텍스트마이닝을 활용한 건설실무정보의 특성 분석*

- 건설기술, 사례, 원가절감 등 정보를 중심으로 -

Analysis on the Characteristics of Construction Practice Information Using Text Mining: Focusing on Information Such as Construction Technology, Cases, and Cost Reduction

정 성 윤 (Seong-Yun Jeong)**
김 진 욱 (Jin-Uk Kim)***

목 차

초 록

본 연구는 전문지식을 갖지 않은 건설기술자와 건설사업 참여자가 건설 실무에서 중요도가 높은 단어와 단어 간의 상호 연관관계를 쉽게 이해할 수 있도록 정보서비스를 개선하고자 하였다. 이를 위해 텍스트마이닝과 네트워크 중심성을 이용하여 건설기술정보시스템에서 가장 많이 사용하고 있는 기술정보, 사례정보 및 원가절감 등 건설실무정보에 대해 단어의 출현 빈도, 주제 모형화, 네트워크 중심성을 분석하였다. 이러한 분석을 통해 도로, 포장, 교량, 터널 등 도로공사와 관련한 설계, 시공, 사업관리, 시방・기준, 유지관리 등이 건설 실무에서 중요한 정보로 파악되었다. 또한, 연결 중심성과 고유벡터 중심성 측정을 통해 중요도가 높은 단어 간의 상관도를 분석하였다. 상관도 분석을 통해 기술정보를 확충한다면 보다 유용한 정보를 제공할 수 있다는 결과를 얻었다. 끝으로, 연구 결과가 갖는 제약과 이에 따른 추가적인 연구를 제시하였다.

ABSTRACT

This study aims to improve the information service so that construction engineers and construction project participants without specialized knowledge can easily understand the important words and the interrelationships between them in construction practice. To this end, using text mining and network centrality, the frequency of occurrence of words, topic modeling, and network centrality in construction practice information such as technical information, case information, and cost reduction, which are most used in the Construction Technology Digital Library, were analyzed. Through this analysis, design, construction, project management, specifications, standards, and maintenance related to road construction such as roads, pavements, bridges, and tunnels were identified as important in construction practice. In addition, correlations were analyzed for words with high importance by measuring Degree Centrality and Eigenvector Centrality. The result was that more useful information could be provided if the technical information was expanded. Finally, we presented the limitations of the study results and additional studies according to the limitations.

키워드: 건설실무정보, 주제 모형화, 네트워크 중심성, 상관관계 분석, 건설기술정보시스템
Construction Practice Information, Topic Modeling, Network Centrality, Correlation Analysis, Construction Technology Digital Library

---

* 본 논문은 교육과학기술부의 재원으로 한국건설기술연구원 “(22주요-대1-목적)미래 건설산업 견인 및 신시장 창출을 위한 스마트 건설기술 연구 (2/2)” 과제와 국토교통부 출연사업인 “22 건설기술정보 DB 및 서비스시스템 운영” 과제의 지원을 받아 수행되었음.

** 한국건설기술연구원 연구위원/공학박사(syjeong@kict.re.kr / ISNI 0000 0004 7640 9363) (제1저자, 교신저자)
*** 한국건설기술연구원 연구위원(jukim@kict.re.kr / ISNI 0000 0004 6511 8106) (공동저자)

논문접수일자: 2022년 10월 18일 최초심사일자: 2022년 11월 13일 게재확정일자: 2022년 11월 24일
한국문헌정보학회지, 56(4): 205-222, 2022. http://dx.doi.org/10.4275/KSLIS.2022.56.4.205

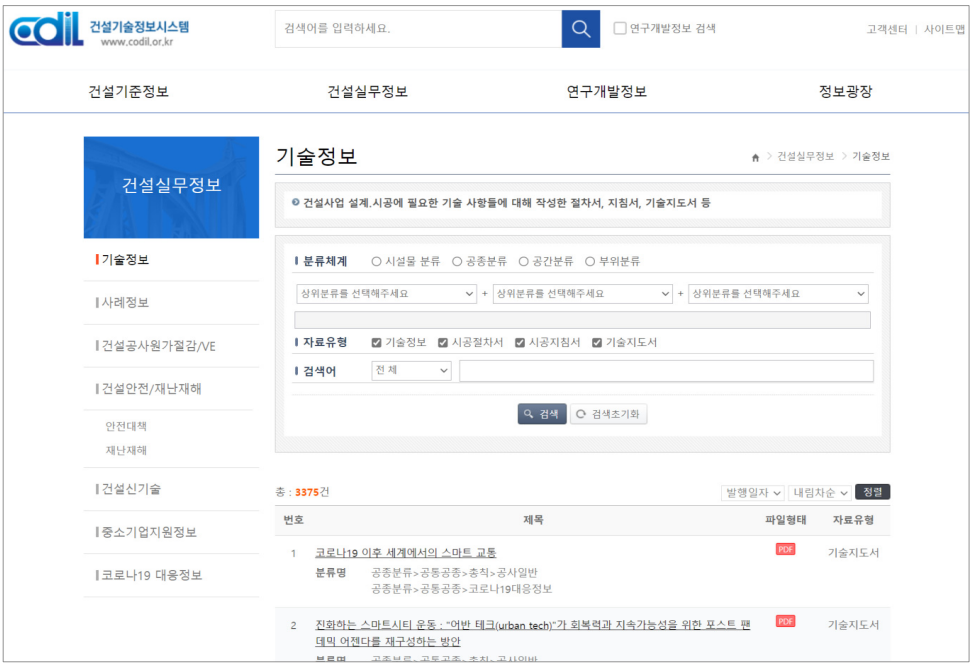
1. 서 론

국내 건설산업은 2000년대에 들어오면서 건설공사의 규모가 대형화, 복잡화되고 있다(KDI 공공투자관리센터, 2021). 최근에는 발주자가 친환경성과 안전성 및 스마트 건설 기반의 편의성을 강조하면서도 건설사업비 감축을 요구하고 있다. 이러한 발주자의 요구사항을 충족시킬 수 있는 새로운 건설기술 개발이 뒷받침되어야 한다. 하지만 현실적으로 중소·중견 건설업체와 엔지니어링업체가 건설기술 개발에 투자하기란 쉽지 않다. 정부는 발주자의 요구사항을 충족시키기 위해서는 국내 건설업체와 엔지니어링업체의 기술경쟁력 확보가 선행되어야 한다고 판단하였다. 이를 위해 건설기술 자료를 누구나 쉽게 획득하고, 활용할 수 있도록 「건설기술진흥법」 제18조(건설기술정보체계의 구축)를 마련하였다. 이 법 조항에 따라 한국건설기술연구원은 건설 현장에서 필요로 하는 각종 건설보고서, 건설공사 실무자료, 건설공사 기준, 품셈 등 각종 건설기술 정보를 건설기술자가 쉽고, 편리하게 이용할 수 있도록 DB로 구축한 후, 건설기술정보시스템(Construction Technology Digital Library, CODIL)을 통해 서비스하고 있다(건설기술정보시스템, 2022). 하지만 건설기술 정보를 DB로 구축하기 위한 예산은 한정되어 있으나, 건설 현장에서는 최신의 건설기술과 활용사례 등 새로운 자료를 확충할 것을 요구하고 있다. 하지만 한정된 예산과 자료의 소재 파악이 쉽지 않다. 비록 소재 파악이 되었더라도 자료 공개를 꺼리는 경향이 심화되고 있다. 이런 이유로 자료의 확충에 대한 건설 현장의 요구사항을 충족시키는 데 한계가 있다. 특히, 전문지식을 갖지 않은 건설기술자와 건설사업 참여자가 서지정보만으로 건설실무정보를 검색하고 조회하는 것은 결코 쉽지 않다. 따라서 현재 CODIL에 축적된 정보에서 건설 현장에 유용하게 사용할 수 있는 정보를 추출하는 방안을 모색하게 되었다. 이러한 정보서비스 개선책으로, CODIL에서 가장 많이 조회되고 있는 기술정보, 사례정보 및 원가절감 등 건설실무정보를 대상으로 하여 서지정보 중 비정형 텍스트로 작성된 자료 제목에 내재한 암묵적인 의미를 갖는 정보를 추출하고자 하였다. 추출한 정보에서 영향력 있는 단어 간의 의미가 있는 맥락을 분석하고 이를 시각화하고자 하였다. 이러한 의미가 있는 맥락 분석과 시각화는 전문지식을 보유하지 않은 건설기술자와 건설사업 참여자에게 건설 실무에서 중요도가 높은 단어가 무엇이고, 이들 단어 간의 상호 연관관계를 직감적으로 이해하는 데 도움을 줄 수 있다고 판단하였다. 더불어, 기술정보, 사례정보 및 원가절감 간의 상관 정도를 분석하여 이들 정보 중 허브 역할을 하는 정보가 무엇인지를 파악하고자 하였다. 상관도가 높은 정보를 선택하여 자료 확충과 서비스 강화를 위한 노력과 예산 배정을 집중한다면 현재보다 유용한 정보서비스가 가능할 것으로 생각하였다. 이러한 목적을 실천하기 위해 텍스트마이닝과 네트워크 이론에 근간하여 기술정보, 사례정보 및 원가절감의 자료 제목에 포함된 단어의 출현 빈도, 주제 모형화, 네트워크 중심성 계수 측정 및 상관관계 분석 등 흐름으로 연구를 진행하였다. 특히, 회귀모형과 상관도 분석을 통해 연결 중심성과 고유벡터 중심성에서 중요도가 높은 단어를 기초로 하여 기술정보,

