측하고자 하였다. 이를 위해 국내 건설기술 분야의 연구 개발을 전문으로 하는 한국건설기술연구원에서 발간한 900여권의 연구보고서의 제목을 수집하였다. 수집한 보고서의 제목을 대상으로 LDA 기법과 word2vec 기법 등을 적용하여 기간에 따라 자주 사용된 건설기술관련 주제어와 연관관계를 갖는 단어 간의 유사도를 계산하였다.

1. 서론

한국건설기술연구원(KICT)은 건설기술 분야의 전문연구 기관으로서 건설 및 국토관리 분야의 원천기술 개발과 성과 확산을 통해 건설산업 발전과 국민 삶의 질 향상 및 국가경제 · 사회 발전에 기여함을 목적으로 하여 1988년에 (구)건설부 산하 정부출연 연구기관으로 승계하였다[1]. KICT는 국가기반시설 성능 고도화, 국토재해 대응, 친환경 국토조성, 건설기반 융·복합 등의 연구 개발을 비롯하여 정책수립 및 기술지원, 품질 인증 및 시험업무 등의 임무를 수행하고 있다[1]. KICT는 1984년부터 2019년까지 약 1,500여권의 연구개발 보고서를 포함하여 수탁 및 기타 보고서 등 약 6,980여권의 보고서를 발간하는 등 국내의 건설기술 분야의 연구 개발을 대표하는 기관이라 말할 수 있다. 본 연구는 한국건설기술연구원에서 발간한 연구보고서가 국내 공공분야에서 추진된 건설기술관련 연구 개발의 산실이라 판단하였다. 따라서 연구보고서에 사용된 단어를 분석하여 앞으로의 건설기술의 발전 동향을 추측하고자 하였다. 이를 위해 연구개발 보고서 제목을 대상으로 하여 불용어 제거, 품사 분석, 단어발생 빈도 분석 등 텍스트 전처리 과정을 거친 후 잠재 디리클레 할당(Latent Dirichlet Allocation) 기법을 이용한 주제어(토픽) 모델링과 word2vec 기법을 이용한 보고서 제목에서 빈번하게 사용된 건설기술관련 단어와 인접하여 출현한 단어 간의 관계를 통계학적 계산과정을 통해 기간에 따라 건설기술관련

주제어의 변화과정으로 분석하였다. 끝으로 본 연구 결과가 갖는 제약 사항과 내실 있는 연구결과를 얻기 위해 향후 추가적으로 필요한 연구 방향을 제시하였다.

2. 건설기술관련 주제어 분석

본 연구는 한국건설기술연구원의 전자도서관에서 서비스 중인 KICT 보고서 중 기본(연구개발) 보고서의 제목과 발간 연도를 수집하였다. 건설기술관련 주제어를 기간별로 분석하기 위해 연차별로 발간된 보고서는 하나로 취합하였다. 취합한 918권의 보고서 제목을 띄어쓰기를 기준으로 하여 단어를 추출하였다. 이 과정에서 건설기술과 관련성이 낮은 특수문자, 문자부호, 숫자, 영단어 등을 불용어로 처리하였다. 이렇게 추출한 단어를 대상으로 Konlgy 패키지를 적용하여 단어의 품사를 정하였다. 명사로 간주될 수 있는 단어를 제외한 관용사, 형용사, 부사 등의 품사에 해당하는 단어를 불용어로 간주하여 분석 대상에서 제거하는 등의 정규화 작업을 하였다. 더불어 조사, 연구, 개발, 기초 등 건설기술과 직접적으로 연관성이 낮은 단어를 필터링하여 분석 대상에 사용되는 군집 단어집에서 제외시켰다. 이러한 군집 단어집용 전처리 과정을 거쳐서 나온 단어를 가지고서 건설기술 동향을 분석하였다. 그림 1은 기간별로 자주 나타난 상위 10위의 단어의 사용 빈도를 도식화하여 나타낸 것이다. 그림 1에서 보듯이 2000년 이전에는 “설계” 단어가 가장 빈번하게 등장하였고 성능은



