KJCEM 17. 5, 089~096
SEPTEMBER 30, 2016

www.kicem.or.kr (pISSN 2005-6095 / eISSN 2465-9703)
http://dx.doi.org/10.6106/KJCEM.2016.17.5.089

텍스트 마이닝을 통한 해외건설공사 입찰정보 분석
- 해외건설공사의 입찰자 질의(Bidder Inquiry) 정보를 대상으로 -

이지희[1] · 이준성* · 손정욱[1]
[1]이화여자대학교 건축공학과

Construction Bid Data Analysis for Overseas Projects Based on Text Mining
- Focusing on Overseas Construction Project's Bidder Inquiry

Lee, JeeHee[1], Yi, June-Seong*, Son, JeongWook[1]
[1]Department of Architectural Engineering, Ewha Womans University

Abstract : Most data generated in construction projects is unstructured text data. Unstructured data analysis is very needed in order for effective analysis on large amounts of text-based documents, such as contracts, specifications, and RFI. This study analysed previously performed project's bid related documents (bidder inquiry) in overseas construction projects; as a results of the analysis frequent words in documents, association rules among the words, and various document topics were derived. This study suggests effective text analysis approach for massive documents with short time using text mining technique, and this approach is expected to extend the unstructured text data analysis in construction industry.

Keywords : Text Mining, Unstructured Text Data, Bidder Inquiry

1. 서론

1.1 연구의 배경 및 목적

기업에서 생산되는 데이터의 80% 이상은 비정형 데이터로 이루어져 있으며, 그 중에서도 텍스트 데이터의 비율은 매우 높다. 이는 건설프로젝트의 경우도 마찬가지인데, 건설공사를 수행하면서 활용하는 대다수의 정보는 계약서, 설계변경 보고서, RFI (Request for Information) 등과 같이 텍스트 기반의 비정형 데이터 형태로 이루어져있다. 그 중에서도 입찰단계에서 발주자가 제공하는 입찰문서의 경우는 계약서, 설계도면, 시방서 등과 같이 발주자의 설계의도를 파악하고, 프로젝트 특성을 판단할 수 있는 중요한 정보라 할 수 있다.

입찰문서에는 목적물을 완성하기 위한 설계와 시공방법, 자재, 품질검사에 대한 기준과 절차가 기술되어 있다. 뿐만 아니라 입찰문서 자체가 계약적으로 효력을 갖기 때문에 추후 클레임 및 분쟁이 발생하였을 경우 중요한 판단의 근거자료로 활용되기도 한다. 입찰단계에서는 입찰안내서(ITB, Invitation to Bid)의 요구사항과 프로젝트 관련 정보 및 아이템을 검토하고 공사비에 반영하게 된다(Seo et al., 2016). 따라서 입찰문서 내에 포함되어 있는 리스크 요인들을 사전에 검토하지 못할 경우 추후 공사 수행 과정에서 공사비 증가요인이 될 수 있으며, 때문에 리스크 관리 차원에서도 입찰문서에 대한 분석은 매우 중요하다고 할 수 있다.

특히 최근 발주되는 해외건설공사에서는 입찰참여자(시공사)가 입찰문서 내의 오류 및 누락사항, 불일치 정보 등에 대해 사전에 충분히 검토하지 못해서 발생한 문제의 경우에는 설계변경이 불가능하도록 정하고 있기 때문에 입찰문서에 대한 철저한 검토가 요구되고 있는 상황이다. 그러나 짧은 입찰기간 동안 방대한 양의 입찰문서를 검토하는 것은 쉽지 않은 상황이며, 이러한 작업을 보다 효과적으로 지원할 수 있는 자동화된 기술의 도움이 필요하다.

이에 본 연구에서는 방대한 양의 문서를 단시간 안에 효과적으로 분석할 수 있는 텍스트 마이닝(text mining)을 비롯한 비정형 텍스트 데이터 분석 방법 기술을 활용하여 해외건설

* Corresponding author: Yi, June-Seong, Department of Architectural Engineering, Ewha Womans University, Seoul 120-750, Korea
E-mail: jsyi@ewha.ac.kr
Received July 29, 2016: revised -
accepted August 3, 2016

공사의 입찰정보를 분석하고, 시사성 있는 정보를 도출하여 향후 활용방안을 모색하고자 한다. 이를 위해 입찰문서 내 불확실한 정보 및 누락사항 등을 사전에 검토하여 입찰 전 발주자에게 질의하는 '입찰자 질의(bidder inquiry)' 정보를 대상으로 텍스트 분석을 실시함으로써 어떤 유형의 리스크 요인이 입찰문서에 존재하고, 입찰문서에서 어떤 부분을 사전에 검토하여 발주자에게 질의하여야 하는지에 대한 전반적인 이해를 돕고자 한다.

1.2 연구의 방법 및 절차

본 연구에서는 텍스트 기반의 해외건설공사 입찰문서 내 정보를 효과적으로 분석하기 위해 텍스트 마이닝, 정보 검색(Information Retrieval, IR), 자연어 처리(Natural Language Processing, NLP) 방법 등과 같은 비정형 데이터 분석 방법을 활용하였다. 또한 통계분석용 오픈 소스 소프트웨어인 R 프로그래밍을 통하여 비정형 텍스트 문서를 구조화하고, 분석 및 시각화하는 작업을 실시하였다. R은 패키지뿐만 아니라 일종의 프로그래밍 언어로서 기본적인 통계 기법부터 모델링, 데이터 마이닝 기법까지 구현이 가능하며, 구현한 결과는 그래프 등으로 시각화할 수 있다. 또한 Java나 C, Python 등 다른 프로그래밍 언어와 연결이 용이하여 프로젝트 특성에 맞는 독창적인 통계기법의 사용이 가능하다(Yim, 2015)는 장점이 있어 본 연구의 텍스트 마이닝을 위한 분석 툴로 선정하였다.