사례정보 및 원가절감 간의 상관도를 분석하였다. 기술정보와 사례정보 간의 상관도가 높았지만, 원가절감과는 상대적으로 상관도가 낮은 것으로 파악되었다. 따라서 기술정보를 집중적으로 확충한다면 보다 유용한 정보서비스가 될 것이라는 결과를 얻었다. 끝으로, 본 연구 결과가 갖는 제약과 이에 따른 추가적인 연구의 필요성을 제시하였다.

2. 건설실무정보의 특성 조사

한국건설기술연구원은 국내 중소·중견 건설 및 건설엔지니어링 업체의 기술경쟁력을 강화하는 목표로 CODIL을 구축하였다(건설기술정보시스템, 2022). CODIL에서 서비스되는 건설기준정보, 건설실무정보, 연구개발정보 및 정보광장 중 건설실무정보에는 〈그림 1〉과 같이 기술정보, 사례정보, 원가절감, 건설안전/재난재해, 건설 신기술, 중소기업지원정보 및 코로나19 대응 정보 등을 포함하고 있다. CODIL에서 주로 다루는 건설은 도로, 교량, 터널, 댐, 건축물 등 시설물을 새로 만드는 과정과 결과물을 말한다. 보편적으로, 하나의 시설물을 건설하기 위해 수년에서 수십 년에 걸쳐 발주자, 설계자, 시공자, 감리자, 인허가자 등 다양한 주체들이 참여한다. 이들 주체 간에 정보의 흐름을 원활하게 관리하고, 계획에 따라 시설물을 만들기 위한 공사와 관련한 공정, 공법, 품질, 원가, 자재, 안전/위험 등 요소를 관리할 수 있는 사업관리(project management)가 필요하다. 건설실무정보는 이러한 사업관리에 필요한



〈그림 1〉 건설실무정보 조회화면 예

정보를 의미한다. 즉, CODIL에서 의미하는 건설실무정보는 사업관리를 통해 산출되는 각종 자료와 정보 중에서 사업관리에 참여하는 사업관리 실무자 또는 건설기술자에게 도움을 줄 수 있는 계획서, 절차서, 기술개발・적용 사례, 건설공사 원가절감사례 등 자료와 정보를 말한다. 건설실무정보 중 기술정보는 시설물을 새로 만들거나 보수・보강하는데 필요한 기술적 기준, 방법 또는 공사 시방을 기술한 각종 절차서, 지침서, 기술 지도서 등을 일컫는 정보와 자료를 말한다. 사업관리 실무자는 하나의 시설물을 완료하는 시점에서 시설물을 만드는 과정에서 작성하였거나 참조하였던 각종 사업관리 자료와 정보를 모아서 준공 도서로 만든다. 사례정보는 작성한 준공 도서 중 다른 건설 현장에서 참조하는데 유용한 각종 시공계획서, 현장 시공 사례, 건설공사지 등 모아둔 사례 위주의 정보와 자료를 의미한다. 끝으로, 원가절감은 2014년부터 한국도로공사의 건설공사 원가절감사례 자료를 비롯하여 하나의 시설물을 만들거나 유지관리하는 과정에서 발주자, 사업관리 실무자 또는 건설기술자가 고안한 새로운 공법, 품질 향상, 공사 기간 단축 등을 통해 건설 비용이나 유지관리 비용을 절감한 사례를 기록한 자료와 정보를 의미한다. 현재 건설기술자와 건설사업 실무자가 CODIL에서 서비스되고 있는 정보 중 기술정보, 사례정보, 원가절감을 가장 많이 조회하고 있다.

3. 선행 연구사례 고찰

본 연구는 새로운 이론을 개발하거나 실증적 분석기법을 고안하기보다는 현재 사용되는 텍스트마이닝 기법 중 주제 모형화와 네트워크 중심성 및 상관도 분석 을 사용하여 건설실무정보에 내포된 암묵지적인 정보의 특성을 파악하고, 이들 정보 간의 상관관계를 분석하는 응용연구라고 할 수 있다. 따라서 본 연구에서는 이론적 고찰보다는 유사한 연구사례를 살펴보고자 하였다. 선행 연구사례를 찾기 위해 네이버와 구글의 학술정보 검색사이트에서 시설물을 만드는데 필요한 사업관리에 관한 의미로서의 건설실무정보와 텍스트마이닝, 네트워크 중심성 및 상관도 분석을 조건어로 지정하여 검색하였다. 여기서 텍스트마이닝은 비정형 텍스트나 데이터로부터 통계학적 계산과정을 통해 의미가 있는 특성이나 맥락을 찾는 알고리즘을 말한다. 한편, 네트워크 중심성은 말뭉치에 포함된 단어의 동시 출현 빈도 등 통계학적으로 계산한 후에 단어와 인접한 단어 간의 연결 관계를 네트워크 형태로 만든 그래프를 말한다. 중심성 계수는 네트워크에서 특정 단어가 갖는 중요 정도 또는 영향 정도를 정량적 수치로 나타낸 것이다. 건설실무정보를 대상으로 하여 텍스트마이닝, 네트워크 중심성 또는 상관도를 분석한 국내 연구사례를 찾지 못하였다. 다만, 건설과 관련해서는 박준용(2021)은 2017년부터 2021년까지 436편의 대한토목학회 논문 제목을 분석하여 유사한 의미를 갖는 주제 단어를 추출하는 방식으로 주제 모형화를 분석하였다. 이때 주제 일관성 지수가 가장 높은 14개를 주제 개수로 결정하였다. 김재준, 정절우(2012)는 단어 빈도-역문서 빈도(Term Frequency-Inverse Document Frequency)와 TF-IDF를 응용한 단어 빈도-데이터 색인(Term Frequency-

