17CTAP-C114956-02

건설 프로젝트 전주기 위험요인 관리 및 대응을 위한 딥러닝 기반 맞춤형 지식정보 시스템 최종보고서

Deep-Learning-Based Information Service System to Support Life-cycle Risk Management of Construction Projects

R&D Report

2018

국토교통과학기술진흥원

국토교통부

보안 과제( ), 일반 과제(○) / 공개(○), 비공개( )

국토교통기술촉진연구사업 제2차 연도 최종 보고서

R&D /

건설 프로젝트 전주기 위험요인 관리 및 대응을 위한 딥러닝 기반 맞춤형 지식정보 시스템 최종보고서

2018 . 07 . 18 .

주관연구기관 / 서울대학교

국 토 교 통 부

(전문기관) 국토교통과학기술진흥원

2. 제출문

제 출 문

국토교통과학기술진흥원장 귀하

'건설 프로젝트 전주기 위험요인 관리 및 대응을 위한 딥러닝 기반 맞춤형 지식정보 서비스 시스템'(연구개발 기간 : 2016. 06. 24 ~ 2018. 06. 23) 과제의 최종보고서 2부를 제출합니다.

2018. 8. 28.

주관연구기관명 : 서울대학교 산학협력단 (대표자) 김 성 철 (인)

주관연구기관책임자: 지 석 호

국토교통부소관 연구개발사업 운영규정 제38조에 따라 최종보고서 열람에 동의합니다.

3. 보고서 요약서

보고서 요약서

<table>
<tr><td>과제 고유 번호</td><td>17CTAP-
C114956-02</td><td>해당 단계
연구 기간</td><td>2016.06.24.
~2018.06.23</td><td>단계구분</td><td>1단계/
1단계</td></tr>
<tr><td rowspan="2">연구사업명</td><td>중사업명</td><td colspan="4"></td></tr>
<tr><td>세부사업명</td><td colspan="4">국토교통기술촉진연구사업</td></tr>
<tr><td rowspan="2">연구과제명</td><td>대과제명</td><td colspan="4"></td></tr>
<tr><td>세부과제명</td><td colspan="4">건설 프로젝트 전주기 위험요인 관리 및 대응을 위한 딥러닝 기반 맞춤형 지식정보 서비스 시스템</td></tr>
<tr><td rowspan="2">연구책임자</td><td rowspan="2">지 석 호</td><td>해당단계
참여
연구원 수</td><td>총: 3 명
내부: 3 명
외부: 0 명</td><td>해당단계
연구개발비</td><td>정부: 50,000천원
기업: 천원
정부 외: 천원
계: 50,000천원</td></tr>
<tr><td>총 연구기간
참여
연구원 수</td><td>총: 3 명
내부: 3 명
외부: 0 명</td><td>총 연구개발비</td><td>정부:100,000천원
기업: 천원
정부 외: 천원
계:100,000천원</td></tr>
<tr><td>연구기관명 및
소속 부서명</td><td colspan="3">서울대학교 산학협력단</td><td colspan="2">참여기업명</td></tr>
<tr><td>국제공동연구</td><td colspan="3">상대국명:</td><td colspan="2">상대국 연구기관명:</td></tr>
<tr><td>위탁연구</td><td colspan="3">연구기관명:</td><td colspan="2">연구책임자:</td></tr>
<tr><td colspan="4">요약</td><td colspan="2">보고서 면수
47</td></tr>
</table>

4. 국문 요약문

연구의 목적 및 내용	○ **연구 목적** ▪ 건설 프로젝트의 전주기 위험요인 관리 및 대응을 위한 맞춤형 지식정보 서비스 시스템 개발 ○ **연구 내용** ▪ **딥러닝 기법을 활용한 의미기반 검색 알고리즘 개발** - 위험요인 관련 비정형 텍스트 데이터베이스(DB) 구축 및 자동 확장 - 사용자 검색어 의미 파악을 위한 코퍼스 구축 - 의미기반 검색 알고리즘 개발, 검증 및 평가 ▪ **맞춤형 검색 결과 요약 및 지식정보 생산 시스템 개발** - 건설 산업 특화 자연어 분석 및 생성 모델 개발 - 검색 결과 요약 기술 개발 - 지식정보 축적 및 제공을 위한 사용자 인터페이스 개발 - 시스템 검증 및 평가				
연구개발성과	○ **연구개발성과 (1) 딥러닝 기반 의미기반 검색 알고리즘** ▪ 건설 프로젝트 전주기에 걸친 지식정보의 재활용을 지원하는, 건설 산업의 언어적 특징을 반영한 검색 알고리즘. ○ **연구개발성과 (2) 맞춤형 검색 결과 요약 및 지식정보 생산 시스템** ▪ 지식정보에 대한 사용자 접근성을 높이기 위해 검색 결과를 자동으로 요약하며 검색된 문서의 내용을 자동으로 비교·분석하는 맞춤형 지식정보 생산 시스템.				
연구개발성과의 활용계획 (기대효과)	○ **활용방안** ▪ 건설 문서 검색, 비교분석, 요약, 시각화. ▪ 건설 산업 지식정보의 재활용 및 지식의 선순환 구조 확립. ○ **기대효과** ▪ 효과적인 위험요인 관리를 통한 기업의 전략적 계획 수립 및 수익성 확보				
핵심어 (5개 이내)	위험요인	딥러닝	의미기반	로봇저널리즘	자연어 처리

5. 영문 요약문

〈 SUMMARY 〉

Purpose & Contents	○ Purpose ▪ Development of Deep-Learning-Based Information Service System to Support Life-cycle Risk Management of Construction Projects ○ Contents ▪ Development of Deep-Learning-Based Semantic Text Searching Algorithms - Construction and automatic extension of project risk related text database (DB) - Corpus for semantic knowledge search - Creation of construction dictionary for semantic meaning discovery ▪ Development of User Interface to Supply Search and Analysis Results - Development of construction specific natural language processing techniques - Robot journalism for document comparison and summarization - User interface for data collection, analysis and supply - System performance analysis and evaluation
Results	○ Outcome (1) Deep-learning-based semantic searching algorithm ▪ An intelligent text mining techniques that consider construction specific natural language and its semantic meaning to support knowledge discovery and reuse for life-cycle risk management during construction projects. ○ Outcome (2) Customized result summarization and knowledge supply system ▪ An application of robot journalism for document analysis, comparison, and summarization that can support decision making for a range of construction management practices.
Expected Contribution	▪ This research proposes a new approach that can automatically discover management knowledge from a number of hidden, irregular, and scattered construction documents. It will emphasize application feasibility of advanced machine learning and Big Data techniques to the construction industry and support strategic planning and management of construction projects.

Keywords	Risk	Deep Learning	Semantic	Robot Journalism	Natural Language Processing

6. 영문 목차

〈 목 차 〉

1. Overview of R&D ·····
2. Technology Development of domestic and overseas ·····
3. Content and Result of R&D ·····
4. Achievement of Research Objective and Contribution ·····
5. Final Research Achievement and Application Plan ·····
6. Related Scientific and Technologic Information ·····
7. Security Rating of R&D ·····
8. Research Facility and Equipment ·····
9. Implementation of Safety Measurement in Laboratory ·····
10. Representative Achievements of R&D ·····
11. Additional Content ·····
12. Reference ·····

7. 본문 목차

〈 목 차 〉

1. 연구개발과제의 개요 ·····
2. 국내외 기술 개발 현황 ·····
3. 연구 수행 내용 및 성과 ·····
4. 목표 달성도 및 관련 분야 기여도 ·····
5. 연구개발성과의 활용 계획 ·····
6. 연구 과정에서 수집한 해외 과학기술 정보 ·····
7. 연구개발성과의 보안등급 ·····
8. 국가과학기술종합정보시스템에 등록한 연구시설 · 장비 현황 ·····
9. 연구개발과제 수행에 따른 연구실 등의 안전 조치 이행 실적 ·····
10. 연구개발과제의 대표적 연구 실적 ·····
11. 기타 사항 ·····
12. 참고 문헌 ·····

8. 본문

제1장. 연구개발과제의 개요

1. 연구개발 목적

○ 신뢰성 있는 과학적 접근법을 통하여 건설 프로젝트의 전주기 위험요인 관리 및 대응을 위한 맞춤형 지식정보 서비스 시스템 개발

- **딥러닝 기법을 활용한 의미기반 검색 알고리즘 개발**
  - 지식정보 서비스 예상 고객 사전 모니터링
  - 위험요인 관련 비정형 텍스트 데이터베이스(DB) 구축
  - 웹크롤링을 통한 데이터베이스(DB) 자동 확장
  - 딥러닝 기법을 활용하여 사용자 검색어 의미 파악을 위한 코퍼스 구축
  - 의미기반 검색 알고리즘 개발
  - 검색 알고리즘 검증 및 평가
- **맞춤형 검색 결과 요약 및 지식정보 생산 시스템 개발**
  - 머신러닝 기법을 활용한 건설산업에 특화된 자연어 분석 및 생성 모델 개발
  - 로봇저널리즘 기법을 활용한 검색 결과 자동 분석 및 요약문 생성 기술 개발
  - 지식정보 축적 및 제공을 위한 사용자 인터페이스 개발
  - 시스템 검증 및 평가

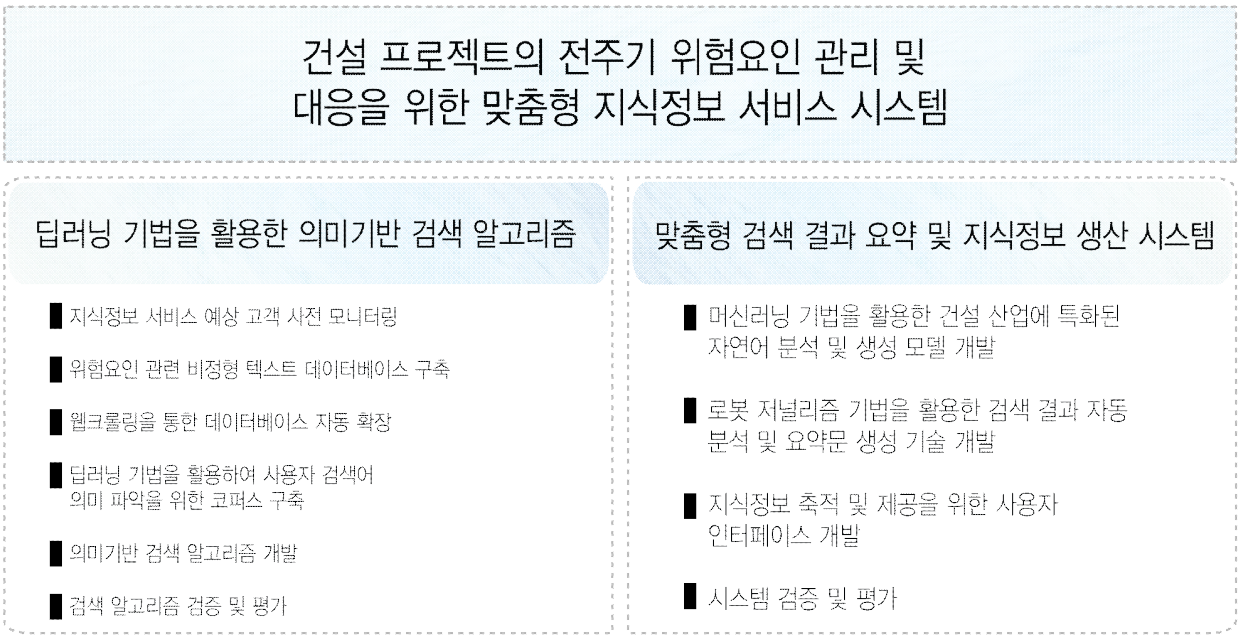


그림 1. 연구개발 목적

2. 연구개발의 필요성

○ 건설 프로젝트 전주기 위험요인(Risk) 관리 및 대응의 필요성

- 위험요인(Risk)이란 프로젝트의 성공적인 수행에 영향을 미치는 불확실성을 내포한 모든 요인을 통칭하며, 건설 산업에서의 위험요인은 결과에 부정적인 영향을 초래할 수 있는 발생 가능한 사건을 의미함.

- 따라서 위험요인은 건설 프로젝트의 전주기(기획, 설계, 발주, 시공, 유지관리)에 걸쳐 존재하며, 대표적인 단계별 위험요인으로는 기획 단계의 해외건설 위험요인, 시공 단계의 건설현장 안전 위험요인, 유지관리 단계의 시설물 유지관리 위험요인 등이 있음.

  - 최근 5년간 국내 건설기업의 높은 시공 기술력을 바탕으로 연간 평균 수주 규모는 과거에 비해 크게 성장했지만, 프로젝트의 관리 역량 면에서는 부족한 실정임 (한국건설산업연구원, 2013). 이에 따라 국내 건설 산업에서는 해외건설사업 리스크 지수(IPRI)와 해외건설공사 통합리스크관리시스템(FIRMS) 등을 개발하여 해외건설 손실에 대한 선제적 모니터링과 전망이 가능하도록 하였음 (유위성, 2015). 하지만 여전히 구체적인 위험요인에 대한 파악이 힘들거나 시공사 위주의 시스템이라는 한계를 지니고 있어 전주기적 해외건설 위험요인을 파악할 수 있는 의사결정 지원 시스템이 필요함.
  - 또한, 국토교통부는 2017년까지 건설현장의 안전사고를 감소시키기 위해 '건설현장 안전 관리체계 개선방안'을 마련했음. 그러나 이렇게 거듭된 건설 안전대책에도 불구하고 높은 재해사고율을 보이는 건설 안전사고를 대폭 감소시키기 위해서는 새로운 건설현장 위험요인 관리 시스템을 마련하는 것이 시급함.
  - 우리나라 건설 산업은 기간시설 확충과 신규 건설 사업에만 주력하여 재해 및 재난의 예방과 준공 후 안전관리를 소홀히 해왔음. 이에 국가차원에서는 시설물안전관리에 관한 특별법이 마련되어 66,000개의 시설물이 관리 대상으로 한국시설안전공단의 총괄하에 관리되고 있으나 모든 시설물을 관리하고 있지는 못하다는 점, 정기점검의 경우 객관성과 안전성 면에서 신뢰도가 떨어진다는 점 등에서 이를 보완해줄 수 있는 객관적이고 체계적인 시스템의 도입이 시급함 (이종세, 2007).

