CONCOR 중심성을 이용한 건설기술 키워드 가중치 분석

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

Construction Technology Keyword Weighting Analysis Using the CONCOR Centrality

Seong-Yun Jeong*
*Dept. of Future & Smart Construction Research,
Korea Institute of Civil Engineering and Building Technology

요 약

본 논문은 건설기술정보시스템에서 서비스되는 건설기술정보에 내재한 유의미한 키워드가 건설기술 네트워크에서 얼마나 영향력을 갖는지를 확인하고자 하였다. 이를 위해 불용어 전처리 작업, LDA 분석 및 토픽 모델링 등의 과정을 통해 건설기술에 영향력이 있는 키워드를 추출하였다. 추출한 키워드 간의 구조적 등위성 기반의 중심성과 상관계수를 추정하여 묶은 4개의 블록을 생성하였다. 생성한 블록을 가지고 가중치가 높은 50개의 키워드를 표시한 네트워크와 키워드의 가중치를 추정하였다.

1. 서론

건설기술정보시스템은 건설기술과 관련한 서지정보와 원문 자료를 서비스하는 시스템이다. 서지정보는 해당 건설기술정보에 대한 의미를 함축적으로 내포하고 있다. 이러한 내포된 정보에서 영향력이 높은 키워드를 뽑을 수 있다고 판단하였다. 이를 위해 건설기술정보시스템에서 서비스하는 30,045건의 건설기술정보를 수집하였다[1]. 수집한 정보는 분석에서 필요하지 않은 불용어 처리 등 전처리 작업을 진행하였다. 잠재 디리클레 할당(Latent Dirichlet Allocation) 분석 기법과 토픽 모델링을 통해 유사한 의미를 갖는 키워드를 하나의 그룹으로 군집화한 토픽을 생성하였다. 토픽에 포함된 키워드가 네트워크에서 얼마나 영향력을 갖는지를 추정한다, 이때 많이 사용하는 중심성 알고리즘으로는 단계(Degree), 근접(Closeness), 매개(Between), 아이겐벡터(Eigenvector) 등이 대표적이다. 본 연구는 이들 중심성 이외에 구조적 등위성(structural equivalence)을 측정하여 군집에 내포된 특성 또는 의미를 추정하는 CONCOR 중심성 알고리즘을 건설기술 정보에 내포된 유의미한 정보를 분석하는 데 활용할 수 있는지를 확인하고자 하였다. 이를 위해서 전처리 과정을 통해 얻은 어휘를 기초 데이터로 하여 CONCOR 중심성에 적용하여 가중치가 상위 50위 안에 포함된 키워드를 중심으로 한 네트워크를 표시하였다. 다음으로, 4개의 블록에 포함된 가중치가 높은 키워드를 추정하였다. 블록에서 상위를 차지한 키워드는 주로 도로와 시설물의 건설공사와 관련한 키워드가 네트워크의 중심에서 영향력을 갖는 것으로 유추할 수 있었다.

2. 이론적 고찰

텍스트 마이닝(Text Mining)은 특정 콘텐츠를 구성하는 어휘의 품사와 이웃 어휘 간의 선·후행 관계, 어휘의 발생빈도 등을 통계적으로 분석하여 유의미적인 키워드를 유추하는 이론을 말한다. 잠재 디리클레 할당 알고리즘을 이용하여 유사한 의미를 갖는 키워드를 묶은 토픽 모델링 과정과 토픽 모델링을 통해 키워드 간의 연관관계를 나타내는 네트워크 이론(Network Theory)을 사용한다. 이때 키워드 간의 네트워크는 꼭짓점과 연결선 등으로 이용하여 키워드 간의 연결 관계를 표현한다. 보통 네트워크 중심성 알고리즘으로 중심성 계산식에 따라 단계, 근접, 매개, 아이겐벡터를 가장 많이 사용하는데 네트워크에 있는 키워드의 가중치를 직접 산출한다[2]. 반면에, CONCOR 중심성은 키워드 간 중심성과 공동 출현 행렬을 반복적인 피어슨 상관계수를 수렴하여 유사한 의미를 갖는 키워드를 서로 묶어준 군집을 찾는다[3]. CONCOR 중심성은 기존에 키워드 간의 연결 개수, 중개 역할 등 키워드의 영향력을 기반으로 하지만 CONCOR 중심성은 유사한 의미를 갖는 군집을 이용한다는 점이 차이가 있다.

3. 네트워크 모델링

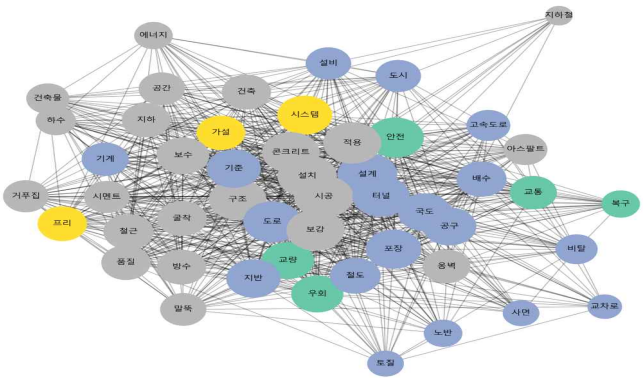
CONCOR 중심성 기반의 네트워크를 모델링하기에 앞서 정성윤[1]에서 수집한 건설기술정보를 분석 대상 데이터로 사용하였다. 전처리 작업을 통해 추출한 어휘 간의 선·후행 발생 관계와 발생빈도를 추정하였다. 주어진 토픽의 개수에 따라 분석 자료에서 일정한 패턴 구조를 갖는 단어 간의 잠재적 관계를 추정하는 잠재 디리클레 할당을 분석하였다. 본 연구에서는 임의로 5개의 모델링(Topic Modelling) 과정을 통해 유의미한 어휘라고 판단하여 토픽별로 군집 형태로 묶었다. 표 1은 5개의 토픽에 포함된 상위 10위까지의 키워드와 가중치를 나타낸 것이다.