Data Index)을 사용하여 2010년에 발표한 「과학기술 미래비전」 보고서 내용 중 건설 분야와 관련된 기술에 대해 동향을 분석하였다. 여기서 단어 빈도-역문서 빈도(TF-IDF)는 특정 단어의 가중치(weight)를 구하기 위해 여러 문서를 모아 놓은 문서군에서 특정 단어가 특정 문서 내에서 얼마나 중요한지를 계산한 통계적 수치를 말한다. 단어 빈도-데이터 색인(TF-DI)은 TF-IDF가 갖는 복잡한 계산방식 및 데이터 추출하는 방법과 범위에 따라 오차율을 개선하기 위해 고안한 알고리즘으로서 특정 단어의 가중치를 문서군에서 계산하지 않고 인터넷의 정보량에 따라 단어의 빈도수를 분석하는 방식을 의미한다. 다음으로, 성현곤 외(2019)는 최막중이 1994～2019년 동안 발표한 133편의 논문 제목을 대상으로 단어출현 빈도, 동시 단어출현, 13개의 주제 모형화, 시기별 연구주제어의 변화추이 등의 분석을 통해 국토 및 도시계획에 관한 연구 경향을 파악하였다. 여기서 국토 및 도시계획에 관한 연구는 최막중 교수가 약 30년 동안의 학술지에 게재한 논문(총 133편, 1994년～2019년) 중 국토 및 도시계획 관련 논문을 말한다. 정근하(2010)는 TF-IDF와 연결, 근접(closeness), 매개(between) 등 중심성 분석을 통해 건설 분야의 미래기술을 예측하였다. 이때 연결, 근접, 매개 등 중심성 분석은 행위자를 표현하는 노드와 상호 연관관계를 의미하는 링크로 구성한 네트워크에서 어떤 노드가 연결된 인접 노드를 찾아가는 경로 탐색 방식에 따라 중심성을 분석하는 방법을 의미한다. 오준석(2015)은 4개의 국내 학술지에서 발표한 2,806편 교통 분야 논문과 363편의 ICT 관련 교통 분야의 논문을 대상으로 단어출현 빈도, 동시 출현 단어, 10개의 주제 모형화 및 연도별 주제 빈도수의 변화를 통해 국내 교통・ICT 융합 분야 연구의 특성을 분석하였다. 여기서 ICT 관련 교통 분야는 정보통신기술(Information Communication Technology)을 접목한 교통 흐름제어 시스템, 교통정보서비스와 같은 정보기술 기반의 교통 서비스를 말한다. 최정묵(2016)은 연결, 고유벡터, 베타(beta) 등의 중심성을 측정하여 행정학・정책학 관련 학술지의 상호인용을 분석하였다. 여기서 연결, 고유벡터, 베타(beta) 등의 중심성은 어떤 노드가 연결된 인접 노드의 가중치를 고려하여 네트워크에서 중심적 영향력 또는 중요도를 계산한 알고리즘을 말한다. 임병학, 전희주(2011)는 항만의 사회 네트워크가 물동량에 미치는 영향을 분석하기 위해 연결, 근접, 매개, 고유벡터 등의 중심성 계수를 측정하였고, 이들 중심성 계수 간에 상관관계를 분석하였다. 이강원, 이정원(2017)은 서울 수도권 지하철망에 대한 특성을 분석하기 위해 매개 중심성과 가중된(weighted) 매개 중심성 간의 상관관계를 분석하였다. 여기서 가중된 매개 중심성은 출발역과 도착역에서 승객의 유동 인력수를 반영하여 출발역과 도착역 간의 최단 경로에 얼마나 많이 존재하는가를 나타내는 지표를 말한다. 이처럼 대부분의 연구에서 분석용 자료를 수집하고, 불용어 제거와 형태소 분석 등 전처리 과정을 거친 후, 단어/단어-문서 간의 출현 빈도, 주제 모형화 등을 분석하였다. 한편, 일부 연구에서는 네트워크 중심성을 분석하여 상관관계나 시간이 흐르면서 나타나는 시계열 기반의 변환 추이를 분석하였다. 이러한 분석을 통해 기술 동향, 미래 예측, 주제 분석, 정보 특성,

우선순위 선정 등 연구 결과를 제시하였다.

4. 주제 모형화 및 네트워크 중심성 분석

4.1 말뭉치의 특성

본 연구는 2022년 9월 15일 기준으로 CODIL에서 서비스되고 있는 기술정보, 사례정보 및 원가절감에 관한 더블린 코어 메타데이터(Dublin Core Metadata) 기반의 서지정보를 수집하였다. 더블린 코어 메타데이터는 디지털화된 정보나 자료를 효율적으로 검색하고 관리하기 위해 국제표준기구(International Standardization Organization)에서 제정한 메타데이터 요소 집합에 관한 표준을 일컫는다. 수집한 서지정보를 〈표 1〉과 같이 연도별로 분류하였다. 서지정보는 크게 자료 제목, 주제분류, 저장 및 발행처, 작성 언어, 발행일 및 원문 보기 및 저장 등 항목으로 구성되어 있다. 이 서지정보 중 분석 작업에 크게 도움이 되지 않는 주제 분류, 저장 및 발행처, 작성 언어 등 항목과 원문 자료가 빠진 경우가 있어 분석 대상에서 제외하였다. 자료 내용을 함축적으로 표시하는 제목과 발행일을 추출하여 말뭉치(Corpus)로 사용하였다.

4.2 말뭉치의 전처리 작업

말뭉치 내용 중 중요도가 낮은 단어를 분석 작업에서 제외하기 위해 길호현(2019)의 한국어 전처리 과정을 준용하여 다음과 같이 불용어 제거와 형태소를 분석하였다. 먼저, 말뭉치에서 사용되는 단어의 형태소를 개별적으로 식별할 수 있는 토큰(token)으로 구분하였다. 구분한 토큰 중에서 독립적으로 의미를 갖지 않는 특수문자와 숫자를 불용어로 처리하였다. 두 번째로, 토큰 중에서 명사는 하나의 독립적 의미를 갖는 실질 형태소의 속성을 가지고 나머지 품사는 의존 또는 형식 형태소를 가진다. 따라서 말뭉치에서 명사에 해당하는 단어를 추출하기 위해 파이썬에서 사용하는 konlpy 패키지를 사용하였다. 여기서 konlpy 패키지는 한국어 말뭉치를 처리를 위해 설계된 파이썬용 한국어 분석기를 말한다. 한국어 분석기와 영어 분석기를 동시에 사용한 사례를 찾지 못하였다. 따라서 konlpy 패키지는 한글의 형태소를 추출하도록 개발되었기 때문에 영어단어에 대해 품사를 구분하지 못하는 한계가 있다. 다만, 본 연구에서 사용한 말뭉치에서 포함된 영어단어는 주로 RC, CALS, CCTV, UHPC, HP-41/CV, P.S, EPS, Smart-Rail, Geotextile과 같이 고유명사나 약어를 사용하였고, 한글 단어

〈표 1〉 수집한 자료의 건수

구분	02-04	05-07	08-10	11-13	14-16	17-19	20-22	합계
기술정보	289	278	272	437	438	1,599	51	3,364
사례정보	397	438	2,294	1,940	3,460	3,888	829	13,246
원가절감	2,717	629	407	217	413	1,183	23	5,589
합계	3,403	1,345	2,973	2,594	4,311	6,670	903	22,199