- 이렇듯 건설 프로젝트는 전주기에 걸쳐 다양한 위험요인에 노출되어 있어 위험요인 관리를 위한 시스템에 대한 요구가 꾸준히 제기되었으며, 최근 국내 건설 산업에서도 위험요인 평가 모델에 대한 개발과 활용이 증가하고 있음.

- 현재까지 개발된 위험요인 평가 및 관리 모델은 주로 프로젝트 초기의 기획(수주 및 입찰) 단계에 활용되고 있어, 전주기에 걸친 위험요인 관리 체계로의 개선과 함께 체계적인 위험요인 관리 및 대응이 필요함.

- 건설산업 실무자들을 대상으로 진행한 설문조사에 따르면, 기존의 위험요인 관리는 주로 실무자들의 개인 역량과 경험지식을 바탕으로 수행되었음. 그렇기 때문에 개인·기업이 보유하고 있는 문헌자료, 네트워크, 지식과 인터넷 상에 공개되어 있는 자료가 위험요인 관리에 사용되는 데이터의 대부분을 차지함.

- 기획 단계에서는 뉴스 기사, 시장 보고서, 기업경제 보고서 등을 통해 발주자, 프로젝트, 시장에 관한 정보를 수집하고 발생 가능한 위험요소를 도출하지만, 필요한 정보를 얻기 위해 파악해야 하는 데이터가 매우 방대해 시간·비용이 많이 소요되고 있음.
- 시공 단계에서의 위험요인은 설계 도면과 공사 진행 상황을 토대로 안전사고가 발생할 수 있는 위험 지역을 도출하거나, 안전 관리자가 직접 현장을 점검하며 위험요소를 발견해내는 방식으로 관리되고는 있지만, 주로 관리자의 경험지식에 근거해 정성적인 의사결정이 이루어지고 있고, 과거사례로부터의 교훈을 충분히 재활용하지 못한다는 한계가 있음.
- 유지관리 단계에서는 전문기관이 정기적으로 점검을 수행함으로써 손상 발생을 감지하고 있으나, 예산의 한계로 인해 점검에 투입할 수 있는 전문 인력의 수가 매우 부족하다는 문제점이 있음.

▪ 빅데이터를 활용해 지식정보를 관리·공유한다면 앞서 설명한 위험요인 관리 기법들의 효율성을 더욱 높일 수 있다는 것이 다양한 연구를 통해 검증되었고, 이에 따라 정보공유 시스템의 필요성이 강조되고 있음(Javernick-Will and Scott, 2010; Taroun, 2014). 본 연구의 필요성 검증을 위해 실시한 인터뷰에 따르면 건설산업의 경험지식은 대부분 텍스트데이터의 형태로 존재하기 때문에, 데이터베이스를 잘 구축하고 문서 안에 담겨 있는 경험지식을 효율적으로 재활용하는 것이 매우 중요함.

▪ 다양한 건설 프로젝트 유형 및 프로젝트 전주기에 걸친 종합적인 위험요인 관리 및 대응에 대한 요구가 높아짐에 따라 빅데이터 기술을 활용하여 새로운 문제를 해결하기 위한 효율적 지식정보 관리 및 정보공유 시스템 구축이 필요함.

○ 건설 산업에서의 위험요인 관련 지식정보 재활용 미비

▪ 최근 건설 산업에서는 위험요인 관리 및 대응을 위한 노력의 일환으로 해외건설공사 통합리스크관리시스템(FIRMS), 건설안전정보시스템(COSMIS) 등을 구축하여 각종 위험요인 관련 데이터 및 정보를 제공하고 있음.

▪ 그러나 건설 산업의 프로젝트는 일시적이고 반복적이지 않은 특징으로 인해 지식 공유와 재사용에 한계가 존재하여 '데이터 축적→지식정보 생산→지식 활용→데이터 재축적'의 선순환 구조 실현에 어려움이 존재함.

▪ 각 기업에서 자체적으로 활용중인 PMIS(Project Management Information Systems)의 경우 입력되는 데이터가 행정 위주의 업무 처리로 현실을 제대로 반영하지 못하는 데이터가 수집되는 경우가 대부분이며, 정형화된 데이터보다는 방대한 양의 문서 위주의 비정형 데이터만 축적되고 있어 시스템으로부터 지식정보 생산을 통한 재활용 사례는 극히 제한적임.

▪ 건설 산업에서 위험요인 관련 완결성 있는 데이터 수집과 지식정보 재활용을 유도하여 건설 프로젝트 수행의 효율성을 높이기 위해서는 건설 프로젝트의 각 특성을 반영할 수 있는 맞춤형 지식정보 서비스 체계가 필요함.

○ 첨단 기법을 활용한 위험요인 지식정보 관리 및 생산 인프라 시급

- 건설 프로젝트 수행 중 발생하는 위험요인 관련 정보들은 다양한 종류의 문서에 산재되어 관리되고 있음 (유위성, 2014).
- 또한, 빅데이터로부터 가치 있는 지식정보를 창출할 수 있는 전문 인력의 부족과 건설 프로젝트의 긴박성으로 인해 프로젝트 수행과정에서 생산되는 다양한 정보를 프로젝트 종료 후 정리·분석할 여력이 없어 대부분의 고급 정보가 데이터베이스 내 사장되어 있음.
- 건설 산업에서도 빅데이터 개념 도입 및 맞춤화를 시도하고 있으나 건설 프로젝트 전주기에 걸쳐 발생되는 비정형 텍스트 데이터를 처리하고 관리할 수 있는 시스템의 부재로 실무 활용에 있어 많은 어려움이 있음.
- 신속한 위험요인 식별 및 보고, 체계적인 통제·모니터링과 대응방안 수립이 가능하도록 방대한 양의 문서정보로부터 건설현장 안전 관리, 시설물 유지관리, 해외건설 위험요인 분석 등에 필요한 지식정보를 자동으로 생산하는 선제 대응형 시스템이 필요함.

○ 전주기 위험요인 관리 및 대응을 위한 시스템이 지녀야 하는 특징

- 기존 프로젝트 사례에 기반을 둔 전주기형 지식정보 서비스
- 위험요인 자동 분석 및 정보제공이 가능한 빅데이터 기반 지식정보 서비스
- 비정형 데이터를 활용하는 신뢰성 있는 과학적 지식정보 서비스
- 최적의 위험요인 관리 및 대응을 위한 사용자 맞춤형 지식정보 서비스
- 기업 경쟁력 강화를 위한 개방형/범용적 지식정보 서비스

3. 연구개발 범위

- 건설 프로젝트에서 위험요인은 전주기(기획, 설계, 발주, 시공, 유지관리)에 걸쳐 존재하며, 현재 존재하는 위험요인 관련 비정형 텍스트 데이터 또한 매우 다양하고 복잡한 형태일 것으로 예측됨. 건설산업에 존재하는 비정형 텍스트데이터 현황은 본 보고서의 제3장(연구 수행 내용 및 성과)에 구체적으로 기술함(보고서 15-16pp).
- 따라서 연구 기간(2년) 내에 모든 위험요인에 관련한 데이터를 수집 및 분석하는 과정에만 상당 시간이 소요될 것으로 판단되어 대표적 단계별 위험요인인 기획 단계의 해외건설 위험요인, 시공 단계의 건설현장 안전 위험요인, 유지관리 단계의 시설물 유지관리 위험요인의 데이터를 중점적으로 활용하여 맞춤형 지식정보 서비스 시스템을 개발함.

- 특히 계약서, 시방서, 설계 문서, 프로젝트 결과 보고서 등 실제 건설 프로젝트의 정보를 담고 있는 데이터는 공개되어 있지 않은 경우가 많기 때문에, 기업의 협조가 없이는 데이터를 수집하고 분석하는 데 현실적인 어려움이 있음. 본 연구개발을 위해 수집된 데이터 목록은 본 보고서의 제3장(연구 수행 내용 및 성과)에 구체적으로 설명함(보고서 15-16pp).

- 본 연구가 완료된 이후에는 국내외 기업 및 연구기관과의 협업을 통해 계약서류, 시방조건, 설계변경 등 공개적으로 수집하기 어려운 데이터를 확보하고, 이를 바탕으로 건설 프로젝트를 수행하는 과정에서 발생할 수 있는 문제 상황을 분석하고 위험요인에 사전적으로 대응하도록 하는 연구를 수행할 계획임.

제2장. 국내외 기술 개발 현황

- 건설산업의 위험요인을 관리하기 위한 지식정보 서비스 시스템에 관련된 국내·외 기술 개발 현황은 다음과 같음.

과거	최근 (3년 이내)
▪ 국가적 차원에서 건설산업 정보통합관리체계를 마련하기 위해, 국토교통부에서는 건설산업기본법 제24조(건설산업정보의 종합관리)에 의거하여 1999년부터 현재까지 건설산업 데이터베이스 구축사업을 진행하고 있음. 하지만 구축된 데이터베이스는 대부분 단순히 데이터를 저장하기 위한 역할만을 수행하기 때문에 지식정보를 관리·공유하기 위한 기반으로서는 적합하지 않음(국토교통부, 2014).	▪ 과거 데이터가 충분히 존재하는 경우 시뮬레이션 등의 기법을 활용해 미래의 상황을 예측하는 연구도 진행되고 있지만(Hyari et al. 2016), 정형화된 데이터만을 사용했다는 점에서 온전한 지식정보를 제공하고 있지 않음.
▪ 국내 건설기업들의 안정적이고 전략적인 해외시장 진출을 지원하기 위해 해외건설협회에서는 해외건설공사 통합리스크관리시스템을 운영 중이지만, 정보의 수요자가 시공사에 한정되어 있기 때문에 활용 범위가 한정적임.	▪ 챗봇(Chat-bot), 로봇 기사(Robot Article) 등 비정형 데이터인 텍스트데이터를 분석·가공하여 지식정보를 제공하는 시스템이 운영되고 있으며, 건설산업에서도 다수의 텍스트데이터가 존재하기 때문에 해당 방법론을 활용할 수 있을 것으로 기대됨.
▪ 건설기업에서 주로 사용해왔던 지식관리 도구들인 지식맵(Knowledge Map), 지식기반 전문가시스템	▪ 새로운 건설 프로젝트를 기획하는 과정에서 텍스트마이닝 방법론을 활용하여 표준 시방서와 공사 시방서의 내용을 비교분석함으로써 시공·운영 중 발생 가능한 위험요인을 사전에 도출하려는 연구가 수행되고 있음.

(Knowledge-based Expert System), 학습공동체(Communities of Practice) 등은 전체 25개의 지식공유 방법론을 대상으로 수행한 효과성 및 사용빈도 조사에서 가장 낮은 점수를 기록함 (Egbu & Botterill, 2002). ▪ 국내에서 활용되고 있는 대표적인 건설산업의 지식정보 서비스 시스템으로는 건축도시연구정보센터(Architecture & Urban Research Information Center; AURIC)와 토목연구정보센터(Civil Engineering Research Information Center; CERIC)가 존재함. 건설산업과 관련된 국내외의 학술논문, 정기간행물, 연구보고서, 전문서적, 최신기술 등을 포괄적으로 제공하고 있지만, 실질적으로는 데이터 저장소로서의 기능만을 수행할 뿐 사용자들이 필요로 하는 지식정보를 생산하기에는 효율적이지 않다는 지적이 있음 (건설산업연구원, 2013). ▪ 미국의 건설산업연구원(Construction Industry Institute; CII)에서는 미국의 해외건설공사 발주자와 계약자들이 사업성과를 높이는 것을 지원하기 위해 International Project Risk Assessment(IPRA) 모델을 개발하여 운영 중임. 하지만 위험요인의 중요도를 평가하는 과정 중 대부분이 전문가의 주관적인 판단에 의거한다는 점이 한계로 지적됨.	▪ 건설 규정집에 등장하는 문장에서 설명하고 있는 대상(문/계단/천장 등), 특징(높이/너비/재료 등), 수치, 단위 등의 조건들을 정형화하는 연구가 다수 수행되었으며, 이렇게 정형화된 데이터들은 시공 결과의 부적합 여부 확인(Compliance Checking) 시 인풋 데이터로써 사용됨(Ilal & Gunaydin, 2017; Salama & El-Gohary, 2016). ▪ 영상 데이터를 기반으로 시공 현장에서 작업 중인 건설 중장비의 생산성 및 안전성을 분석함으로써 프로젝트 운영에 필요한 지식정보를 자동으로 생산해내고 이를 통해 안전뿐만 아니라 프로젝트 수행과 관련된 위험요인을 관리하는 연구가 수행됨(Kim & Chi, 2017). ▪ 국도교량유지관리시스템(Bridge Management System; BMS)에 탑재되어 있는 교량시설물의 손상발생 이력 및 점검 이력 데이터를 활용하여 교량 유지관리 실무에 필요한 의사결정에 도움을 주는 지식정보 생산 연구가 수행되고 있음.

<table>
<tr><th>구분</th><th>연구 범위</th><th>세부 연구 범위</th><th>구체적인 내용</th></tr>
<tr><td>1차 년도</td><td>딥러닝 기법을 활용한 의미기반 검색 알고리즘 개발</td><td>지식정보 서비스 예상 고객 사전 모니터링</td><td>

○ 예상 주요 고객 조사

▪ 본 연구의 범위인 기획 단계(해외건설 위험요인), 시공 단계(건설현장 안전 위험요인), 유지관리 단계(시설물 유지관리 위험요인) 에 포함되는 예상 수요자를 문헌고찰을 통해 조사함.

○ 사전 설문조사 및 인터뷰 수행

▪ 본 연구 범위에 포함되는 예상 수요자 중 다음을 대상으로 인터뷰를 실시하였음.

<table>
<tr><th>소속</th><th>사업단계</th></tr>
<tr><td>한미글로벌</td><td>기획</td></tr>
<tr><td>대림산업</td><td>시공</td></tr>
<tr><td>삼보기술단</td><td>시공</td></tr>
<tr><td>한국도로공사</td><td>유지관리</td></tr>
</table>

<표 3-1. 인터뷰 대상>

○ 설문조사 결과 분석

▪ 설문조사 결과 파악한 시스템 필수요소는 다음과 같음.