[표 1] 토픽별 키워드의 가중치

토픽번호	키워드별 가중치
1	콘크리트(0.04), 보강(0.02), 보수(0.02), 구조(0.02), 방수(0.02), 시공(0.01), 아스팔트(0.01), 적용(0.01), 철근(0.01), 말뚝(0.01)
2	도로(0.20), 포장(0.12), 우회(0.05), 안전(0.04), 지반(0.02), 교통(0.02), 토질(0.02), 비탈(0.02), 연약지반(0.01), 터널(0.01)
3	설계(0.18), 교량(0.07), 공구(0.04), 기준(0.03), 철도(0.02), 설치(0.02), 안전(0.01), 흙막이(0.01), 에너지(0.01), 도시(0.01)
4	국도(0.07), 시공(0.05), 설치(0.03), 거푸집(0.02), 콘크리트(0.02), 설비(0.02), 가설(0.01), 강관(0.01), 프리(0.01), 배관(0.01), 시스템(0.05)
5	터널(0.04), 배수(0.02), 발파(0.02), 설치(0.01), 굴착(0.01), 질소(0.01), 미생물(0.01), 유입(0.01), 오염(0.01)

4. CONCOR 중심성 기반의 네트워크 분석

앞에서 전처리 작업과 토픽 모델링을 통해 얻는 키워드가 네트워크에서 어느 위치의에 분포하고 있는지를 분석하기 위해 CONCOR를 이용한 네트워크 중심성을 분석하였다. 보통 영향력이 큰 키워드일수록 네트워크 중심에 위치하고 연결선이 많으며 노드의 크기를 크게 표시하여 시각적으로 표현한다. 그림 1은 CONCOR 중심성 가중치가 상위 50위 안에 포함된 키워드로 구성한 네트워크를 표시한 것이다.



[그림 1] CONCOR 중심성 기반의 네트워크 시각화

이 그림에서 보면 설계, 도로, 구조, 시공, 포장, 터널, 콘크리트, 보상, 지반, 기준 등의 키워드가 네트워크 중심에 위치하였다. 표 2는 CONCOR 중심성 기반의 블록에 포함된 가중치가 높은 키워드에 대한 가중치를 추정한 결과이다. 블록 1과 2는 도로와 도로 건설공사와 관련한 키워드가 네트워크의 중심에 위치하여 영향력을 갖는 것으로 추정하였다. 블록 3과 4는 주로 시설물 건설공사에 사용하는 키워드가 네트워크에서 중요한 역할을 하는 것으로 판단할 수 있다. 이러한 CONCOR 중심성 기반의 키워드 가중치는 LDA 기반의 키워드 가중치와 겹치는 부분이 많았다. 이는 CONCOR 중심성은 건설기술 키워드 분석에 적용할 수 있다고 유추할 수 있다.

[표 2] CONCOR 중심성 키워드의 가중치

블록번호	키워드별 가중치
1	교량(0.76), 교통(0.62), 복구(0.41), 안전(0.9), 우회(0.76)
2	고속도로(0.53), 공구(0.80), 교차로(0.33), 국도(0.80), 기계(0.61), 기준(0.82), 노반(0.41), 도로(0.94), 도시(0.57), 배수(0.69), 비탈(0.49), 사면(0.37), 설계(1.00), 설비(0.57), 지반(0.82), 철도(0.71), 터널(0.92), 토질(0.37), 포장(0.86)
3	가설(0.65), 시스템(0.84), 프리(0.67)
4	거푸집(0.57), 건축(0.67), 건축물(0.51), 공간(0.59), 구조(0.92), 굴착(0.67), 말뚝(0.59), 방수(0.67), 보강(0.94), 보수(0.76), 설치(0.98), 시공(0.98), 시멘트(0.57), 아스팔트(0.53), 에너지(0.4), 옹벽(0.63), 적용(0.94), 지하(0.65), 지하철(0.20), 철근(0.69), 콘크리트(0.92), 품질(0.67), 하수(0.45)

5. 결언

LDA 분석과 토픽 모델링을 적용하여 건설기술정보시스템에서 서비스되는 건설기술정보에 내재한 유의미한 키워드를 찾았다. CONCOR 중심성을 이용하여 찾은 키워드가 건설기술 네트워크에서 어느 위치에 차지하는지를 표현하였다. 또한, 키워드의 영향력을 정량적 가중치로 추정하였다.

참고문헌

[1] 정성윤, “미래 예측 모형을 활용한 건설기술 동향 예측에 관한 연구”, 산학기술학회논문지, 제26권 12호, pp. 964-973, 12월, 2005년.

[2] 정성윤, “텍스트 마이닝과 네트워크 이론을 활용한 권역별 국가하천 점용허가 키워드 분석”, 스마트미디어저널, 제12권 11호, pp. 185-197, 2월, 2023년.

[3] 김후정, 김성희, “키워드 네트워크 분석을 통한 문헌정보학 분야 생성형 AI에 관한 연구 동향 분석”, 한국비블리아학회지 제36권 3호, 227-245, 9월, 2025년.

※ 이 논문은 2026년도 정부(국토교통부)의 출연금으로 한국건설기술연구원이 수행한 “26년 건설기술정보 DB 및 서비스 시스템 운영” 과제의 연구 결과임