에 비해 출현 빈도가 현저히 낮아 한글 단어를 분석 대상으로 사용하였기 때문에 영어단어를 불용어로 지정하더라도 분석 결과에는 큰 영향을 미치지 않을 것으로 판단하였다. 세 번째로, '일반', '분야', '다양' 등과 같이 단독으로는 의미 전달이 낮은 단어를 불용어로 처리하였다. 네 번째로, '교', '벽', '산' 등 한자리 음절 단어도 단독으로 의미가 거의 없으므로 불용어로 지정하였다. 다섯 번째로, 일반 및 고유명사를 제외한 동사, 형용사, 부사, 대명사 등 품사에 해당하는 단어는 단독으로는 고유한 의미가 없다. 따라서 일반 및 고유명사에 해당하는 단어만을 추출하기 위해 파이썬의 konlpy 패키지 중 'Hannanum'이라는 클래스를 적용하였다. 〈표 2〉는 불용어 제거와 형태소 분석을 통해 추출한 명사인 단어의 개수를 나타낸 것이다. 추출한 명사에 해당하는 단어를 최종 분석 데이터로 사용하였다. 김재준, 정절우 (2012)의 사례와 같이 보통 출현 빈도가 높은 단어는 분석 데이터에서 중요한 역할을 한다고 판단할 수 있다. 본 연구에서 단어의 출현 빈도를 분석한 결과로, '공사' 단어는 기술정보에서 1,402번 출현하였고, 사례정보에서는 15,202번 출현하였다. 원가절감은 1,667번 출현하였다. 이처럼 동일 단어가 평균 17.3번이 중복한 것으로 파악되었다. 이렇게 중복이 많은 것은 제목에서 사용된 단어는 자료 내용을 함축적으로 표현하는 단어의 종류가 한정되어서 나타난 결과라고 판단된다.

〈표 3〉은 기술정보, 사례정보 및 원가절감 별로 출현 빈도가 높은 상위 10개의 단어를 나타낸 것이다. 〈표 3〉에서 보면 기술정보와 사례정보 간, 사례정보와 원가절감 간에 상충하는 단어가 많았다. 이는 기술정보와 사례정보가 건설기술을 대상으로 한 단어를 많이 사용한 결과라고 생각된다. 반면에, 사례정보와 원가절감은 건설공사 사례를 대상으로 하였기 때문에 상충한 단어가 많은 것으로 판단된다. 도로 확장·포장 건설공사의 설계, 시공 및 사업관리, 유지

〈표 2〉 분석용 단어

구분	전체 단어 수	중복제거 단어 수	중복비율(%)
기술정보	20,978	2,207	951
사례정보	110,662	3,419	3,237
원가절감	27,165	3,553	765
합계	158,805	9,179	1,730

〈표 3〉 출현 빈도가 높은 단어 및 비율

구분	상위 10개 단어(비율 %)
기술정보	공사(6.68), 절차(3.76), 지침서(3.23), 유지관리(3.07), 관리(3.03), 시공(2.95), 지침(2.22), 도로(2.16), 전문(1.64), 시험(1.63)
사례정보	공사(13.74), 도로(4.33), 보고서(3.86), 시방서(3.77), 설계(3.69), 확장(2.59), 국도(2.45), 실시(2.33), 시공(2.31), 포장(1.95)
원가절감	개선(6.14), 설치(2.20), 변경(2.16), 시공(1.57), 설계(1.55), 공법(1.44), 검토(1.38), 공사(1.15), 콘크리트(1.15), 터널(1.12)

관리 및 시방・기준 등과 관련한 단어의 출현 빈도가 높은 것으로 파악되었다.

4.3 주제 모형화(topic modeling) 분석

텍스트마이닝에서는 비정형 텍스트에서 사용되는 단어 간의 선행과 후행 관계에 따라 나타나는 일정한 형태를 분석하여 의미가 있는 맥락을 파악한다. 의미가 있는 맥락에 해당하는 단어를 모으면 특정 주제에 대한 모형화를 생성할 수 있다. 주제 모형화는 주제 개수에 따라 주제별로 유사한 의미를 갖는 주제 단어가 달라질 수 있다. 이때 분석가가 직접 주제 개수를 지정해야 한다. 따라서 분석 데이터에 적합한 주제 개수를 지정하지 않으면 자칫 의도하지 않은 결과를 얻을 수 있다. 텍스트마이닝에서는 적정한 주제 개수를 추정하기 위해 통계적 방법을 기반으로 한 혼란도(perplexity)와 주제 일관성(topic coherence)을 많이 사용한다. 혼란도는 주제 모형화가 실제로 관측한 값을 얼마나 잘 예측하는지를 측정하는 데 사용한다. 혼란도의 값이 작을수록 주제 모형화를 잘 예측하였다고 판단한다. 하지만 혼란도의 값이 작다고 반드시 예측 결과가 정확하다고 할 수는 없다. 이러한 혼란도의 한계를 보완하기 위해 주제 일관성을 적용할 수 있다. 주제 일관성은 주제별로 포함한 주제 단어 간의 의미적 유사도를 측정한다. 주제 일관성이 높을수록 해당 주제 모형화는 의미론적으로 일관성이 높다고 판단한다. 본 연구에서는 건설실무정보에 부합하는 주제 개수를 추정하기 위해 파이썬 3.10 분석 도구를 사용하였고, 서대호(2019)에서 제시한 주제 개수 지정 방식을 적용하여 〈표 4〉와 같이 혼란도와 주제 일관성을 측정하였다. 원가절감에 대한 혼란도가 가장 낮았지만, 주제 일관성은 가장 높았다. 따라서 상대적으로 주제 모형화를 잘 예측하였다고 생각할 수 있다. 나머지 기술정보와 사례정보도 대체로 의미론적 맥락을 갖는다고 판단할 수 있다.

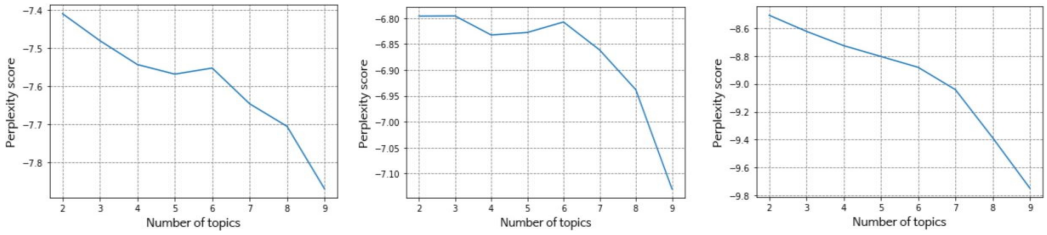
〈표 4〉 혼란도 및 주제 일관성 측정 결과

구분	혼란도	주제 일관성
기술정보	-7.44	0.40
사례정보	-6.77	0.28
원가절감	-8.64	0.44

다음으로, 박준용(2021)의 사례를 준용하여 주체 개수를 2에서 9까지를 지정하였을 때 〈그림 2〉와 〈그림 3〉과 같이 혼란도와 주제 일관성의 변화 정도를 분석하였다. 〈그림 2〉에서 보면 혼란도는 4~6개의 주제 개수에서 혼란도의 값이 가장 컸다가 급격히 감소하는 경향을 보였다. 또한, 〈그림 3〉에서 보면 주제 일관성은 4 또는 5개의 주제 개수에서 주제 일관성의 값이 가장 크다가 다시 급격히 감소하였다. 이러한 변화를 고려하여 토픽 개수를 4개로 지정할 때 건설실무정보에 대한 주제 모형화가 반영되었다고 판단하였다.

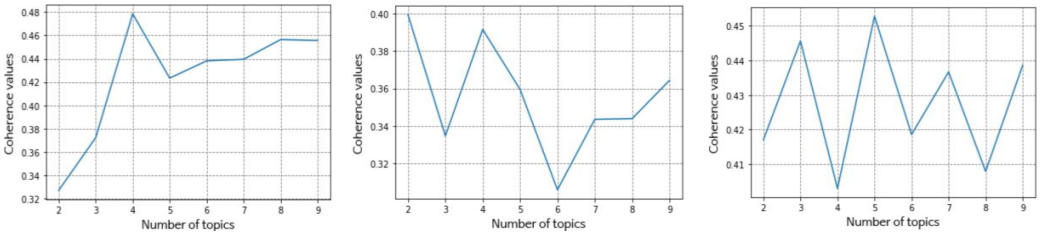
〈표 5〉는 주제 개수를 4로 지정하였을 때 주제별 유사도가 높은 상위 10개의 단어를 나타낸 것이다. 〈그림 4〉는 주제별로 유사도가 높은 10개의 단어 간의 연결 관계를 시각화한 것이다.