- 질문자의 의도를 파악하여 검색 성능을 향상시키기 위한 Semantic 검색 기술
- 각종 문서 작성 형식 및 매뉴얼
- 공고문(Invitation To Bid, ITB)등 데이터 재분류 / 재구조화 / 재구성
- 사용자의 검색 요구에 대한 유사 사례 제공 기술 및 의사결정 지원 정보 압축 제공을 위한 검색 결과 요약 기술
- 설계/시공 전 사전조사를 위한 유사 사례 검색 및 제공 기술

</td></tr>
</table>

구분	연구 범위	세부 연구 범위	구체적인 내용
		위험요인 관련 비정형 텍스트 데이터베이스(DB) 구축	

○ 건설산업에 존재하는 위험요인 관련 비정형 텍스트데이터 현황 분석

- 건설산업에 존재하는 위험요인 관련 비정형 텍스트데이터 현황은 다음과 같음.

건설 프로젝트 생애주기	위험요인 관련 텍스트데이터
기획	공사표준계약서, 공사견적서, 과업지시서, 시공계획서, 원가산출서 등
설계 및 시공	공사 계획서, 공사시방서, 공사감독일지, 공사실적내역, 공사일지, 공사현장안전관리수칙, 공정관리점검표, 설계변경요청서 등
운영 및 유지관리	공사완료보고서, 준공검사조서, 하자보수요청서 등

<표 3-2. 건설산업에 존재하는 위험요인 관련 비정형 텍스트데이터 현황>

○ 건설산업의 위험요인 관리에 적합한 DB 구조 설계

- 전통적인 정형 데이터 대상 RDBMS(Relational DataBase Management System) 대신 비정형 텍스트데이터를 다룰 수 있는 NoSQL(Not Only SQL) 과 같은 데이터베이스를 선택하여 데이터베이스 설계를 시도하였음.

○ 위험요인 관련 비정형 텍스트데이터 수집 및 DB 구축

- 기획, 시공, 유지관리 단계별로 발생하는 데이터의 형태가 상이하므로 각 단계마다 별도로 DB를 구축하여 데이터를 수집함.

단계	수집 데이터
기획	시장 분석 보고서 (Engineering News-Record; ENR, Construction Intelligence Center; CIC)
시공	안전 사고사례 (건설안전정보시스템)
유지관리	시설물 점검 보고서 (교량 정밀점검 및 정밀안전진단 보고서)

<표 3-3. 사업단계별 수집 데이터 목록>

구분	연구 범위	세부 연구 범위	구체적인 내용
		빅데이터 오픈소스 기술을 활용한 데이터베이스(DB) 자동 확장	

○ 빅데이터 오픈소스 기술인 웹크롤링을 활용한 웹상에 존재하는 건설산업 위험요인 관련 비정형 데이터 수집

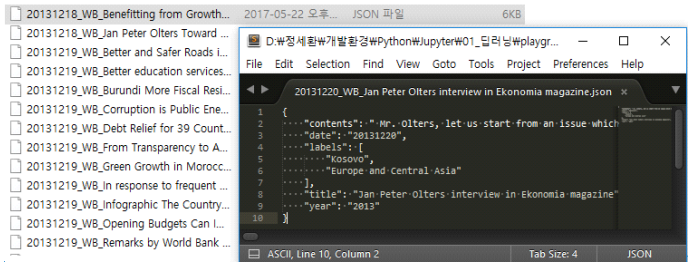
- 웹상에 존재하는 비정형 텍스트데이터를 수집하기 위해 빅데이터 오픈소스 기술 (웹크롤링)을 사용함.

단계	수집 데이터
기획	뉴스 기사 (Worldbank)
시공	안전 사고사례 (안전보건공단)
유지관리	뉴스 기사

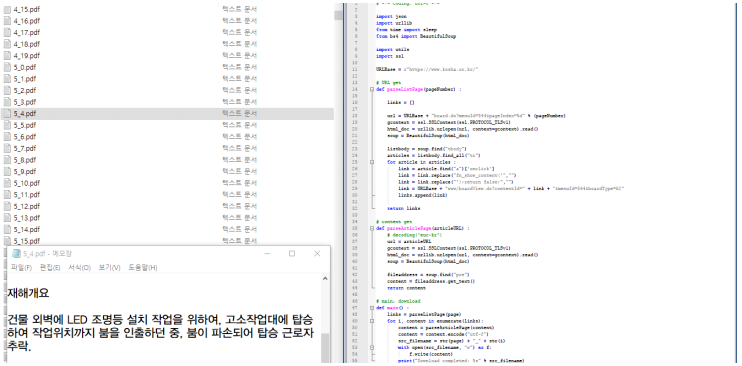
<표 3-4. 사업단계별 웹크롤링을 활용한 데이터 수집 목록>

○ 데이터 입력 자동화를 통한 DB 확장

- 비정형 텍스트데이터의 입력을 위해 위에서 설계한 DB 구조에 맞추어 json 형식으로 데이터를 저장하였음.
  - json 형식은 인터넷에서 자료를 주고받을 때 폭넓게 활용되는 방법이며, 속성-값 쌍으로 이루어진 데이터 오브젝트를 저장하기 위한 자료 형식임.
  - 특히 json 형식은 자료의 종류에 제한이 없다는 특징이 있어 비정형 텍스트데이터 입력에 적합함.



<그림 3-1. Worldbank 크롤링 데이터 수집 예시>

<table>
<tr><th>구분</th><th>연구 범위</th><th>세부 연구 범위</th><th>구체적인 내용</th></tr>
<tr><td></td><td></td><td></td><td> 

<그림 3-2. 건설현장 안전 사고사례 데이터 크롤링 예시></td></tr>
<tr><td></td><td></td><td>딥러닝 기법을 활용하여 사용자 검색어 의미 파악을 위한 코퍼스 구축</td><td>

○ 사용자 검색어의 의미기반 처리를 위하여 딥러닝 기법을 활용한 자연어 분석 기술 개발

▪ 단어의 의미기반 처리를 위하여 Word Embedding 기법과 관련된 자연어 분석 기술 문헌고찰을 수행함.

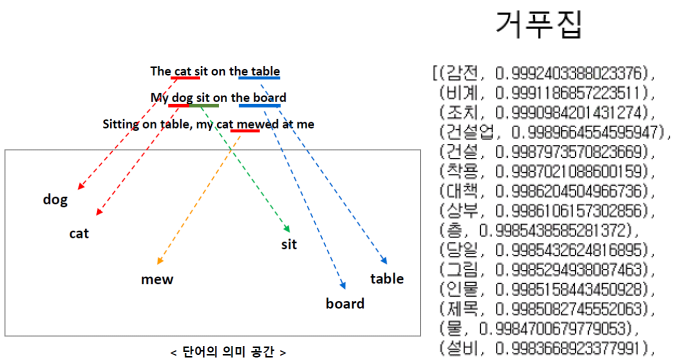
- Word Embedding이란 단어의 특정 정보(주변 단어의 맥락, co-occurrence 등)를 보존하면서 단어를 저차원의 실수 벡터로 표현하는 기법임.

<table>
<tr><th>알고리즘</th><th>특징</th></tr>
<tr><td>Word2Vec</td><td>단어 주변에 등장하는 다른 단어들의 분포의 유사성을 보존</td></tr>
<tr><td>Doc2Vec</td><td>Word2Vec 알고리즘에 더해 단어가 나타난 문서 정보를 맥락에 포함시킴</td></tr>
<tr><td>GloVe</td><td>단어의 co-occurrence frequency를 보존</td></tr>
<tr><td>Fasttext</td><td>Word2Vec이 응용된 모델이며, 노이즈(오타 등)가 많은 데이터에 적합</td></tr>
</table>

<표 3-5. Word Embedding 알고리즘>

</td></tr>
</table>

구분	연구 범위	세부 연구 범위	구체적인 내용
			▪ 본 연구는 Word2Vec 알고리즘을 적용하여 건설 문서에서 나타나는 단어들의 의미 유사도를 정성적으로 검증하였음.



<그림 3-3. Word2Vec 알고리즘 적용 예시>

○ 건설산업 특유의 언어적 특징을 반영할 수 있는 코퍼스 구축

▪ 후술할 tokenizing 및 Word2Vec 알고리즘을 통해 자연어 전처리 과정을 거친 코퍼스를 구축한 결과는 다음과 같음.

단계	데이터 수	용량
기획	12,000 이상 (계속 확장 중)	1.34GB
시공	3,849	1.07GB
유지관리	27,728	216GB

<표 3-6. 코퍼스 구축 결과>

<table>
<tr><th>구분</th><th>연구 범위</th><th>세부 연구 범위</th><th>구체적인 내용</th></tr>
<tr><td></td><td>맞춤형 검색 결과 요약 및 지식정보 생산 시스템 개발</td><td>머신러닝 기법을 활용한 건설산업에 특화된 자연어 분석 및 생성 모델 개발</td><td>

○ 건설산업에 존재하는 비정형 텍스트데이터의 언어적 특징 분석

▪ Domain-specific한 단어들이 자주 사용되는 데 반해 해당 단어들의 목록을 체계적으로 정리한 자료가 없어 통상의 ‘단어 사전 구축-형태소 분석을 통한 자연어 처리’의 자연어 처리 기법을 적용하기 어려움

▪ 보고서, 뉴스, 사례집 등 공적인 기관에 의해 작성되는 자료가 많아 상대적으로 노이즈(띄어쓰기 오류, 오탈자 등)가 SNS 데이터 등에 비해 적음

○ 수집된 데이터 기반 건설 문서의 효율적이고 정확한 자연어 처리를 위한 머신러닝 기법 비교·분석 및 적용

▪ 건설산업 문서 텍스트의 특징 및 한국어 텍스트의 특징을 고려하여 다음의 머신러닝 기반 자연어 처리 알고리즘을 비교·분석 및 적용하였음.

<table>
<tr><th>알고리즘</th><th>원리</th></tr>
<tr><td>Cohesion Probability</td><td>음절 간 결합도가 클수록 해당 음절들이 한 단어일 확률이 큼</td></tr>
<tr><td>Branching Entropy</td><td>단어의 경계 부분에서는 이전/다음 글자의 불확실성(Entropy)이 증가하므로 Entropy가 급격히 증가하는 부분이 단어의 경계일 가능성이 높음</td></tr>
<tr><td>L-R tokenizer</td><td>한국어 단어는 의미부와 문법부로 구성되며 텍스트 분석에 실질적으로 필요한 의미부는 대부분 단어의 왼쪽에 존재함</td></tr>
</table>

<표 3-7. 자연어 처리를 위한 머신러닝 기법 비교>

</td></tr>
</table>

<table>
<tr><th>구분</th><th>연구 범위</th><th>세부 연구 범위</th><th>구체적인 내용</th></tr>
<tr><td></td><td></td><td></td><td>

○ 건설산업에 특화된 자연어 처리를 위한 머신러닝 기반 모델 개발

- 수집한 코퍼스를 대상으로 머신러닝 기반 자연어 처리 알고리즘을 적용함으로써 코퍼스에 최적 fit의 자연어 처리 모델을 개발함.
- 모델의 성능은 추출되는 단어의 결과를 바탕으로 정성적으로 검증함.

<table>
<tr><th>균열</th><th>누수</th><th>백태</th></tr>
<tr><td>주입</td><td>신축이음</td><td>바닥판</td></tr>
<tr><td>폭</td><td>하부</td><td>캔틸레버</td></tr>
<tr><td>건조수축</td><td>교면</td><td>하면</td></tr>
<tr><td>온도</td><td>배수</td><td>우수유입</td></tr>
<tr><td>사용</td><td>A2</td><td>흔적</td></tr>
<tr><td>측정</td><td>설치</td><td></td></tr>
<tr><td></td><td>교체</td><td></td></tr>
<tr><td></td><td>본체</td><td></td></tr>
<tr><td></td><td>노후화</td><td></td></tr>
</table>

기존 tokenizer (twitter)

EPS T TF V 가스 걸고 검진 경보 고르 고수부지 관 광 구 기 기구 기반
단자 데크 땅 로 리 막 메우기 물 미타 발판 방지 방호 배합 백 벨브 벨브실
변압 변전 본 빔 설 스위치 시설 실 알곤 알람 앵카 역 역류 열전 영구 예
오니 오탁 와이어 용 우프 유공 음 음부 이 이스트 인용 작업 잔 저장 적 조
조물 조치 콘넥터 큐비 크로스 클 트 패널 풋 플레이 함 형 호 홈 회전 후지

개발된 tokenizer

미타설 검진기구 광단자함 배합작업 백열전 큐비클(변압기반) 앵카시설
데크플레이트 콘넥터 영구용 와이어로우프 고수부지 알람벨브실
오니저장조 이음부 예인용 크로스빔 홈메우기 변전실 땅고르기
오탁방지막 TF관 V벨브 호이스트 걸고리 본구조물 역회전방지용 역T형
풋스위치 잔적물 EPS패널 역류방호조치 유공발판 알곤가스 후지경보음

<그림 3-4. 머신러닝 기반 자연어 처리 알고리즘 적용 예시>

</td></tr>
<tr><td>2차 년도</td><td>딥러닝 기법을 활용한 의미기반 검색 알고리즘 개발</td><td>의미기반 검색 알고리즘 개발</td><td>

○ 코퍼스 기반 자연어 분석 기술로 처리된 검색어의 의미 파악 기술 개발

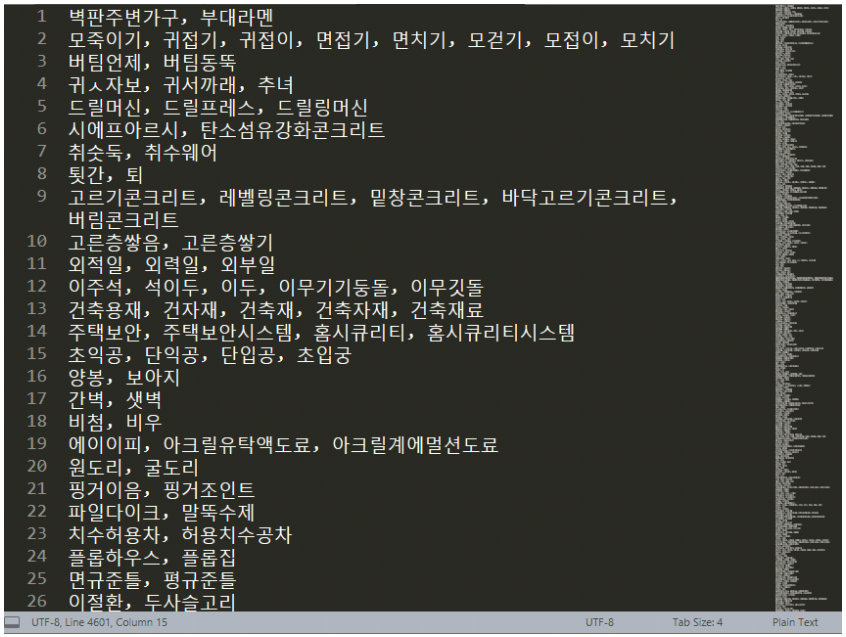
- 건설 분야 문서에서 사용되는 용어들의 사전인 ‘건설 시소러스(Construction thesaurus)’를 구축함.
  - 시소러스(Thesaurus)란 특정 분야 용어들의 유의관계, 상·하의관계 등을 표현하는 단어 사전의 일종임.
  - 국립국어원에서 발간하는 토목 및 건축 용어 사전을 바탕으로 건설 유의어 사전, 약어 사전, 상의어 사전, 하의어 사전, 참조어 사전을 구축함 (ref. – 국립국어원).
- 또한 Word2Vec을 통해 생성된 단어 벡터 간 코사인 유사도(Cosine similarity)를 바탕으로 사용되는 맥락이 비슷한 단어들을 묶은 ‘Word2Vec 시소러스’를 구축함.