[그림 1] 기간별 사용단어의 빈도

상대적으로 빈도수가 적었다. 이는 건설기술에 대한 연구개발이 기초적인 수준에 있었다고 추측할 수 있다. 하지만 2000년에 들어오면서 성능이 강조된 것을 알 수 있었다. 이는 1990년대에 추진된 건설기술의 설계가 구현되면서 성능에 대한 연구개발이 활발해진 것으로 해석할 수 있다. 2010년대에는 성능이 확보된 건설기술을 보급, 확산하기 위한 실용화에 대한 관심이 높은 것으로 판단된다. 이처럼 보고서 제목에 사용되는 단어의 빈도를 가지고서 건설기술 변화과정을 간접적으로 알 수 있다. 하지만 단어의 빈도만을 가지고서 기술 변화를 추측하는 것에는 한계가 있다. 그래서 보고서 제목에서 연관성을 갖는 주제군과 잠재된 제어의 가중치를 추론하는 LDA 기법을 적용하였다. 그 결과로 주제군을 4개로 하였을 경우에 혼란도(perplexity)로는 -8.66을 얻었고, 주제 연관성(topic coherence)은 0.729로 계산되었다. 여기서 혼란도는 토픽 모델이 실제로 관측된 값을 얼마나 잘 측정하였는가를 평가하는 지표로서 수치가 낮을수록 언어 모델의 성능이 좋고 해석할 수 있다. 주제 연관성은 토픽 모델에서 특정 주제군에 유사한 의미를 갖는 단어가 얼마나 잘 모였는지를 평가하는 지표로서 수치가 높을수록 의미론적 일관성이 높다고 측정한다. 따라서 이들 수치는 주제군과 제어어가 적절하게 반영된 것으로 판단할 수 있다. 또한 주제 연관성의 그래프에서 주제의 개수가 3에서 변동하기 시작하였다. 따라서 본 연구에서의 주제군으로는 3개를 사용하는 것이 바람직한 것으로 분석되었다. 다음으로 word2vec 라이브러리를 사용하여 제목에서 불용어로 제거된 두개 단어들 간의 코사인 유사도(Cosine Similarity)와 가중치를 계산하였다. 표 1은 KICT의 보고서와 국토교통과학기술진흥원(KAIA)[2]에서 발간한 5,316권의 보고서의 제목에서 가장 빈번하게 사용된 4개의 단어와 연관되는 단어들 중 유사도가 높은 상위 4위의 단어를 추출한 결과를 비교한 것이다[3]. 여기서 유사도는 제시된 단어와 연관된 단어 간 유사도 정도를 나타내기 위한 코사인(cosine) 거리를 계산한 것이다. 표 1에서 보듯이 보고서 제목에서 빈번하게 사용된 상위 4위의 단어를 추출하였다. 순위의 차이는 있으나 KICT와 KAIA의 보고서에서 거의 동일한 단어가 빈번하게 사용되었다. 추출된 단어와 인접한 단어 간의 유사도가 높은 상위 5위인 단어의 유사도를 계산하였다. 2000년대에는 추출된 단어들이 서로 연관되었으나 2010년대에서는 요소기술로 세분화 된 단어가 많이 사용된 것으로 분석되었다.

[표 1] 상위 5위의 연관 단어와 유사도 분석

구분 순위	KICT		
	1984년~1999년	2000년~2009년	2010년~2019년
설계	토목(0.215), 섬유(0.146), 예측(0.127), 포장(0.084)	성능(0.081), 구조물(0.078), 소재(0.007), 회계(0.014)	저감(0.217), 융복합(0.198), 마이크로그리드(0.197), 가상화(0.196)
성능	방지(0.172), 연약(0.147), 수문(0.117), 건설사업(0.114)	회계(0.359), 자산관리(0.117), 소재(0.086), 설계(0.081)	수자원(0.322), 타당성(0.214), 재료(0.210), 도로시설물(0.174)
콘크리트	자동화(0.131), 에너지(0.121), 공동주택(0.112), 도시(0.110)	구조물(0.038), 소재(0.021), 바닥판(0.011), 건설공사(0.048)	내화구조(0.283), 스마트시티(0.238), 저비용(0.218), 제거기술(0.211)
구조물	홍수(0.225), 공동주택(0.223), 모형(0.212), 공간(0.155)	콘크리트(0.135), 성능(0.117), 회계(0.036), 교량(0.034)	폐기물(0.240), 홍수(0.225), 공동주택(0.223), 고효율(0.177)

구분 순위	KAIA		
	1994년~1999년	2000년~2009년	2010년~2019년
콘크리트	안전진단(0.234), 에너지(0.216), 최적화(0.184), 감시시스템(0.118)	지반정수(0.258), 공기단축형(0.245), 안전진단(0.234), 표준화(0.226)	초정밀(0.249), 국산화(0.247), 경감(0.237), 안전진단(0.234)
실용화	지하공간(0.223), 건물(0.143), 공동주택(0.132), 개발용(0.131)	기본계획(0.270), 측정(0.259), 센서(0.253), 신재료(0.246)	통합안전관리(0.273), 고수압(0.261), 측정(0.259), 센서(0.253)
설계	최적화(0.278), 철근콘크리트(0.178), 터널(0.156), 공동주택(0.124)	다중추진시스템(0.308), 최적화(0.278), 자동인식(0.267), 객차용(0.244)	실내공간정보(0.324), 최적화(0.279), 지반함몰(0.278), 자원순환체계(0.261)
구조물	건설공사(0.151), 시공(0.143), 안정성(0.137), 터널(0.130), 설계(0.065)	공사(0.314), 계차(0.237), 한국형(0.235), 고성능(0.195), 예측(0.192)	건설현장(0.293), 곡선(0.278), 구조용(0.275), 도로시스템(0.272), 인간(0.268)

3. 결론

본 연구는 한국건설기술연구원에서 발간한 연구보고서의 제목을 대상으로 하여 건설기술 분야와 관련한 연구개발 동향을 파악하고자 하였다. 이를 위해 텍스트 마이닝에서 사용하는 불용어 전처리 과정, LDA 기법과 word2vec 기법을 적용하여 보고서 제목에서 빈번하게 사용된 4개의 제어어와 연관어들 간의 유사도를 계산하였다. 본 연구 결과가 보다 객관적이고 합리성을 갖기 위해서는 요약문과 본문까지의 분석 범위를 확대해야 하고, 연구개발 당시의 사회적 요구(need)에 대한 영향 등도 함께 연구할 필요가 있을 것으로 사료된다.

참고문헌

- [1] 한국건설기술연구원, "https://www.kict.re.kr"
- [2] 국토교통과학기술진흥원, "https://www.kaia.re.kr"
- [3] 정성윤, 김남곤, "텍스트마이닝을 이용한 국토교통 R&D 기술동향 분석" 한국스마트미디어학회 2019년도 추계학술대회, Vol.8, No.1, pp.245-247, 11월, 2019년.