기술정보에 대한 주제 모형화에서는 건설공사 시방, 구조물 보수, 국도 유지보수 및 도로 안전 공사 등과 관련된 유사한 의미를 갖는 단어가 의미론적인 맥락을 갖는 것으로 파악되었



(a) 기술정보 (b) 사례정보 (c) 원가절감

〈그림 2〉 혼란도의 변화

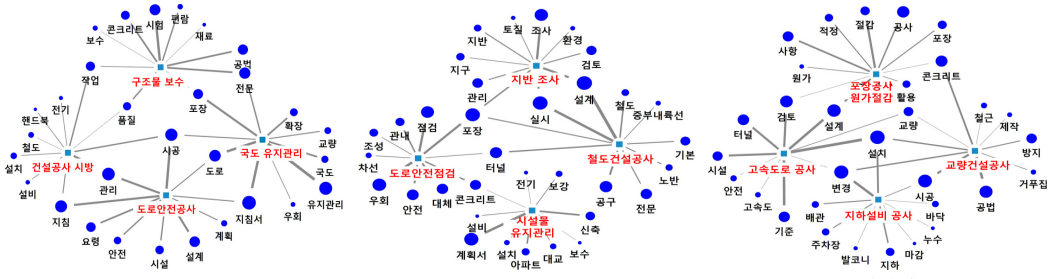


(a) 기술정보 (b) 사례정보 (c) 원가절감

〈그림 3〉 주제 일관성의 변화

〈표 5〉 주제별 상위 10개 단어

구분	기술정보	사례정보	원가절감
주제1	지침서, 유지관리, 포장, 도로, 국도, 확장, 시공, 전문, 우회, 교량	우회, 포장, 점검, 터널, 조성, 차선, 안전, 대체, 관내, 콘크리트	절감, 설계, 사항, 적정, 공사, 원가, 포장, 검토, 활용, 콘크리트
주제2	공법, 시험, 전문, 콘크리트, 편람, 품질, 포장, 작업, 재료, 보수	설계, 포장, 노반, 전문, 기본, 실시, 공구, 터널, 철도, 중부내륙선	설치, 설계, 기준, 검토, 터널, 교량, 시설, 고속도, 안전, 변경
주제3	작업, 관리, 지침, 시공, 설치, 설비, 핸드북, 전기, 철도, 품질	설치, 전기, 콘크리트, 보강, 대교, 설비, 보수, 계획서, 신축, 아파트	거푸집, 방지, 제작, 철근, 교량, 공법, 콘크리트, 시공, 설치, 변경
주제4	계획, 지침서, 지침, 시공, 관리, 설계, 안전, 요령, 도로, 시설	설계, 실시, 포장, 검토, 조사, 토질, 지반, 관리, 환경, 지구	변경, 지하, 설치, 배관, 주차장, 바닥, 시공, 누수, 발코니, 마감



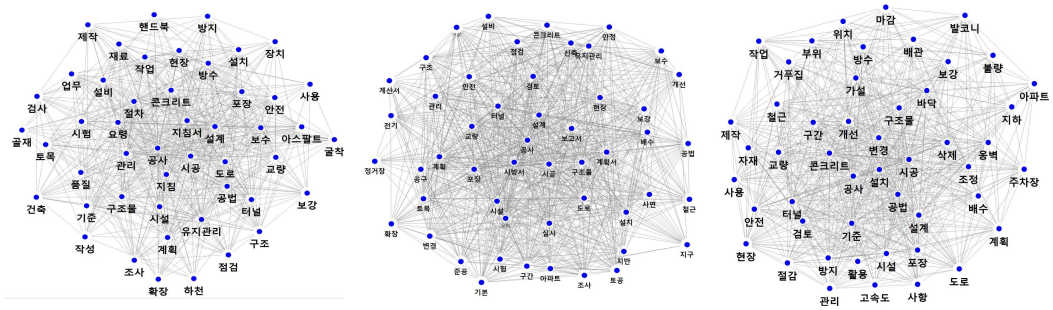
(a) 기술정보 (b) 사례정보 (c) 원가절감

〈그림 4〉 주제 모형화

다. ‘시공’ 단어는 3개의 주제에서 함께 연결 관계를 맺었다. 이외에 ‘작업’, ‘관리’, ‘포장’, ‘전문’, ‘도로’, ‘품질’, ‘지침’ 등 도로공사와 공사 시방에 관련된 단어는 2개의 주제 사이에 연결되었다. 사례정보에 대한 주제 모형화에서는 도로 안전 점검, 지반공사, 철도건설공사, 시설물 유지관리 등 주제로 구분할 수 있다. 3개의 주제에서 ‘포장’ 단어를 공동으로 사용하였다. ‘터널’, ‘설계’, ‘실시’, ‘콘크리트’ 등 도로와 철도 공사의 유지관리에 관련된 단어는 2개의 주제 간에 연결되었다. 끝으로, 원가절감에서는 고속도로, 포장, 교량 및 지하 설비 등 건설공사에 관한 주제로 지정할 수 있다. 3개의 주제에서 ‘설치’, ‘변경’ 단어가 연결 관계를 맺었다. ‘설계’, ‘시공’, ‘교량’, ‘콘크리트’, ‘검토’ 등 도로 및 지하 설비와 원가절감과 관련된 단어는 2개의 주제 사이에 연결되었다.

4.4 네트워크 중심성(network centrality) 분석

주제 모형화를 통해 특정 주제마다 유사한 의미를 갖는 단어 간의 연결 관계를 알 수 있으나, 특정 단어가 말뭉치에서 얼마나 중요한 위치에 있는지는 알 수 없다. 네트워크에서 특정 단어가 차지하는 위치를 정량적 계수로 측정하기 위해 중심성을 사용할 수 있다. 임병학, 전주희(2011) 및 김동희, 이원형, 조성일(2018) 사례와 같이 네트워크에서 단어 간의 직·간접적인 연결 정도, 연결된 단어 간의 최단 거리 또는 최소 연결 순서 등에 따라 연결(degree), 근접(closeness), 매개(between), 고유벡터(eigenvector), 페이지랭크(pagerank) 등 여러 모형의 중심성(centrality)을 적용할 수 있다. 본 연구에서는 중심성 측정에 기본으로 사용하는 연결 중심성과 단어 간의 직접적인 연결뿐만 아니라 간접적으로 연결된 이웃 단어 간의 최단 거리를 연쇄적으로 누적한 계수를 반영하는 고유벡터 중심성을 적용하여 말뭉치에 포함된 단어로 구성한 네트워크에서 차지하는 단어의 위치를 측정하였다. 본 연구는 다음과 같은 절차로 중심성 계수를 측정하였다. 첫 번째로, Netminer 4.1 분석 도구를 사용하여 〈표 2〉에서 제시한 단어 간의 연결 관계를 네트워크로 표출하였다. 표출된 네트워크를 구성한 단어가 많아서 단어 간의 연계 관계가 하나의 점처럼 나타났다. 따라서 〈그림 5〉



(a) 기술정보 (b) 사례정보 (c) 원가절감

〈그림 5〉 건설실무정보 네트워크

와 같이 출현 빈도가 높은 상위 50개의 단어를 추출하여 네트워크를 구성하였다. 이때 단어는 노드 모양으로, 단어 간의 연결은 선으로 표시하였다.