</td></tr>
</table>

구분	연구 범위	세부 연구 범위	구체적인 내용

Thesaurus	예시
(1) 건설 시소러스	
유의어	콘크리트 - 혼응토 / 진압롤러 - 다짐롤러
약어	모르타르 - 몰탈 / 동바리 - 동발
상의어	건축용재 - 골재, 내장재, 벽동
하의어	발코니 - 바닥 / 외벽 - 벽
참조어	용적률 - 건폐율 / 취수탑 - 배수탑
(2) Word2Vec 시소러스	
Word2Vec	슬래브 - 슬라브 - SLAB - Slab / 타워크레인 - T/C

<표 3-8. 시소러스 및 예시 단어>



```
벽판주변가구, 부대라멘
모죽이기, 귀접기, 귀접이, 면접기, 면치기, 모걷기, 모접이, 모치기
버팀언제, 버팀동뚝
귀ㅅ자보, 귀서까래, 추녀
드릴머신, 드릴프레스, 드릴링머신
시에프아르시, 탄소섬유강화콘크리트
취숫둑, 취수웨어
툇간, 퇴
고르기콘크리트, 레벨링콘크리트, 밑창콘크리트, 바닥고르기콘크리트,
버림콘크리트
고른층쌓음, 고른층쌓기
외적일, 외력일, 외부일
이주석, 석이두, 이두, 이무기기둥돌, 이무깃돌
건축용재, 건자재, 건축재, 건축자재, 건축재료
주택보안, 주택보안시스템, 홈시큐리티, 홈시큐리티시스템
초익공, 단익공, 단입공, 초입궁
양봉, 보아지
간벽, 샛벽
비첨, 비우
에이이피, 아크릴유탁액도료, 아크릴계에멀션도료
원도리, 굴도리
핑거이음, 핑거조인트
파일다이크, 말뚝수제
치수허용차, 허용치수공차
플롭하우스, 플롭집
면규준틀, 평규준틀
이절환, 두사슬고리
UTF-8, Line 4601, Column 15        UTF-8        Tab Size: 4        Plain Text
```

<그림 3-5. 건설 유의어 사전 예시>

구분	연구 범위	세부 연구 범위	구체적인 내용
			▪ 구축된 시소러스를 바탕으로, 입력된 검색어와 의미적으로 유사한 다른 단어들을 함께 검색함으로써 검색어의 의미를 검색 알고리즘에 반영하는 검색어 확장(Query expansion) 방법을 적용함. - 예를 들어, “슬래브”라는 단어를 검색할 경우, “슬래브”라는 단어가 포함된 문서를 검색할 뿐만 아니라, “슬래브”와 유사한 “슬라브”, “SLAB”, “Slab” 등을 포함하는 문서들을 함께 검색함으로써, 검색어와 문서 간 단순 단어 일치 여부만을 판단하는 것이 아닌 검색어의 의미를 파악하는 검색 결과를 출력함. ○ **신뢰성 있는 검색 결과를 제공하기 위하여 검색어와 데이터베이스 내부의 데이터 간의 유사도 계산을 하는 검색 알고리즘 개발** ▪ 검색어와 유사도가 높은 문서일수록 검색어와 관련이 높다고 할 수 있으며, 유사도가 높은 문서가 검색 결과의 상위에 출력될수록 검색 결과의 신뢰성이 증가함. ▪ 검색어와 DB 내부 문서 데이터 간 유사도 계산을 위해 Okapi BM25 방법론을 응용함. - Okapi BM25는 TF-IDF의 확장 버전으로, 특정 검색어가 많이 사용된 문서일수록 해당 단어의 Term frequency (TF) 가 높아져서 해당 검색어와 문서 간 유사도가 증가함. - 반면, 예를 들어 건설현장 안전 사고사례 DB에서 “건설”, “안전”과 같은 단어를 검색하는 경우, 해당 단어들은 DB 내부의 거의 모든 문서에서 등장하기 때문에 검색어로서의 중요도가 낮음. - DB 내 문서 중 특정 단어를 포함하는 문서가 많을수록 해당 단어의 Document frequency (DF) 는 증가하며, DF의 역수인 Inverse document frequency (IDF)를 통해 해당 단어의 가중치를 감소시킴. $$score(D,Q)=\sum_{i=1}^{n} idf(q_1)\frac{f(q_i,D)(k_1+1)}{f(q_i,D)+k_1(1-b+b\frac{\lvert D\rvert}{avg.dl})}$$ <수식 3-1. Okapi BM25>

구분	연구 범위	세부 연구 범위	구체적인 내용

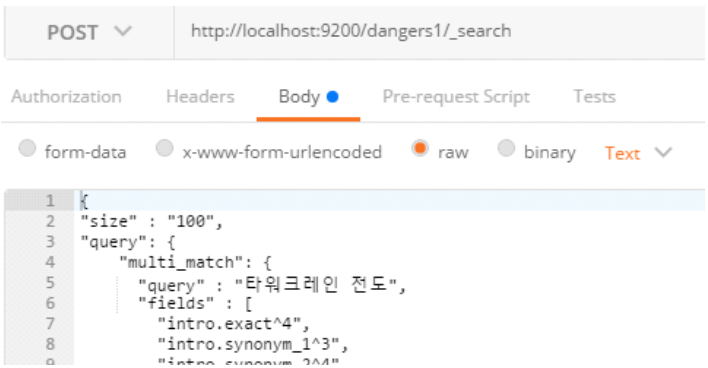
▪ 본 연구에서는 Okapi BM25의 유사도 산정 식에 Thesaurus weight를 추가함으로써 검색어 확장에 의해 추가되는 검색어들의 중요도를 유사도 계산에 반영함.

$$score(D, Q) = \sum_{i=1}^{n} idf(q_1) \frac{f(q_i, D)(k_1 + 1)}{f(q_i, D) + k_1(1 - b + b\frac{|D|}{avg.dl})} \cdot Thesaurus\ weight$$

<수식 3-2. 수정 Okapi BM25>

Thesaurus	의미관계			가중치
	유의관계	상·하위관계	참조관계	
Construction thesaurus				
유의어	○			4
약어	○			4
상의어		○		3
하의어		○		3
참조어			○	2
Word2vec thesaurus				
Word2vec	○	○	○	3

<표 3-9. Thesaurus weight>



<그림 3-6. 검색 알고리즘 구현 예시>

구분	연구 범위	세부 연구 범위	구체적인 내용
		검색 알고리즘 검증 및 평가	○ **전문가 자문 및 설문조사를 통한 알고리즘의 타당성 검토 및 성능 검증** ▪ 16명의 건설 분야 종사자들을 대상으로 본 연구에서 개발한 검색 알고리즘에 의한 건설현장 안전 사고사례 검색 결과의 타당성을 검증함. - 기획단계 및 유지관리 단계에서 생성되는 문서 데이터가 대부분 분량이 긴 보고서이므로 검색어와 문서 간 연관성을 직관적으로 파악하기 어려운 데 반해, 안전 사고사례는 상대적으로 내용이 간략하기 때문에 사용자 입장에서 검색어와 검색 결과 간 연관성을 판단하기 용이함. - 따라서 본 연구의 설문조사에서는 건설안전정보시스템으로부터 수집한 안전 사고사례 데이터를 활용하여 본 연구 검색 알고리즘의 타당성을 검증함. ▪ 설문조사 결과, 본 연구의 검색 알고리즘에 의한 검색 결과는 검색어와의 연관성이 높은 것으로 확인되었으며, 특히 검색 결과의 상위에 위치한 문서일수록 검색어와의 연관성이 높은 것을 확인함. ○ **타 검색엔진을 이용한 건설 산업 위험요인 검색 결과 비교를 통한 검색 알고리즘의 정량적 성능 평가** ▪ 검색 알고리즘의 검색 성능을 정량적으로 평가하기 위해, 본 연구는 Normalized discounted cumulated gain (NDCG) 방법을 활용함. - 검색 결과로 출력된 문서 중 검색어와 연관성이 큰 문서가 상위에 위치해 있을수록 Discounted cumulated gain (DCG) 이 증가하며, 반대로 연관성이 낮은 문서가 상위에 위치하거나 연관성이 큰 문서가 하위에 위치할 경우 DCG는 감소함. $$DCG_P = \sum_{i=1}^{P} \frac{2^{rel_i} - 1}{\log_2(1+i)}$$ <수식 3-3. DCG>

<table>
<tr><th>구분</th><th>연구 범위</th><th>세부 연구 범위</th><th>구체적인 내용</th></tr>
<tr><td></td><td></td><td></td><td>- NDCG는 모든 문서가 관련성이 높은 순서대로 출력된 이상적인 경우의 Ideal DCG (IDCG) 에 대해 DCG 값을 정규화함으로써 검색 알고리즘의 성능을 0~1 스케일로 표현하는 방법임.

▪ 다양한 검색어의 조합에 대해 NDCG는 평균적으로 94~98% 정도로 산정되었으며, 따라서 본 연구에서 개발한 의미 기반 검색 알고리즘은 다양한 검색어에 대해 관련성이 높은 문서들을 효과적으로 검색할 수 있음을 확인함.
<table>
<tr><th>검색어</th><th>교량 콘크리트 타설 작업 중 붕괴 추락</th><th>터널 천공작업 붕괴 매몰</th><th>가설공사 이동작업 중 작업발판 추락</th><th>하수도 이설작업 토사 붕괴 매몰</th></tr>
<tr><td>평균 NDCG</td><td>98%</td><td>98%</td><td>94%</td><td>98%</td></tr>
</table>
<표 3-10. 검색어 별 NDCG 결과></td></tr>
<tr><td></td><td>맞춤형 검색 결과 요약 및 지식정보 생산 시스템 개발</td><td>로봇저널리즘 기법을 활용한 검색 결과 자동 분석 및 요약문 생성 기술 개발</td><td>○ 사용자 검색 결과의 문서 내용을 자동으로 비교·분석하여 그래프, 도표 등의 시각화 정보로 요약하는 기술 개발

▪ 검색 결과의 문서 내용을 자동으로 비교·분석하기 위해 문서에서 자동으로 위험요인을 도출하는 다양한 방법론을 비교·분석 및 적용함.
- 위험요인은 문서의 성격에 따라 다양하게 정의됨. 예를 들어, 건설현장 안전 사고사례에서 도출할 수 있는 위험요인은 위험발생객체, 위험발생위치, 작업프로세스, 사고결과 등이 있으며, 시설물 점검 보고서에서 도출할 수 있는 위험요인은 손상 발생 위치(부재), 손상 종류, 인자 등이 있음.
▪ 규칙 기반(Rule-based) 위험요인 도출 방법론은 미리 정의된 규칙을 바탕으로 문서에서 위험요인을 도출하는 방법임.
- 예를 들어, “~ 작업중” 이라는 단어 앞에 위치한 단어는 ‘작업프로세스’에 해당한다는 규칙으로부터, “아파트 현장에서 비계공 0명이 측벽 외부비계 해체 작업중 (중략)” 과 같은 문장에서 “(측벽 외부비계) 해체”가 ‘작업프로세스’에 해당한다는 것을 판별하는 방법임.</td></tr>
</table>

구분	연구 범위	세부 연구 범위	구체적인 내용

- 규칙 기반 방법론은 규칙이 잘 정의되어 있고 문서의 형식이 획일화된 경우 높은 인식 정확도를 보이나, 규칙을 정의하는 과정에 인력과 시간이 많이 소요될 뿐만 아니라 다양한 형식의 문서에 대해 일률적인 규칙을 적용하기 어렵다는 한계점이 있음.