두 번째로, 임병학, 전희주(2011)에서 기술한 연결 및 고유벡터 중심성 계수를 계산하는 식을 적용하여 건설실무정보에 대한 중심성 계수를 측정하였다. 세 번째로, 〈표 6〉과 같이 연결 및 고유벡터 중심성 계수에 대한 통계적 특성을 분석하였다. 사례정보의 연결 및 고유벡터 중심성의 최댓값이 가장 높았다. 다음으로, 기술정보의 연결 중심성의 최댓값이 다음으로 높았으나, 고유벡터 중심성 계수에서는 가장 낮은 것으로 측정되었다. 하지만 중심성 계수의 평균과 표준편차를 보면 모든 건설실무정보는 큰 차이가 없는 것으로 측정되었다.

네 번째로, 〈표 7〉 및 〈표 8〉과 같이 네트워크에서 중심적 위치에서 있는 상위 10개의 단어를 추출하였다. 연결 중심성에서는 '공사' 단어가 가장 중심적 위치에 있고, 이는 출현 빈도와 주제 모형화의 분석 결과와 같았다. '공사', '설계', '시공' 단어는 모든 건설실무정보에서 상위 10위 안에 포함되었다. 고유벡터 중심성에 있어서는 연결 중심성과 유사하게 '공사'와 '시공' 단어가 가장 중심적 위치에 있었다.

한편, 기술정보에 있어서는 연결 및 고유벡터 중심성을 비교하면 '공사', '시공', '관리', '절차', '지침', '작업' 등 단어는 모든 중심성에서 상위 10위 내에 모두 포함되었다. 사례정보에 있어서는 연결 및 고유벡터 중심성을 비교하면

〈표 6〉 연결 및 고유벡터 중심성의 통계적 분포 비교

구분	연결 중심성			고유벡터 중심성		
	최댓값	평균	표준편차	최댓값	평균	표준편차
기술정보	0.149	0.002	0.006	0.544	0.002	0.022
사례정보	0.204	0.002	0.006	0.641	0.001	0.018
원가절감	0.101	0.002	0.005	0.579	0.002	0.017

〈표 7〉 상위 10개 연결 중심성 계수

구분	기술정보		사례정보		원가절감	
	단어	계수	단어	계수	단어	계수
1	공사	0.149	공사	0.204	개선	0.101
2	시공	0.082	보고서	0.117	설치	0.093
3	관리	0.067	시방서	0.116	변경	0.071
4	절차	0.066	시공	0.110	공사	0.067
5	설계	0.061	설계	0.074	설계	0.065
6	콘크리트	0.061	터널	0.051	시공	0.065
7	시험	0.056	시설	0.047	콘크리트	0.052
8	지침	0.055	계획	0.046	공법	0.046
9	공법	0.049	도로	0.045	터널	0.043
10	작업	0.042	검토	0.039	시설	0.040

〈표 8〉 상위 10개 고유벡터 중심성

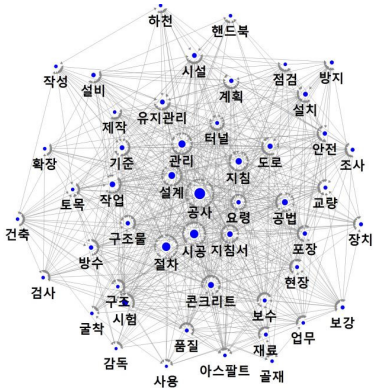
구분	기술정보		사례정보		원가절감	
	단어	계수	단어	계수	단어	계수
1	유지관리	0.544	공사	0.641	개선	0.579
2	지침서	0.459	시방서	0.468	기준	0.418
3	공사	0.449	포장	0.304	공법	0.354
4	시공	0.294	실시	0.277	설치	0.301
5	관리	0.222	확장	0.272	설계	0.249
6	절차	0.156	설계	0.239	시공	0.236
7	포장	0.149	시공	0.140	변경	0.185
8	지침	0.125	보고서	0.092	검토	0.154
9	작업	0.118	도로	0.070	시설	0.136
10	품질	0.116	신축	0.057	공사	0.075

'공사', '설계', '시공', '시방서', '도로' 등 단어는 모든 중심성에서 상위 10위 내에 포함되었다. 원가절감에서는 '콘크리트', '터널', '기준', '검토' 등 단어를 제외한 나머지 단어는 연결과 고유벡터 중심성에서 모두 상위 10위 내에 포함되었다. 기술정보와 사례정보 간에는 같은 단어가 많았지만, 원가절감과는 서로 다른 단어가 많은 것으로 파악되었다. 다섯 번째로, 〈그림 5〉에서 적용한 것과 동일하게 연결 및 고유벡터 중심성에서 중심에 위치하는 상위 50개 단어 간에 연결 관계를 〈그림 6〉과 같이 시각화하였다. 여기서 중심에 가까울수록 노드의 크기를 크게 표시하였다. 예를 들어, 중심성 계수가 가장 높은 단어인 '공사'는 네트워크에서 가장 크게 노드가 표시되었다.

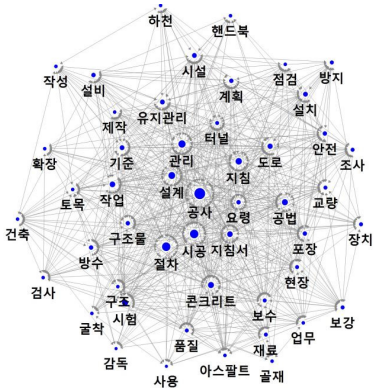
〈그림 6〉에서 보듯이 〈표 7〉과 〈표 8〉에서 기술된 단어가 네트워크에 중심적 위치에 있는 것을 알 수 있다. 이처럼 기술정보와 사례정보는 상호 연계성이 높았고, 원가절감은 다소 독립적인 경향을 보였다. 또한, 도로, 교량, 터널 등 시설물 건설공사의 설계, 시공 및 공법・시방・지침 등과 관련한 단어에 관심이 높은 것으로 유추할 수 있다.

5. 건설실무정보 간의 상관도 분석

본 연구에서는 기술정보, 사례정보 및 원가절감 중 하나를 독립변수로 두고, 다른 하나는 종속변수로 둔 후에 독립변수가 변함에 따라 종속변수가 얼마나 변화하는지를 파악할 수 있는 상관관계를 분석하였다. 여기서 종속변수가 얼마나 변화하는지를 수치로 나타낸 것이 상관계수이다. 본 연구에서는 다음과 같은 절차로 상관계수를 측정하였다. 먼저, 기술정보, 사례정보 및 원가절감에서 모두 존재하는 단어를 추출하였다. 다음으로, 이강원, 이정원(2017), 임병학, 전희주(2011)의 사례에서는 중심성 모형 간에 상관관계를 분석하였으나, 〈표 9〉는 연결 중심성과 고유벡터 중심성에 모두 존재하면서 중심성 계수가 높은 상위 10개의 단어를 비교한 것이다.