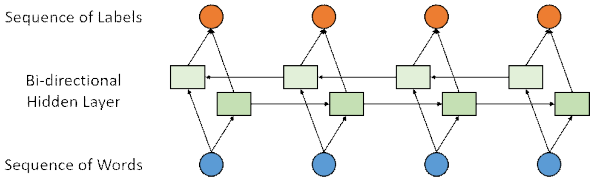
Rules	Tacit Knowledge (Semantic Role)			
	Hazard Object (Effector)	Hazard Position (Location)	Work Process (Purpose)	Accident Result (Predicate)
Critical Rule	{A이} or {A가} (A is/are)	{A에} or {A에서} ({to/at/on … A})	{A} + [중] or [작업] ([during] {A} or {A} [work])	{A된} or {A하여} or {A되} or {A됐} or {A됨} or {A하고} or {A한} or {A하다가} (The verb states that the result {A} is generated)
Plus Rules	{A를} + [밟다] ([step on] {A}) {A에} + [맞다/부딪히] ([be hit/bumped by] {A}) {A에} + [감전/접촉/균열/매몰] ([Electric shock to/ contact/ crack in/ be buried under] {A}) {A의} + [붕괴/침하/탈락] ([Collapse of/ subsidence of/ drop out of] {A}) {A에} + [사람신체] ([human body] with {A})	{A=-면} + [B이/B가] ~ ([B is/are] ~ {on - surface}) {A} + [주위에] ([around] {A}) {A을/를} + [이동] ([moing] {A})	{A진행} or {A하던} + [중], {A을/를} + [하던] + [중] ([in] {progress A})	{A} + [현상/발생] {A} [phenomenon / occurrence]
Minus Rules	A=[사람이름/숫자/그룹/사람신체/추상명사/단위] + {이} or {가} (A=[person name/number/group/human body/abstract noun/unit] is/are)	A=[사람신체/지시] + {에} or {에서} ({to} [human body] / [instructions]) {A에} + [맞다/부딪히] ([be hit/bumped by] {A}) {A에} + [감전/접촉/균열/매몰] ([Electric shock to/ contact/ crack in/ be buried under] {A}) {A에} + [사람신체] ([human body] {with A})		

Note. [ ] = Relatum, { } = Tacit knowledge

<표 3-11. 위험요인 인식 규칙 예시 – 건설현장 안전 사고사례 대상>

▪ 머신러닝 기반 위험요인 도출 방법론 중 본 연구에서는 Conditional random field (CRF) 방법과 Recurrent neural network (RNN) 방법이 있음.

- CRF는 특정 단어가 어떤 위험요인에 해당하는지 판별하기 위해, 직전에 사용된 단어 및 그 단어의 속성(class)를 바탕으로, 특정 단어의 속성(class)을 판별하는 방법임.

구분	연구 범위	세부 연구 범위	구체적인 내용
			- RNN은 특정 단어가 어떤 위험요인에 해당하는지 판별하기 위해 이전 단어들의 정보를 활용한다는 점은 CRF와 유사하나, 직전 단어뿐만 아니라 앞뒤에 있는 모든 단어들의 정보를 활용한다는 점과, 단어들의 정보를 Word embedding 등을 통해 추상화하여 활용한다는 차이점이 있음.  rnn_model.train(input_fn=lambda:input_fn(sentences=x_tokens_sample, labels=y_tokens_sample, input_mode='TRAIN', batch_size=1000), steps=100) Tensor("embedding_lookup:0", shape=(10, 48, 50), dtype=float32, device=/device:CPU:0) INFO:tensorflow:Create CheckpointSaverHook. 2018-04-26 13:59:21,396 : INFO : Create CheckpointSaverHook. <그림 3-7. RNN 개념도 및 구현 코드 샘플> ▪ 검색 결과로부터 도출된 위험요인을 한눈에 볼 수 있도록 Word cloud, 히스토그램, 파이 차트 등을 활용하여 주요 위험요인을 직관적으로 파악할 수 있도록 함. <그림 3-5. 검색 인터페이스 및 검색 결과 요약 시스템>

구분	연구 범위	세부 연구 범위	구체적인 내용

			○ **로봇저널리즘 기법을 활용하여 시각화 정보를 해석하고, 결과를 자동으로 요약해주는 기술 개발**
			
			▪ 텍스트데이터를 자동으로 요약하기 위해 우선 문서들을 유사한 주제별로 분류하는 토픽 모델링(Topic Modeling)을 수행함.
			- 토픽 모델링은 전체 문서집단을 유사한 주제에 따라 분류하는 기법임. 본 연구에서 사용한 알고지름은 토픽의 단어 분포와 문서의 토픽 분포가 결합하여 문서 내 단어들이 생성된다고 가정하는 잠재 디리클레 모형(Latent Dirichlet Allocation; LDA)으로써, 디리클레 분포에 기반하여 토픽별 단어의 분포, 문서별 토픽의 분포를 추정함.
			- 사용자가 문서집단에 존재하는 토픽의 개수(k)를 정하면, 모든 단어에 대해 임의의 토픽을 할당함. 이후에는 한 단어를 제외한 다른 모든 단어의 토픽을 고정한 상태에서 해당 단어의 토픽을 다시 계산함. 이러한 과정은 깁스 샘플링(Gibbs Sampling)을 통해 수행되며 토픽의 단어 분포가 조건부확률처럼 계산됨.
			- 깁스 샘플링은 변수의 결합확률분포를 알지 못 할 때 사용하는 확률분포 추정 기법임. 결합확률분포를 임의로 설정한 뒤, 주변의 변수를 고정시킨 채 변수를 하나씩 변경해가며 확률분포의 표본을 확보하다보면 표본의 분포가 전체 결합확률분포에 수렴한다고 알려져 있음.
			- 샘플링이 완료된 이후에는 각 문서의 모든 단어마다 어떤 토픽으로부터 등장했는지에 대한 확률(토픽의 단어 분포)이 존재함. 이러한 확률의 결합으로 각 문서의 토픽 분포를 계산할 수 있음(문서의 토픽 분포). 각 문서의 토픽 분포에 따라 확률이 높은 토픽 그룹에 문서를 할당하면 LDA가 완료됨.
			
			`docs["topic"] = doc_tp_assign(docs["tknd"], D, doc_tp_cnt)`
			
			`Assigning docs to topic ...11491 docs (1.000)`
			`doc-topic Assignment: done!!!`
			
			<그림 3-6. 토픽 모델링(LDA) 코드 실행 결과>

구분	연구 범위	세부 연구 범위	구체적인 내용
			▪ 토픽 별로 구분된 문서 집단에 대해 TextRank 알고리즘이라는 로봇저널리즘 기법을 사용하여 주요 문창을 추출함으로써 텍스트 요약을 수행함. - TextRank 알고리즘은 “다른 단어/문장과 강하게/많이 연결되어 있을수록 중요한 단어/문장일 것이다”라는 논리에 기반한 요약 방법임. - 전체 문서집단으로부터 모든 문장을 추출하고, 공통된 단어를 포함하는 문장들을 연결함. 공통된 단어가 얼마나 많이 등장하느냐에 따라 문장 간 유사도를 계산하고, 유사도가 높은 문장들과 많이 연결된 문장일수록 중요도가 높아지도록 중요도를 계산함. 마지막으로는 중요도가 높은 문장들을 추출하여 조합함으로써 요약문을 생성함. <그림 3-7. TextRank 구현 코드 및 실행 결과 예시>
		보안성 자료 수집 전략 수립	**○ 확보 데이터와 확보 가능한 데이터를 활용하여 제안한 방법론의 타당성 및 결과물의 활용 가능성 검증** ▪ 월드뱅크(Worldbank) 웹사이트에서 수집한 11,491건의 뉴스기사에 대해 토픽 모델링을 수행했으며, 토픽의 개수는 10개로 설정함. 토픽 모델링 결과, 첫 번째 문서의 토픽 비중은 표 3-10과 같음. 4번 토픽의 비중이 가장 높기 때문에 이 문서는 4번 토픽에 해당한다고 할 수 있음.

구분	연구 범위	세부 연구 범위	구체적인 내용

Topic	0	1	2	3	4	5	6	7	8	9
Score	0	10	0	0	308	0	2	16	0	0

<표 3-10. 문서의 토픽 비중>

- 4번 토픽의 단어 비중은 표 3-11.과 같음. buisiness, report, improve, bank, financial 등의 키워드가 상위에 위치한 것으로 보아 국제적인 금융 사업과 관련된 토픽임을 파악할 수 있음.
- 4번 토픽의 내용에 대해 TextRank 알고리즘을 적용해 요약한 결과는 그림 00.과 같음. 일반적인 학술 논문의 초록이 300자 내외인 것을 감안하여 요약문의 길이도 300자로 제한했으며, 그림 00.는 그 중 일부를 발췌한 것임. 개발도상국에 국제적인 경제적 지원이 활발하게 이루어지고 있다는 내용을 확인 할 수 있음.

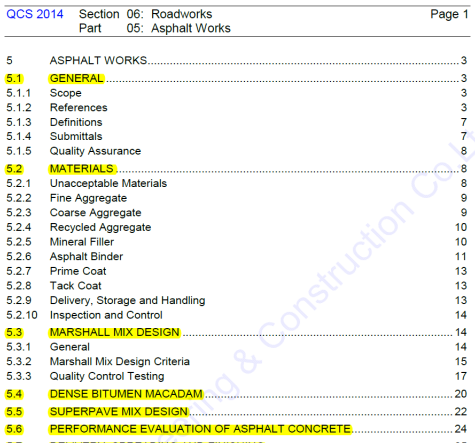
Rank	Keyword	Score
1	Business	6,604
2	Report	5,139
3	World	4,582
4	New	4,470
5	Data	4,137
6	(	3,747
7	Improve	3,597
8	Bank	3,454
9	)	3,085
10	Financial	3,006
...	...	...

<표 3-11. 토픽의 단어 비중>

...two projects approved today aim boosting economic opportunities eastern caribbean improving education social protection services young people saint vincent grenadines creating new opportunities farmers fishers grenada saint vincent grenadines linking larger markets products said tahseen sayed country director caribbean. project facilitate social inclusion poor vulnerable rural communities lima december 11th 2015 board executive directors approved us 50 million financing today maintain 5,000 kilometers rural roads rehabilitate 2,200 additional kilometers intention improving country competitiveness increasing access markets public services among poor remote rural populations...

<그림 3-6. TextRank 결과>

구분	연구 범위	세부 연구 범위	구체적인 내용
			○ 검증된 결과를 바탕으로 보안성 자료에 대한 접근 전략 수립 및 데이터 수집 ▪ 뉴스기사 데이터만으로는 건설 프로젝트의 위험요소를 충분히 파악할 수 없고 대략적인 정보만을 확인할 수 있기 때문에, 결과물의 활용성을 위해서는 보안성 자료의 역할이 중요함. 현재 D건설, C연구원, H정부기관 등 건설산업의 다양한 실무 기관들과의 협업을 통해 교량유지관리 보고서, 해외건설공사 시방서, 건설현장 안전사고 보고서 등 보안성 자료를 일부 수집함. 추후에는 이렇게 수집한 자료를 바탕으로 본 연구 결과를 고도화할 예정임.



QCS 2014 Section 06: Roadworks Page 1
Part 05: Asphalt Works

5 ASPHALT WORKS 3
5.1 GENERAL 3
5.1.1 Scope 3
5.1.2 References 3
5.1.3 Definitions 7
5.1.4 Submittals 7
5.1.5 Quality Assurance 8
5.2 MATERIALS 8
5.2.1 Unacceptable Materials 8
5.2.2 Fine Aggregate 9
5.2.3 Coarse Aggregate 9
5.2.4 Recycled Aggregate 10
5.2.5 Mineral Filler 10
5.2.6 Asphalt Binder 11
5.2.7 Prime Coat 13
5.2.8 Tack Coat 13
5.2.9 Delivery, Storage and Handling 13
5.2.10 Inspection and Control 14
5.3 MARSHALL MIX DESIGN 14
5.3.1 General 14
5.3.2 Marshall Mix Design Criteria 15
5.3.3 Quality Control Testing 17
5.4 DENSE BITUMEN MACADAM 20
5.5 SUPERPAVE MIX DESIGN 22
5.6 PERFORMANCE EVALUATION OF ASPHALT CONCRETE 24

REPORT (NCR)

NCR No: EXW-P023-0004-QM-AEC-NC-00356 | Date Opened: 15-07-2017 | Date Closed: 03-07-2017

PART A: DETAILS OF NON-CONFORMANCE

Details:
Total thickness of Asphalt paving course (combined) is less than specified. As per QCS 2014 Section 6 Part 5 Clause 5.11.1 (1) at MCW, South Bound, Ch. 119+129 Lane 3 & 119+140 Lane 4. , Vide EXW-P023-0004-QM-DEC-BL-72166.

Classification of Defect:
[ ] Major [x] Minor [ ] Unknown

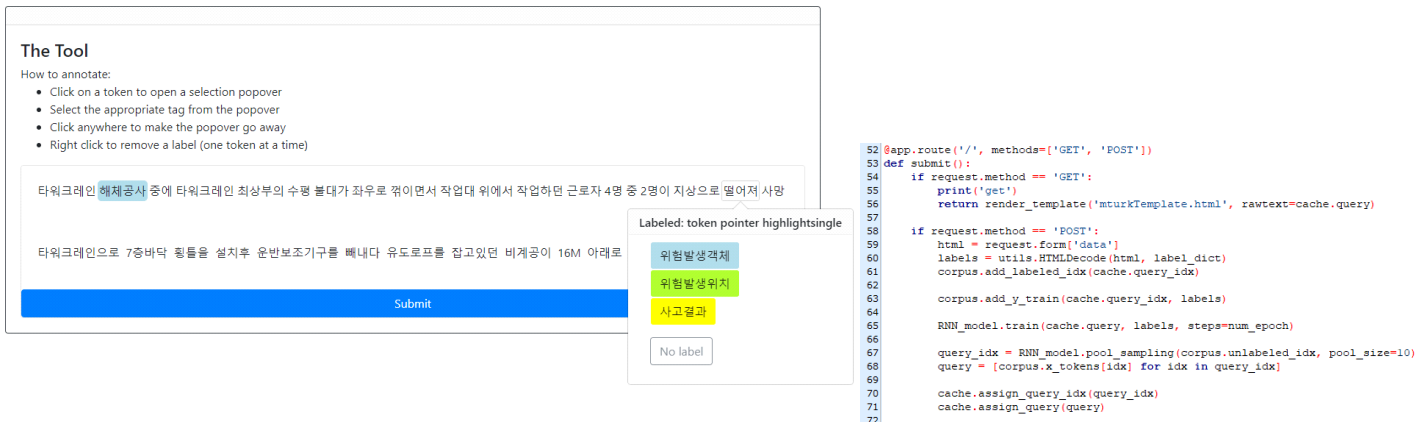
Cause of Non-Conformance:
[ ] Material [x] Process [ ] Equipment [ ] Documentation

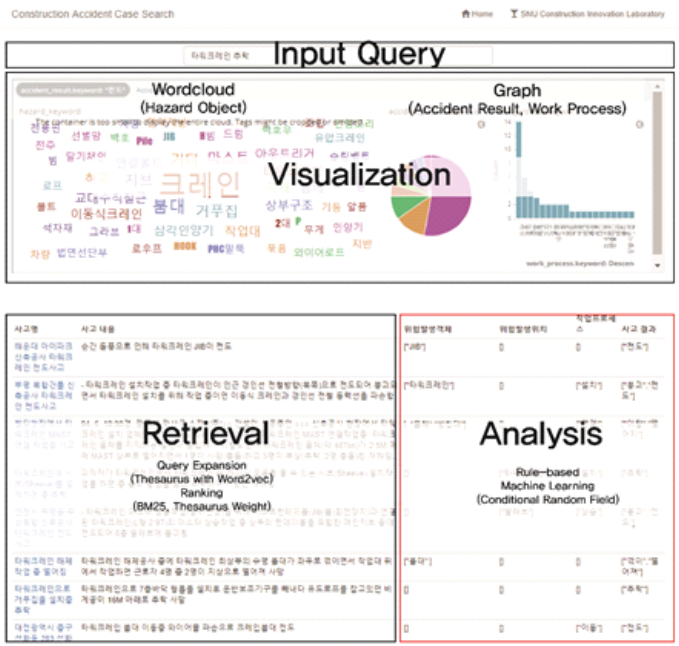
Discipline:
[x] Roads [ ] Structures [ ] Utilities [ ] Quality [ ] Others (Please Specify)

Documentation Ref. (Specifications, ITP, Drawings, Standard, etc.,):
1. QCS 2014 Section 6 Part 5 Clause 5.11.1 (1)
2. Test Reports.