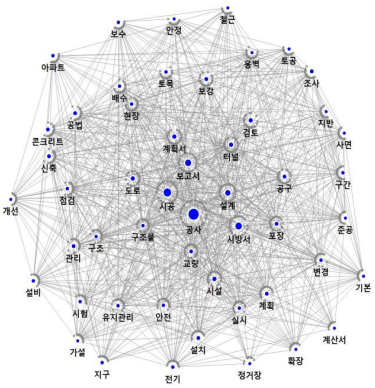


1) 연결 중심성

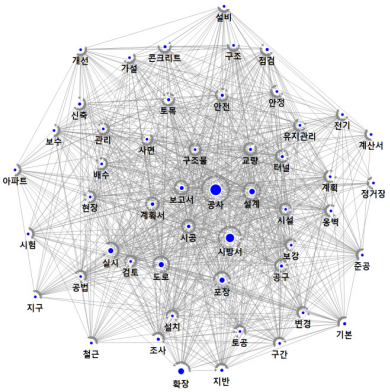


2) 고유벡터 중심성

(a) 기술정보

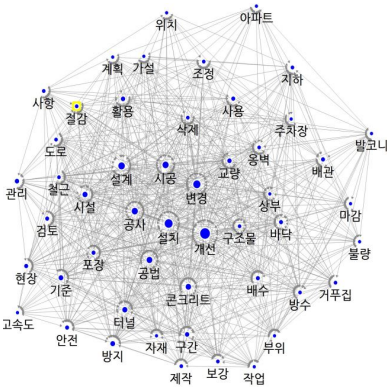


1) 연결 중심성

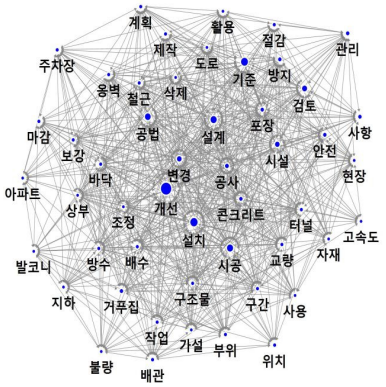


2) 고유벡터 중심성

(b) 사례정보



1) 연결 중심성



2) 고유벡터 중심성

(c) 원가절감

〈그림 6〉 중심성 기반의 네트워크 구성

〈표 9〉 상위 10개의 단어의 중심성 계수 비교

구분	연결 중심성				고유벡터 중심성			
	단어	기술정보	사례정보	원가절감	단어	기술정보	사례정보	원가절감
1	공사	0.149	0.204	0.067	유지관리	0.544	0.015	0.013
2	시공	0.082	0.110	0.065	공사	0.449	0.641	0.075
3	관리	0.067	0.038	0.033	시공	0.294	0.140	0.236
4	절차	0.066	0.002	0.003	관리	0.222	0.011	0.064
5	설계	0.061	0.074	0.065	절차	0.156	0.000	0.008
6	콘크리트	0.061	0.034	0.052	포장	0.149	0.304	0.059
7	시험	0.056	0.025	0.012	지침	0.125	0.000	0.011
8	지침	0.055	0.004	0.006	작업	0.118	0.000	0.040
9	공법	0.049	0.029	0.046	품질	0.116	0.005	0.033
10	작업	0.042	0.007	0.018	확장	0.097	0.272	0.007

두 번째로, 상관도와 회귀모형에 대한 유의성을 측정하기 위해 엑셀의 통계 데이터 분석 기능 중 상관분석과 회귀분석 기능을 사용하였다. 〈표 10〉은 연결 중심성을 기반으로 한 기술정보, 사례정보 및 원가절감 간의 상관계수를 나타낸 것이다.

〈표 10〉에서 보면 기술정보는 사례정보와 0.661이라는 상관계수(R)를 얻었고, 원가절감과는 0.526을 얻었다. 사례정보와 원가절감의 상관계수로는 0.571로 나타났다. 이러한 상관계수는 건설실무정보 간의 기울기가 90도의 대각선보다 조금 직선에 가까운 연관성을 가졌다. 또한, 결정계수(R2)는 독립변수가 종속변수에 미치는 영향에 대해 설명력이 있다고 판단하였다. 결정계수의 값이 1에 가까울수록 해당 회귀모형은 설명력이 좋다고 해석한다. 기술정보와 사례정보 간의 결정계수는 0.437이었고, 사례정보와 원가절감 간에는 0.326을 얻었고, 원가절감과 기술정보 간에는 0.276으로 측정되었다. 따라서 결정계수가 모두 0.5보다 낮은 수치이므로 회귀모형이 다소 낮은 설명력을 갖는다고 판단할 수 있다. 다음으로, 회귀모형의 가설 검정을 위해 유의확률인 P-값을 측정하였다. 유의확률인 P-값은 독립변수와 종속변수 간의

〈표 10〉 상관분석 결과

구분	기술정보 ↔ 사례정보	사례정보 ↔ 원가절감	원가절감 ↔ 기술정보
상관계수	0.661	0.571	0.526
결정계수	0.437	0.326	0.276
조정된 결정계수	0.436	0.325	0.275
표준오차	0.009	0.007	0.008
유의한 F	0.000	0.000	0.000
P-값	0.000	0.000	0.000
관측수	876	876	876

유의적인 상관관계를 나타낸다. 이때 보통 유의성 기준으로 0.05를 사용한다. 만약 유의확률이 0.05 미만일 경우에는 평균 간에 차이가 없다는 귀무가설을 기각하고 유의한 차이가 있다고 해석한다. 또한, 독립변수의 종속변수에 대한 영향력의 유의성을 판단하는데 F-통계량을 계산하였다. F-통계량이 0.05보다 작으면 회귀모형은 종속변수에 대한 유의성을 갖는다고 해석한다. 회귀분석을 통해 〈표 10〉과 같은 결과를 얻었다. 〈표 10〉에서 보면 유의확률인 P-값과 F-통계량이 거의 0에 가까움으로 독립변수의 상관관계는 통계적으로 유의미하다고 할 수 있다. 따라서 연구가설에 사용된 중심성 계수가 높은 단어 간의 상관관계는 요인분석에 적절하다고 해석할 수 있다. 〈표 10〉에서 보면 기술정보와 사례정보 간의 상관관계가 높았지만, 원가절감과는 상대적으로 상관관계가 낮은 것으로 파악되었다.

6. 결 론

한국건설기술연구원은 국내 건설 관련 기업의 기술경쟁력을 강화하기 위해 각종 건설보고서, 건설공사 실무자료, 건설공사 기준, 품셈 등 각종 건설기술 정보를 DB로 구축하여 CODIL을 통해 서비스하고 있다. 최근 들어 발주자의 요구가 다양화, 전문화, 고품질화 되면서 건설현장에서는 실무에서 필요로 하는 자료를 요구하고 있다. 하지만 한정된 예산과 자료 공개를 꺼리는 경향이 커지고 있어 현재의 서비스 방식으로는 건설 현장의 요구를 충족시키는 데 한계가 있다. 따라서 건설 현장의 요구를 어느 정도 충족시키면서 전문지식을 갖지 않은 건설기술자와 건설사업 참여자에게 건설 실무에서 중요도가 높은 단어에 대한 정보를 제공하는 서비스 개선책을 제시하고자 하였다. 이를 위해 CODIL에서 가장 많이 조회되고 있는 기술정보, 사례정보 및 원가절감 등 건설실무정보를 대상으로 하여 텍스트마이닝과 네트워크 중심성을 이용하여 중요도 높은 단어를 추출하였고, 이를 네트워크로 시각화하였다. 또한, 기술정보, 사례정보 및 원가절감 간의 상관도를 분석하여 이들 정보 중 허브 역할을 하는 정보가 무엇인지를 분석하였다. 분석 결과, 기술정보가 사례정보와 원가절감보다 중요한 것으로 파악하였다. 이러한 일련의 과정은 새로운 이론을 개발하거나 실증 분석기법을 고안하기보다는 현재 선행연구에서 사용한 텍스트마이닝의 적용 사례를 준용하여 건설실무정보의 암묵지적인 특성과 상관관계를 파악하였는데 의의를 둘 수 있다. 특히, 더블린코어 메타데이터 기반의 서지정보 중 자료의 제목을 하나로 취합한 말뭉치에서 중요도가 높은 암묵지적인 정보를 제공할 수 있다는 점이 기존 CODIL의 정보서비스 방식과 차이가 있다. 또한, 주제의 개수에 따라 주제별로 유사한 의미를 갖는 주제 단어가 달라진다. 일부 선행연구에서는 주제의 개수에 대한 판단기준을 명확히 제시하지 않은 경우가 있었으나 본 연구에서는 통계를 기반으로 한 혼란도와 주제 일관성을 계산하여 건설실무정보에 적정한 주제의 개수를 추정하였다. 게다가, 본 연구에서는 중심성 모형 간에 상관도를 분석하기보다는 연결과 고유벡터 중심성 모형에 모두 존재하는 단어 중 중요도가 높은 단어를 대상으로 상관도를 분석하였다는 점이 기존