Name & Signature: S.M. Hilmy | Date:13-07-2017

<그림 3-7. 해외건설공사 시방서 및 건설현장 안전사고 보고서 샘플>

구분	연구 범위	세부 연구 범위	구체적인 내용
		지식정보 축적 및 제공을 위한 사용자 인터페이스 개발	○ **사용자에게 제공한 지식정보를 데이터베이스에 자동으로 축적** ▪ 검색 결과의 문서 데이터에 존재하는 위험요인 도출 결과가 잘못된 경우 사용자의 피드백을 반영하기 위한 Active learning 프로세스를 정립함. - 머신러닝 기반 위험요인 도출 모델은 학습에 사용하는 데이터의 양이 증가할수록 인식 성능이 증가하며, 특히 모델 입장에서 인식을 어려워하는 데이터가 제공될수록 성능 향상 폭이 증가함. - Active learning 방법은 모델의 인식 확신도(Recognition confidence)가 낮은 데이터를 우선적으로 선택하여 사용자에게 피드백을 요청함으로써, 데이터 축적에 의한 모델의 성능 향상 효율을 극대화함. - 모델의 예측 불확실성이 높을수록 인식의 확신도(confidence)가 낮으며, 예측 불확실성은 엔트로피(Entropy)를 통해 산정함. 엔트로피가 클수록 모델 입장에서 인식을 어려워하는 데이터라 할 수 있으므로, 엔트로피가 큰 데이터에 대해 우선적으로 사용자에게 피드백을 요청함. ▪ 사용자가 위험요인 도출 결과에 대해 손쉽게 피드백을 줄 수 있도록 본 연구에서는 사용자가 직접 위험요인을 수정할 수 있도록 하는 GUI 기반 인터페이스를 개발함.  <그림 3-7. 위험요인 도출 결과 피드백을 위한 GUI 인터페이스 및 인터페이스-모델 간 통신 모듈>

구분	연구 범위	세부 연구 범위	구체적인 내용
			▪ 인터페이스를 통한 사용자 피드백은 서버 측의 위험요인 인식 모델로 전달되며, 모델은 사용자의 피드백을 바탕으로 지속적으로 인식 성능을 향상시킴. ○ 사용자 맞춤형 지식정보 제공을 위한 사용자 인터페이스 개발 ▪ 본 연구는 의미기반 검색 알고리즘에 의한 검색 결과 및 검색 결과의 요약 내용을 사용자에게 맞춤형으로 제공하는 사용자 인터페이스를 개발함. - 검색 시스템의 기본 골격은 오픈소스 검색 엔진인 Elasticsearch를 활용하였으며, 시각화 도구는 Kibana에서 제공하는 Wordcloud, 히스토그램, 파이차트 제작 도구를 활용함.  <그림 3-8. 지식 제공을 위한 사용자 인터페이스>

구분	연구 범위	세부 연구 범위	구체적인 내용
		시스템 검증 및 평가	▪ 위의 인터페이스는 로컬호스트(개별 데스크톱에서만 구현되는 일종의 웹페이지) 기반으로 만들어졌으며, 현장의 기술자 등 수요자들이 연구결과를 쉽게 활용할 수 있도록 동영상을 제작함. 동영상에는 제공되는 서비스뿐만 아니라 사용자 인터페이스에 검색어를 입력하여 결과를 출력하는 과정과 결과물 페이지의 구성에 관한 설명이 담겨있기 때문에, 사용자를 교육하고 연구결과를 홍보하는데 활용될 수 있음. (동영상 별도 첨부) **○ 시스템의 활용 가능 시나리오 분석** ▪ 본 연구에서 개발한 의미기반 검색 알고리즘 및 검색 결과 요약 시스템은 기획, 시공 및 유지관리에 이르는 건설 프로젝트의 전 주기에 걸쳐 활용될 수 있음. ▪ 특히, 유지관리 단계에서 생성되는 시설물 점검 보고서로부터 도출한 위험요인 분석 결과는 추후 새로운 시설물 건설 사업의 초기 단계(기획, 설계 및 시공)에서 활용될 수 있음. - 예를 들어, 교량시설물에 발생한 주요 손상 및 주요 영향인자 분석 결과를 바탕으로, 해당 손상의 발생을 사전에 방지할 수 있도록 설계 및 시공 프로세스를 수립하거나 설계기준을 개선할 수 있음. **○ 자문단을 통한 사전 설문조사 및 자문 바탕으로 타당성 및 신뢰성 확보** ▪ 본 연구의 목적은 건설산업에 존재하는 경험 지식정보를 효율적·효과적으로 저장/관리/공유하기 위한 서비스 시스템을 개발하는 것이기 때문에, 결과물의 타당성 및 실효성을 확보하기 위해 실무자들을 대상으로 인터뷰를 실시함. ▪ 딥러닝·머신러닝 방법론을 활용한 빅데이터 분석 연구를 수행하고 있는 서울대학교 컴퓨터공학부, 산업공학과, 언론정보학과 소속 연구원들에게 기술적 타당성 검토를 위한 인터뷰를 수행하였고, 본 시스템에서 활용한 데이터와 알고리즘이 결과물로 이어지는 과정에 문제가 없다는 검증을 받음. 또한 본 연구가 활용한 검색 및 자동 요약 방법론을 타 분야에 적용하거나, 본 연구에서 저장/관리/공유하고자 하는 건설 프로젝트 전주기 위험요인 관련 지식정보가 리스크 커뮤니케이션 관리 등 다른 연구에서 활용될 수 있다는 것을 확인함.

구분	연구 범위	세부 연구 범위	구체적인 내용
			▪ 한국건설기술연구원과 데이터분석 전문회사 지쓰리웨이 소속 전문가들로 자문단을 구성하여 본 연구의 검색 알고리즘 및 검색 결과 요약 시스템 중 교량 유지관리와 관련된 부분을 중점적으로 타당성과 신뢰성을 검토하기 위한 인터뷰를 실시하였음. 연구 결과는 충분히 타당하며, 데이터 전처리를 더욱 엄밀하게 수행하고 사용자 인터페이스의 편의성을 증진시킨다면 교량 유지관리를 위한 점검과정에서 충분히 활용될 수 있을 것이라는 평가를 받음. ○ 평가 결과 기반 시스템 보완 ▪ 자문 결과 교량 점검보고서에서 추출할 수 있는 위험요인으로 손상 발생위치(부재), 손상 종류, 영향인자 등의 정보가 도출되었음. ▪ 본 연구는 자문 결과를 바탕으로 검색 결과 요약 기술을 확장시켜 교량 점검보고서의 검색 결과에 대해 주요 발생 손상 및 영향인자, 손상 발생 위치 등을 요약함.

1. 목표 달성도

○ 핵심성과 1: 딥러닝 기법을 활용한 의미기반 검색 알고리즘

- **목표 성과:** 건설 프로젝트 전주기에 걸친 위험요인 관리 및 대응에 필요한 지식정보를 효과적으로 재활용할 수 있도록 시스템의 검색 기능을 지원하는 알고리즘 개발. 또한, 사용자가 의도한 검색 결과를 정확하게 제공하는 건설산업 특화 의미기반 검색 알고리즘으로 지식정보 습득에 소요되는 시간과 노력을 절감시킴으로써 업무의 효율성 향상.

- **목표 성능:** 도메인의 영향을 크게 받는 의미기반 검색 알고리즘을 건설산업에서 실용적으로 활용할 수 있는 수준의 알고리즘. 건설산업에서의 위험요인 관련 검색에 대하여 건설안전정보시스템과 같은 현행 검색 엔진보다 높은 정확도(Precision, Recall) 결과를 보이는 수준의 알고리즘

- **목표 달성 확인 방법 및 결과:** 의미기반 검색 알고리즘을 통해서 추출된 정보에 대한 정확도(Precision, Recall)를 평가한 결과, precision 93.25%와 recall 80.00%라는 준수한 성능을 보임. 또한 전문가들의 검증을 통해 알고리즘의 정확성과 적절성을 평가했으며, 16명의 참여자들이 평균 74%의 만족도를 보임. 본 핵심성과와 관련해 1개의 국내 특허와 1개의 국제전문학술지 논문의 실적을 달성함.

○ 핵심성과 2: 맞춤형 검색 결과 요약 및 지식정보 생산 시스템

- **목표 성과:** 건설 프로젝트의 위험요인을 효율적·효과적으로 관리하고 선제적으로 대응할 수 있도록 사용자의 검색 결과를 자동으로 요약하는 맞춤형 지식정보 생산 시스템 개발. 사용자가 건설 위험요인과 관련된 지식정보를 수집, 분석 및 평가하는데 소요되는 시간과 노력을 절감시킴으로써 리스크 대응 업무의 효율성 향상

- **목표 성능:** 사용자의 특성 및 요구조건에 맞게 검색 결과를 요약·비교·분석하여 제공. 지식정보의 활용성을 제고하기 위해 건설 위험요인에 관한 검색 결과를 요약하는 과정에서 건설산업의 특성을 반영

- **목표 달성 확인 방법 및 결과:** 검색 결과 요약문의 실효성을 검증하기 위해 K연구원의 전문가 2명에게 자문을 수행하였고, 결과물의 현재 수준으로도 현장에서 충분히 활용 가능하지만 텍스트 데이터의 전처리 과정에서 자동화 수준을 제고할 수 있다면 활용 가능성이 더욱 높아질 것이라는 평가를 받음. 이 외에 정량적인 결과로는 소프트웨어 등록 1건, 국제전문학술지 논문 2건 (게재 1건, 게재 승인 1건), 학술대회 발표 6건 등의 실적을 달성함.

2. 관련 분야 기여도

○ 기술적 측면

- 첨단 기술인 빅데이터 및 텍스트마이닝, 딥러닝 기술을 건설산업에 특화하여 개발하며, 궁극적으로는 타 산업분야에도 적용할 수 있는 지식정보를 자동으로 생산하는 시스템을 개발하는 연구로 이와 관련한 원천기술 확보가 가능함.
- 특히 본 연구는 건설산업에 산재하여 사장되어있는 비정형 텍스트데이터를 활용하여 새로운 건설 지식정보를 자동으로 생산하는 시도로 건설산업의 새로운 가치를 창출하는 원천기술을 확보할 수 있을 것으로 판단됨.
- 또한 본 연구의 기술은 저장된 데이터가 증가할수록 건설 프로젝트의 복잡하고 다양한 속성을 표현할 수 있는 분석 도구로써 뿐만 아니라 향후 건설산업의 다양한 빅데이터 연구에 있어 기본이 되는 지식정보를 생산하는 기술로써 활용될 수 있을 것임.
- 본 연구는 첨단 기술과 건설산업에 오랜 기간 동안 축적되어 온 지식정보의 융합으로 제안하는 시스템 구축을 통해 위험요인 관련 건설 지식정보의 재활용과 지식정보의 선순환 구조를 유도할 것임.
- 더불어 기존에 산재한 방대한 양의 전주기 위험요인 관련 데이터를 자동 입력하여 사용함으로써 상당한 정보가 입력되어야만 사용자가 유의미한 결과를 확보할 수 있었던 사업관리시스템의 현실적 한계를 극복하는 데 이바지 할 것임.
- 본 연구의 결과물은 국가기술지도의 기반주력산업 가치 창출에서의 건설 정보화 기술과 첨단 인프라 건설 기술에 속하여 높은 수준의 건설비용 절감과 더불어 기존 재래식 건설 시공 프로세스를 혁신하고 대외 경쟁력을 확보할 수 있음.

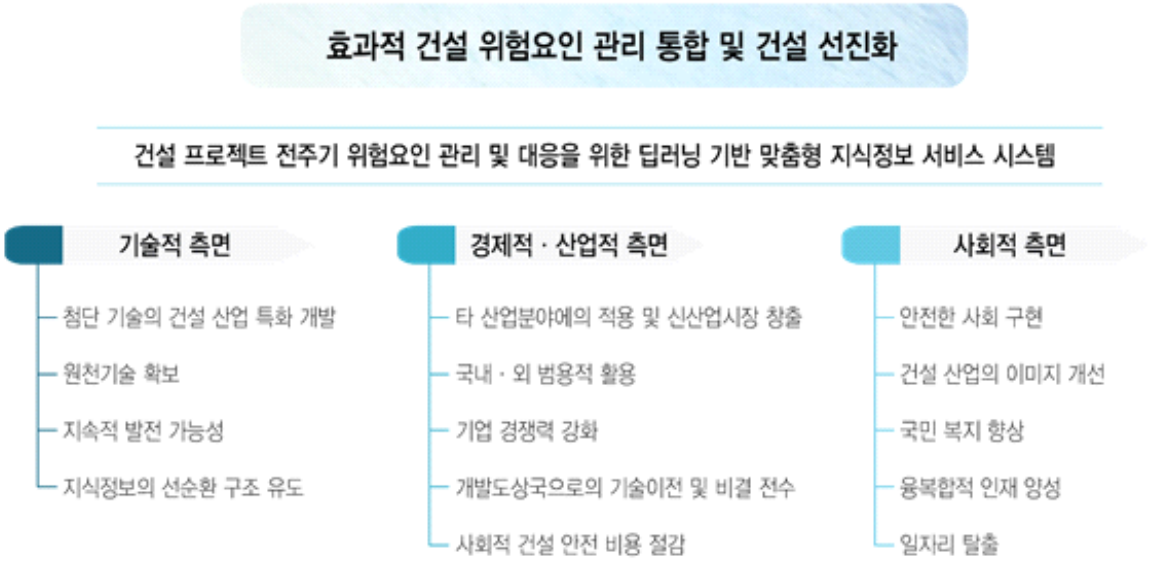
○ 경제적・산업적 측면

- 시스템을 구성하는 핵심 기능인 의미기반 검색과 복잡하고 다양한 형태의 비정형 데이터를 처리하여 사용자 맞춤형 지식정보를 자동으로 생산하는 기술을 타 산업분야에 적용함으로써 신산업시장 창출이 가능함.
- 또한, 비정형 텍스트데이터를 적극적으로 활용한 사례는 매우 제한적이기 때문에 본 연구의 범용적 시스템은 국내뿐만 아니라 해외에서도 유용하게 활용될 수 있을 것으로 판단되며, 이는 건설산업에서 새로운 건설 지식정보를 생산하여 부가가치를 창출함.
- 건설 프로젝트 전주기 위험요인 관리 및 대응을 위한 지식정보 서비스 시스템을 통해 기획 단계의 해외건설 위험요인을 효율적이고 효과적으로 관리함으로써 국내 건설 기업의 경쟁력을 강화하여 적극적인 해외건설 시장 진출이 기대되며, 체계적인 지식정보의 활용 및 재사용을 가능하도록 하여 실제 사업의 전략적 계획 수립 및 수익성 확보에 도움이 될 것으로 판단됨.
- 더불어 건설 프로젝트 수행을 위한 인프라가 열악하고 위험요인 관리 및 대응에 대한 지식정보가 상대적으로 부족한 개발도상국으로의 기술이전 및 비결 전수가 가능할 것임.

- 적시·적재·적소에 건설 프로젝트의 전주기 위험요인 관리 및 대응이 가능하게 하는 연구 결과물은 건설 현장의 안전 비용뿐만 아니라 전주기에 걸친 시설물의 유지관리 비용을 감소시켜 주어 천문학적인 사회적 건설 안전 비용을 효과적으로 감소시킬 것임.

○ 사회적 측면

- 연구의 최종성과물은 건설 프로젝트의 전주기 위험요인을 관리하고 대응할 수 있도록 함으로써 안전한 건설 현장뿐만 아니라 재난·재해에 대한 시설물 안전 진단 및 관리를 지원함으로써 안전한 사회 구현에 기여함.
- 건설산업의 디지털화가 보여줄 수 있는 첨단성을 본 지식정보 서비스 시스템을 통하여 인터넷상의 공간에서 국민에게 보여줌으로써 건설산업에 대한 국민들의 이미지 개선에 기여할 뿐만 아니라, 건설 기술과 문화에 대한 중요성을 국민들에게 각인시킬 수 있음.
- 이는 장기적인 관점에서 건설산업에 대한 국민적 반감을 줄이고 “건설이 곧 복지”라는 인식을 넓혀 SOC 투자 확대를 통한 국민 복지 향상에 기여할 수 있음.
- 또한, 본 연구에서 제안하는 빅데이터 활용에 대한 연구 방법론은 건설산업 빅데이터 분석의 전문지식을 갖춘 융·복합적 인재를 양성하는데 이바지할 것임.
- 본 연구는 효율적이고 효과적인 건설 프로젝트의 위험요인 관리 및 대응 서비스 제공을 통해 기업의 경쟁력을 강화해주는데 이바지할 것으로 예상되며, 이는 해외 진출 확대로 이어져 일자리를 추가로 창출하는 간접적인 기대효과도 있을 것으로 전망됨.



<그림 4-1. 연구 기대성과 및 파급효과>

○ 활용분야

- 본 연구의 결과물인 맞춤형 지식정보 서비스 시스템은 정부부처, 지방자치단체, 민간 기업 등 다양한 주체가 건설 프로젝트의 위험요인을 관리 및 대응하는 과정에서 활용할 수 있음.

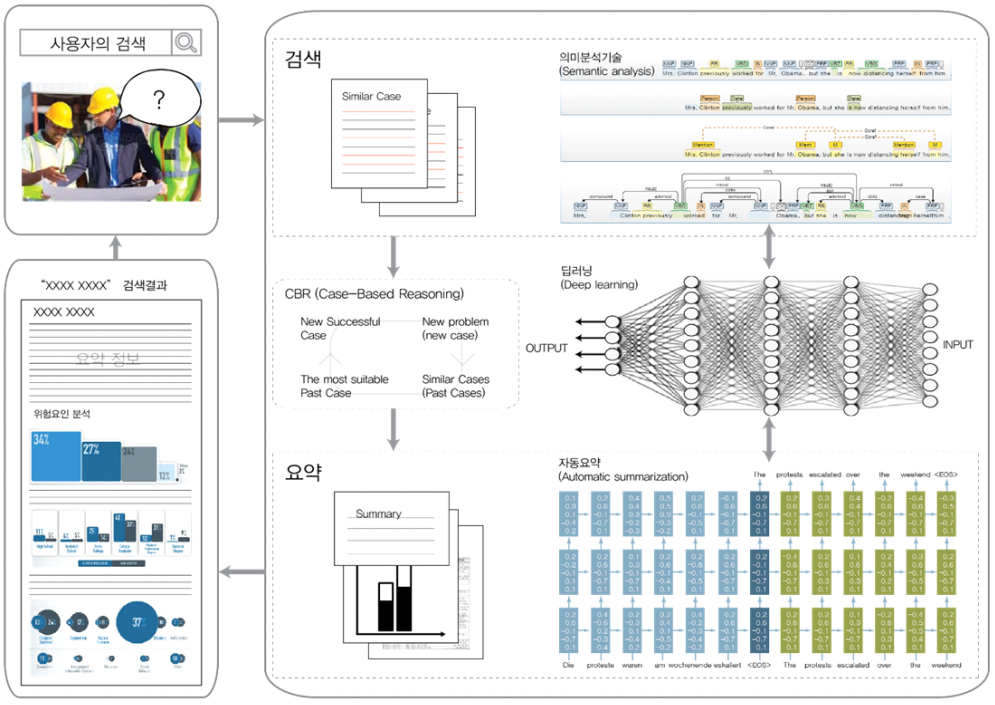
  ▷ 정부부처: 국토교통부 산하 5개, 산업통상자원부 산하 5개, 고용노동부 산하 1개
  - 지방자치단체: 광역자치단체 17개, 기초자치단체 226개
  - 민간 기업: 종합 건설산업체 11,220개, 전문 건설산업체 37,899개

- 궁극적으로 본 연구는 건설 프로젝트의 유형에 한정되지 않으며 전주기에 걸친 위험요인에 대하여 관리 및 대응할 수 있는 개방적이고 범용적인 시스템 개발을 목적으로 하여 실제 연구개발성과의 활용분야는 더욱 다양할 것으로 기대됨.

○ 활용방안

- 본 연구는 건설산업에 산재하여 사장되어 있는 비정형 텍스트데이터를 활용하여 새로운 건설 지식정보를 자동으로 생산하는 시도로 건설산업의 새로운 가치를 창출하는 원천기술을 확보할 수 있음. 이를 통해 효과적 건설 프로젝트 관리를 위한 각종 문서 검색, 비교분석, 요약, 시각화 등을 가능하도록 하여 지식정보의 재활용 및 지식의 선순환 구조 확립을 유도할 것임.

- 또한, 검색 결과를 자동으로 요약하고, 검색된 문서의 내용을 자동으로 비교·분석하여 제공하는 맞춤형 지식정보 생산 시스템으로 해외건설 위험요인관리, 건설현장 안전관리, 시설물 유지관리 등에 필요한 지식정보에 대한 사용자 접근성을 높임.



<그림 5-1. 연구개발 결과물의 현장 활용 개념도>

▪ 건설 프로젝트의 대표적인 단계별 위험요인인 기획 단계의 해외건설 위험요인, 시공 단계의 건설현장 안전 위험요인, 유지관리 단계의 시설물 유지관리 위험요인 각각에 대한 구체적 활용방안은 아래와 같음.

- **해외건설 위험요인 관리**

  ▷ 해외 건설 프로젝트 전주기에 걸친 주요 위험요인 파악을 통한 위험요인 집중 관리 및 선제적 대응에 활용됨.

  ▷ 해외 건설 프로젝트 투입 인력의 해외 건설 사전 교육용으로 활용됨.

  ▷ 해외 건설 프로젝트의 다양한 속성과 사례에 기반한 해외 진출 전략 수립에 활용됨.

- **건설현장 안전 관리**

  ▷ 건설 프로젝트의 특정 공법 및 공종에 대한 위험요인 사전 검토로 건설현장의 안전사고 발생을 미연에 방지함.

  ▷ 건설현장 안전사고 데이터베이스 공유를 통한 건설 안전사고 예방체계 확립에 활용됨.

  ▷ 건설현장 안전사고 데이터베이스에 기반한 건설 안전 교육 프로그램 개발을 통해 건설산업의 안전 기술력을 제고함.

- **시설물 유지관리**

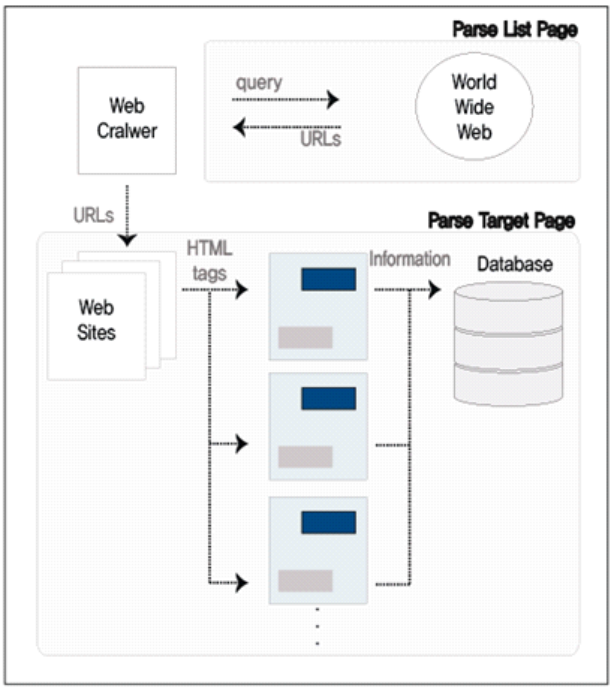
  ▷ 시설물 유지관리 시 발생하는 주요 위험요인 파악을 통하여 최적의 시설물 유지관리를 위한 의사결정 지원에 활용됨.

  ▷ 시설물 관리주체의 중장기 유지관리 전략수립에 대한 의사결정 지원 도구로써 활용됨.

  ▷ 시설물 유지관리 관련 정책수립에 대한 의사결정 지원 도구로써 활용됨.

○ 웹 크롤링(Web Crawling)

- 웹크롤링(Web Crawling)이란 일정한 구조의 웹사이트로부터 필요한 정보를 자동적으로 추출해내는 행위를 뜻 함(Cho 2002). 웹은 기본적으로 문서의 글자크기, 글자색, 글자모양, 그래픽, 하이퍼링크 등을 정의하는 명령어인Hypertext Markup Language(HTML) 형태로 구축되어 있음. 각 정보에는 <size>, <font>, <text> 등 태그가 부착되어 있기 때문에, 웹페이지의 구조를 파악할 수 있다면 같은 구조를 가지는 수많은 웹페이지로부터 특정 정보를 추출할 수 있음.



<그림 6-1. 웹크롤링 개요>

- 웹크롤링은 (1)웹사이트들의 Uniform Resource Locator(URL) 리스트를 받아오는 과정(Parse List Page)과 (2)각 URL에 접근하여 정보를 추출하는 과정(Parse Target Page)으로 구분됨(Manning et al. 2008). Parse list page 단계에서는 사용자가 질의어(query)를 입력하면 웹페이지에는 검색 결과의 리스트가 나타나게 되고, HTML 태그 중 하이퍼링크 태그를 이용하여 각 검색 결과의 URL을 추출함. Parse target page 단계에서는 앞서 추출한 URL에 각각 접근하여 필요한 정보를 추출하는 과정으로, 본 연구에서는 네이버 뉴스 기사의 댓글 및 트위터의 트윗 본문의 텍스트 데이터를 추출함.

○ 텍스트마이닝(Text Mining)

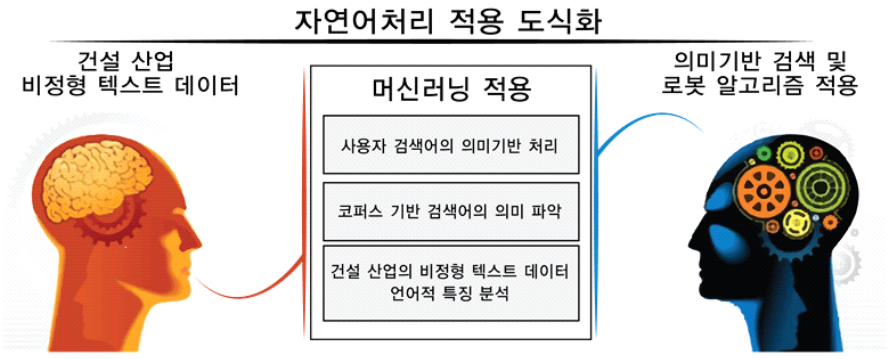
- 텍스트마이닝은 비정형 데이터에 대하여 자연어처리(Natural Language Processing) 기술과 문서 처리 기술을 적용하여 유용한 정보를 추출하여 제공하는 빅데이터 분석의 주요 방법론 중 하나로, 데이터 수집, 정제, 변환의 과정 다음 전처리 과정을 거쳐 정보를 추출하고, 문서 처리, 색인 작성, 문서 요약 등의 많은 응용 분야에 활용됨.