선행연구와의 차별이라 할 수 있다. 끝으로, 본 연구는 다음과 같은 제약과 이에 따른 추가적인 연구가 필요할 것으로 사료된다. 첫 번째로, 연구 결과의 신뢰성을 확보하기 위해서는 〈표 7〉과 〈표 8〉에서 보듯이 '공사'와 '시방', '시방'과 '시방서', '지침'과 '지침서'와 같이 유사한 의미를 갖는 단어를 하나의 동의어(thesaurus) 사전으로 구축할 필요가 있다. 특히, 연결된 개방형 데이터(Linked Open Data)(한국정보화진흥원, 2015)와 온톨로지(ontology)를 응용하여 단어가 갖는 의미론적 연관도와 연관도에 따른 유사 단어 간의 가중치를 구축한다면 좀 더 정확한 의미론적 분석이 가능할 것으로 생각된다. 둘째로, 본 연구는 한국어 말뭉치에 포함된 단어의 형태소 처리를 위해 설계된 konlpy 패키지를 사용하다보니 영어단어를 불용어로 지정하였다. 따라서 영어로 표기되는 성향이 강한 주제개념이 빠지게 되었다. 이를 보완하기 위해서는 한국어와 영어를 함께 분석할 수 있는 패키지 개발에 관한 연구가 추가로 필요하다. 셋째로, 본 연구는 건설실무정보의 제목을 가지고서 분석을 위한 말뭉치로 사용하였다. 하지만 내실 있는 연구 결과를 얻기 위해서는 텍스트마이닝의 특성을 고려할 때 단문 형태의 제목보다는 요약문이나 결론과 같이 장문의 비정형 텍스트를 분석용 말뭉치로 사용할 필요가 있다. 이를 위해서는 요약문이나 결론에 대한 원문서비스가 선행되어야 한다.

참 고 문 헌

건설기술정보시스템 (2022. 09. 20). 출처: http://www.codil.or.kr

길호현 (2019). 텍스트마이닝을 위한 한국어 불용어 목록 연구. 우리말글, 78, 1-25.

김동회, 이원형, 조성일 (2018). 트위터에서 형태소 분석과 PageRank 기반 화제단어 추출 방법 제안. 디지털콘텐츠학회 논문지, 19(1), 157-163. https://doi.org/10.9728/dcs.2018.19.1.157

김재준, 정절우 (2012). 텍스트마이닝을 활용한 건설분야 트랜드 분석. 한국 디지털 건축・인테리어학회 논문지, 12(2), 53-60.

박준용 (2021). 텍스트 마이닝을 활용한 토목분야 연구토픽 분석. 서울기술연구, 11, 50-52.

서대호 (2019). 잡아라! 텍스트마이닝 with 파이썬. 서울: 비제이퍼블릭.

성현곤, 박인권, 지규현, 김진유, 장경석, 김준형, 고진수, 김승화 (2019). 텍스트마이닝을 활용한 최막중의 국토 및 도시계획 연구경향 분석: 학술지 게재논문을 중심으로, 학술지 게재논문을 중심으로. 국토연구, 103, 3-26. http://dx.doi.org/10.15793/kspr.2019.103..01

오준석 (2015). 텍스트마이닝 방법을 통한 국내 교통・ICT 융합 분야 연구기회 발견. 교통연구, 22(4), 93-110. http://dx.doi.org/10.34143/jtr.2015.22.4.93

이강원, 이정원 (2017). 네트워크 중심성 지표를 이용한 서울 수도권 지하철망 특성 분석. 한국철도학회

논문집, 20(3), 413-422. https://doi.org/10.7782/JKSR.2017.20.3.413
임병학, 전희주 (2011). 항만의 사회 네트워크가 물동량에 미치는 영향에 대한 연구: 항만간 협력지수를 중심으로. POSRI경영경제연구, 11(3), 289-305.
정근하 (2010). 텍스트마이닝과 네트워크 분석을 활용한 미래예측 방법 연구. 한국과학기술기획평가원.
최정묵 (2016). 중심성지수를 이용한 행정학・정책학 관련 학술지의 상호인용 네트워크 분석. 한국디지털정책학회・디지털복합연구, 14(9), 301-308. http://dx.doi.org/10.14400/JDC.2016.14.9.301
한국정보화진흥원 (2015). 알기 쉬운 Linked Open Data.
KDI 공공투자관리센터 (2021). 2020년도 KDI 공공투자관리센터 연차보고서. 한국개발연구원, 31-33.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Choe, Jong-Mook (2016). Investigating journal citation network with centrality measures in the public administration and policy field. Journal of Digital Convergence, 14(9), 301-308. http://dx.doi.org/10.14400/JDC.2016.14.9.301
Construction Technology Digital Library (2022. 09. 20). Available: http://www.codil.or.kr
Jeong, Geun-Ha (2010). A Study of Foresight Method Based on Textmining and Complexity Network Analysis. Korea Institute of Science and Technology Evaluation and Planning.
KDI Public Investment Management Center (2021). 2020 KDI Public Investment Management Center Annual Report. Korea Development Institute.
Kil, Ho-Hyun (2019). The study of Korean stopwords list for text mining. Urimalgeul The Korean Language and Literature, 78, 1-25.
Kim, Dong-Hoi, Lee, Won-Hyung, & Cho, Sung-Il (2018). Proposal of keyword extraction method based on morphological analysis and PageRank in Tweeter. Journal of Digital Contents Society, 19(1), 157-163. https://doi.org/10.9728/dcs.2018.19.1.157
Kim, Jae-Jun & Jeong, Cheol-Woo (2012). Analysis of trend in construction using textmining method. Journal of The Korean Digital Architecture・Interior Association, 12(2), 53-60.
Lee, Kang-Won & Lee, Jeong-Won (2017). Analysis of Seoul metropolitan subway network characteristics using network centrality measures. Journal of The Korean Society for Railway, 20(3), 413-422. https://doi.org/10.7782/JKSR.2017.20.3.413
Leem, Byung-Hak & Chun, Heui-Ju (2011). A study on the impact of social network at ports on throughput with a focus on port cooperation index. POSRI Business Economics Research, 11(3), 289-305.

National Information Society Agency (2015). Easy to know Linked Open Data.

Oh, Jun-Seok (2015). Identifying research opportunities in the convergence of transportation and ICT using text mining techniques. Transportation Research, 22(4), 93-110. http://dx.doi.org/10.34143/jtr.2015.22.4.93

Park, Jun-Yong (2021). Research topic analysis in civil engineering field using text mining. Seoul Technology Research, 11, 50-52.

Seo, Dae-Ho (2019). Catch it! Text Mining with Python. Seoul: Bjpublic.

Sung, Hyun-Gun, Park, In-Kwon, Ji, Kyu-Hyun, Kim, Jin-Yoo, Jang, Kyoung-Seok, Kim, Jun-Hyung, Ko, Jinsoo, & Jin, Chenghua (2019). The analysis on research trend of territorial and urban planning for Mack Joong Choi through text mining: focused on his academic papers. The Korea Spatial Planning Review, 103, 3-26. http://dx.doi.org/10.15793/kspr.2019.103..01