- 본 연구는 비정형 텍스트데이터를 활용하는 시스템 개발을 최종 목표로 진행되며, 따라서 텍스트데이터를 분석할 수 있는 텍스트마이닝 기술은 의미기반 검색기능을 구현하고 맞춤형 지식정보를 서비스함에 있어 필수적임.

○ 자연어처리(Natural Language Processing)

- 텍스트마이닝 기법을 적용하기 위해서는 가장 먼저 비정형 텍스트데이터를 컴퓨터가 이해할 수 있도록 처리하는 과정이 필요하며, 자연어처리는 이 과정에서 활용되는 기술로써, 사람의 언어를 컴퓨터가 이해하여 자연어를 분석 및 생성할 수 있도록 하는 기술임. 사람이 사용하는 언어를 형태 분석, 의미 분석, 대화 분석 등을 통해 컴퓨터가 인식해서 처리할 수 있도록 바꾸어주는 작업을 의미하며, 처리한 결과물은 사람의 편의에 따라 텍스트, 음성, 그래픽 등으로 생성됨. 그러나 자연어의 특성 상 언어별, 산업 분야별 사용되는 자연어의 특징이 각기 다르기 때문에 범용적인 자연어 처리 기법이 존재하지 않아 해당 언어 및 산업 분야에 적합한 자연어처리 기법을 적용하는 것이 매우 중요함.

- 본 연구에서는 건설산업에서 발생하는 비정형 텍스트데이터의 언어적 특징을 분석하고, 건설산업의 문서 이해에 특화된 알고리즘을 개발하는 기반이 됨. 이 과정에서 수집된 데이터로부터 효율적이고 정확한 자연어 처리를 할 수 있도록 머신러닝 기법을 적용시킴.



<그림 6-2. 건설 산업 텍스트 데이터 분석을 위한 자연어처리 적용>

○ 자동 요약(Automatic Summarization)

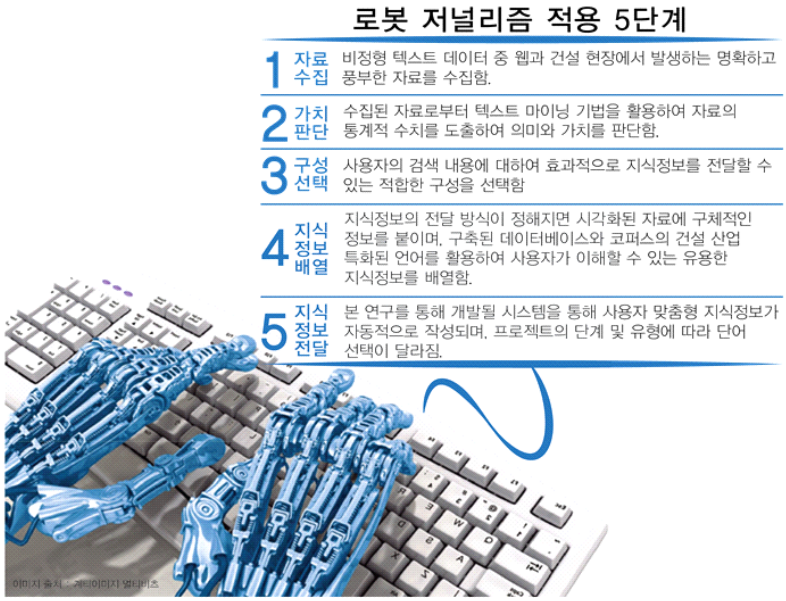
- 텍스트데이터를 자동으로 요약하는 방법은 크게 2가지, Extraction 방식과 Abstraction 방식으로 구분됨. Extraction 방식은 원래의 문서에서 일부 텍스트를 추출하고, 추출한 텍스트를 조합하여 요약하는 기법임. 계산 속도가 빠르고 주요 키워드가 잘 보존된다는 장점이 있지만, 요약된 문장의 가독성이 떨어지며 표 또는 그림과 같이 텍스트가 아닌 형태로 존재하는 정보에 대해서는 요약할 수 없다는 단점이 있음. Abstraction 방식은 원래의 문서에서 키워드를 찾고, 일반적인 용어들을 사용하여 키워드를 연결함으로써 매끄러운 문장을 만들어내는 방식임. Extraction 방식에 비해 그럴 듯한 문장이 생성되지만, 프로그램 구현 및 결과물 검증이 어렵다는 단점이 존재함 (Habn and Mani 2000).

- 본 연구에서는 해외건설시장과 관련된 뉴스기사를 데이터로 사용하고 있는데, 결과물의 가독성보다는 정보의 보존이 중요하다고 판단되며 뉴스기사의 특성 상 표/그림의 정보 또한 텍스트로 설명되기 때문에 Extraction 방식을 적용함.

○ 로봇저널리즘(Robot Journalism)

- 로봇저널리즘은 텍스트마이닝 기술을 활용하여 개발한 소프트웨어를 통해 자동으로 작성되는 기사를 의미하며, 이렇게 작성되는 기사는 웹상의 데이터 수집 및 정리, 알고리즘에 기반한 분류, 의미해석의 단계를 거침. 로봇저널리즘의 기사는 일반적으로 5단계를 통해 작성되며, 각 단계는 '1. 자료 수집, 2. 가치 판단, 3. 구성 선택, 4. 기사 배열, 5. 기사 작성'임 (컬럼비아 저널리즘스쿨 토우 센터).

- 본 연구에서 로봇저널리즘 기법은 사용자 검색의 결과로 추출된 문서 내용을 자동으로 비교·분석하여 그래프, 도표 등의 시각화 정보로 요약하는 기술과 함께 이를 해석하고 요약하는 등의 지식정보를 생산하는 과정에서 사용됨. 본 연구의 로봇저널리즘 적용은 다음 그림과 같은 단계로 이루어짐.

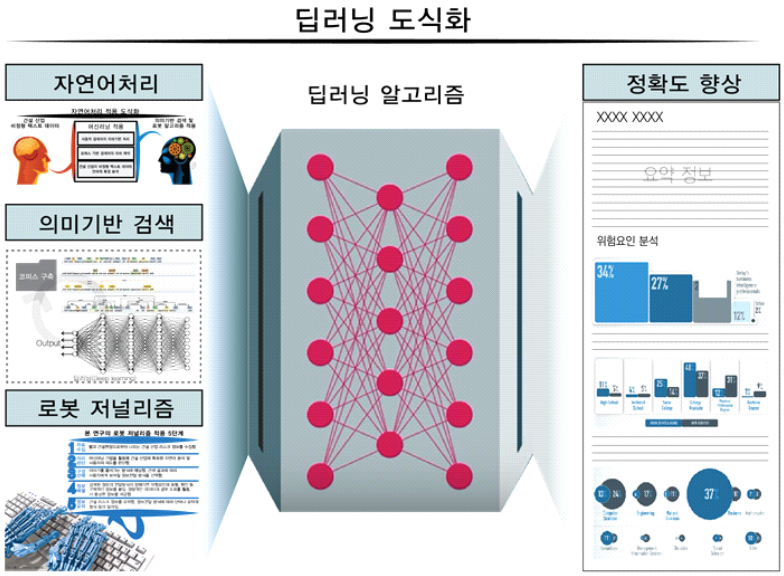


<그림 6-3. 로봇저널리즘 적용방안 5단계>

○ 딥러닝(Deep Learning)

- 딥러닝은 머신러닝의 한 분야로 다층구조 형태의 신경망 알고리즘을 기반으로 높은 수준의 추상화 모델을 구축하는 기법이며, 대표적인 딥러닝 기법을 활용하여 개발한 인공지능 프로그램으로는 알파고(AlphaGo)가 있음.

- 본 연구는 다양한 머신러닝 기법 중에서도 딥러닝을 활용하여 인공지능형 건설 프로젝트 전주기 위험요인 관리 및 대응을 위한 딥러닝 기반 맞춤형 지식정보 서비스 시스템을 구현하고자 함. 특히 본 연구에서는 딥러닝 기법을 활용하여 사용자 검색어의 의미를 파악하며, 맞춤형 검색 결과 요약 및 지식정보를 생산함.



<그림 6-4. 딥러닝 연구 적용방안>

제7장. 연구개발성과의 보안등급 (해당사항 없음)

제8장. 국가과학기술종합정보시스템에 등록한 연구시설 · 장비 현황 (해당사항 없음)

구입 기관	연구시설/ 연구장비명	규격 (모델명)	수량	구입 연월일	구입 가격 (천원)	구입처 (전화번호)	비고 (설치 장소)	NTIS장비 등록 번호

제9장. 연구개발과제 수행에 따른 연구실 등의 안전 조치 이행 실적

○ 본 연구실은 서울대학교 환경안전원의 연구실안전환경관리규정에 의거하여 연구실 안전을 확보함.

○ 본 연구는 컴퓨터를 주요 연구 장비로 사용하기 때문에 기술적 위험 요소는 존재하지 않음.

○ 연구실 안전 관리 이행 사항

- 연구실 정기점검
  - 서울대학교 연구실안전환경관리규정에 의거하여 연구실 정기점검 및 정밀안전진단을 실시함.
- 연구실안전환경관리규정 연구실 비치 및 준수
  - 연구실에 서울대학교 연구실안전환경관리규정을 비치하고 과제참여자에게 규정 숙지 및 교육을 정기적으로 실시함.
  - 안전 관리 담당자를 지정하여 매달 안전점검을 실시하며, 서울대학교 환경안전원에서 정기적으로 실시하는 각종 안전점검에 성실하게 협조하여 안전 확보를 도모함.
- 건설현장 방문 시 기본적 안전장비 착용함을 원칙으로 하며, 이를 대비하여 연구실에 안전모 및 안전화를 구비함.
- 건설현장 방문 시 현장 안전에 대한 훈련에 반드시 응하고, 훈련내용을 숙지하도록 교육함.

제10장. 연구개발과제의 대표적 연구 실적

번호	구분	논문명/특허명/기타	소속 기관명	역할	논문 게재지/ 특허 등록 국가	영향력 지수	논문 게재일 /특허 등록일	사사 여부	특기 사항
1	논문	Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks	서울대학교	주관	Automation in Construction	2.919	2018.03.01	중복사사	SCI-E
2	논문	Predicting Construction Cost Index Using the Autoregressive Fractional Integrated Moving Average (ARFIMA) Model	서울대학교	주관	Journal of Management in Engineering	2.011	2018.03.01	단독사사	SCI-E
3	논문	Document Management System Using Text Mining for Information Acquisition of International Construction	서울대학교	주관	KSCE Journal of Civil Engineering	0.812	2018.04.29 (accepted)	단독사사	SCI-E
4	특허	건설 프로젝트의 전주기 위험요인 지식정보 서비스 시스템	서울대학교	주관	대한민국	-	2017.04.28 (출원)	-	-
5	S/W	의미기반 건설 안전사고 위험요소 검색 시스템	서울대학교	주관	대한민국	-	2018.05.15	-	-

제11장. 기타 사항 (해당사항 없음)

제12장. 참고 문헌

- 건설산업연구원 (2013) 2014년 해외건설 잠재리스크 최소화를 위한 긴급토론회. 한국건설산업연구원.
- 국토교통부 (2014) 2014년 건설사업정보화 시행계획 추진(기술정책과). 국토교통뉴스, http://www.molit.go. kr/USR/NEWS/m_71/dtl.jsp?id=95073562.
- 유위성 (2014) 해외사업 리스크 관리에 빅 데이터 기술 활용할 수 없나. CERIK 저널, 특집2, 한국건설산업 연구원
- 유위성, 김우영 (2015) 해외건설사업 리스크지수(IPRI) 개발 및 시사점. 건설산업연구원.
- 이종세, 황인호, and 이후석 (2007) "교량유지관리를 위한 첨단굴절로봇차.". 제 33 회 대한토목학회 정기 학술대회 CIVIL EXPO 2007 (2007): 3120-3123.
- Egbu, C.O. and Botterill, C. (2002) “Information technologies for knowledge management: their usage and effectiveness.” ITcon 7(Special Issue ICT for Knowledge Management in Construction), 7(1).
- Hyari, K.H., Tarawneh, Z.S., and Katkhuda, H.N. (2016) “Detection Model for Unbalanced Pricing in Construction Projects: A Risk-Based Approach.” Journal of Construction Engineering and Management, 142(12).
- Macit İlal, S. and Günaydın, H.M. (2017) “Computer representation of building codes for automated compliance checking.” Automation in Construction, 82(1).
- Javernick-Will, A.N. and Scott, W.R. (2010) “Who Needs to Know What?” Institutional Knowledge and International Projects: Journal of Construction Engineering and Management, 136(5).
- Kim, J. and Chi, S. (2017) “Adaptive Detector and Tracker on Construction Sites Using Functional Integration and Online Learning.” Journal of Computing in Civil Engineering, 31(5).
- Salama, D. and El-Gohary, N.M. (2016) “Semantic Text Classification for Supporting Automated Compliance Checking in Construction.” Journal of Computing in Civil Engineering, 30(1).
- Taroun, A. (2014) “Towards a better modelling and assessment of construction risk: Insights from a literature review.” International Journal of Project Management, 32.

9. 뒷면지

주 의

1. 이 최종보고서는 국토교통부에서 시행한 국토교통기술촉진연구사업의 연구보고서입니다.

2. 이 최종보고서 내용을 발표하는 때에는 반드시 국토교통부에서 시행한 사업의 연구개발성과임을 밝혀야 합니다.

3. 국가과학기술 기밀 유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 안 됩니다.