본 연구의 수행절차는 다음과 같다.

첫째, 텍스트 마이닝 및 비정형 데이터 분석 방법에 대한 이론적 고찰을 바탕으로 해외건설공사 입찰정보 분석을 위한 방향을 수립한다. 둘째, 텍스트 마이닝 분석을 위한 입찰 질의 문서를 수집하되, 유의미한 분석 결과를 제시할 수 있도록 여러 프로젝트의 충분한 양의 데이터를 확보한다. 셋째, 텍스트 데이터를 구조화된 형태로 전환시키기 위해 텍스트 데이터 전처리(pre-processing) 과정을 거친다. 넷째, 입찰 질의서에서 반복적으로 발생하는 빈출 용어 및 주요 토픽에 대한 정보를 토대로 입찰자 질의 문서에서 공통적으로 지적되는 입찰문서의 주요 문제요인들을 파악한다. 다섯째, 연구의 결론을 도출한 후 향후 활용방안에 대해 제시한다.

2. 예비적 고찰

2.1 비정형 텍스트 데이터 분석

최근 빅 데이터 기술이 각광을 받으면서 방대한 양의 데이터를 다루는 기술뿐만 아니라 텍스트, 이미지, 음성 데이터와 같이 정형화 되지 않은 비정형 데이터를 다루는 기술이 빠르게 발전하고 있다(Lee et al., 2016). 건설 활동에서 발생하는 대다수의 정보들도 비정형 데이터의 형태를 띠고 있다고 할 수 있는데, Simoff and Maher (1998)는 건설 정보를 관리한다는 것은 곧 다양한 형태의 데이터 유형과 관련이 있음을 지적하며 건설 데이터의 유형을 다음과 같이 분류하였다.

- 정형 데이터 파일(structured data files) : 데이터 베이스(database)나 데이터 웨어하우스(data warehouse), 또는 ERP(Enterprise Resource Planning) 등과 같은 특정 어플리케이션에 저장된 구조화된 파일
- 준정형 데이터 파일(semi-structured data files) : HTML, XML, SGML 등의 파일
- 비정형 텍스트 데이터 파일(unstructured text data files) : 계약서, 시방서, 재료 카탈로그, 설계변경보고서, RFI, 회의록 등의 파일
- 비정형 그래픽 파일(unstructured graphic files) : 2D, 3D 설계도면
- 비정형 멀티미디어 파일(unstructured multimedia files) : 사진, 이미지, 오디오, 비디오 파일

일반적으로 텍스트 데이터를 분석하는 방법으로는 텍스트 마이닝이 많이 활용되고 있는데, 텍스트 마이닝은 자연어로 구성된 비정형 데이터에서 패턴 또는 관계를 추출하여 의미있는 정보를 찾아내는 기법으로, 컴퓨터가 사람들이 말하는 언어를 이해할 수 있도록 하는 자연어 처리(NLP)에 기반을 둔 기술이다(Yim, 2015).

건설 분야에서는 건설 문서의 자동 분류를 위해 텍스트 마이닝 기술을 활용한 연구들이 북미권을 중심으로 일부 수행되어 왔다. Caldas et al. (2002)의 연구에서는 PMIS와 같은 건설 프로젝트 정보 시스템에 저장된 수많은 건설문서들을 자동으로 분류하기 위해 텍스트 마이닝 기반의 기계 학습(machine learning)을 통해 건설 문서 자동분류 시스템의 프로로타입을 개발하였다. 또한 문서 내의 컨텐츠와 관련이 있는 정보를 상호 연결하기 위해 텍스트 마이닝을 활용한 연구들도 수행되었는데 Mao et al. (2007)의 연구에서는 문서 내에 작성된 개별 컨텐츠와 관련된 정보를 연결하기 위해 비정형 데이터의 메타데이터 모델(metadata model)을 사용하기도 하였다. 이처럼 건설 분야의 비정형 텍스트 데이터 분석을 위한 연구는 다양한 측면에서 이루어지고 있으며, 비정형 데이터와 정형 데이터를 통합하여 정보 관리의 효율을 꾀하기 위한 움직임도 나타나고 있다.

2.2 해외건설공사 입찰정보 분석의 필요성

본 연구에서 분석의 대상으로 선정한 해외건설공사 입찰단계의 입찰자 질의 정보는 서신 또는 입찰 사전 회의(pre-bid meeting)를 통해 이루어지며, 발주자가 제공한 입찰문서(계약서, 도면, 시방서 등)의 내용이 명확하지 않거나, 누락된 정보가 있는 경우 입찰 참여자가 발주자에게 질의하는 것을 말한다. 이때 입찰 참여자는 입찰지시서의 규정에 따라 서면으

로 질의사항을 작성하여 발주자 또는 엔지니어에게 제출하여야 하며, 발주자 및 엔지니어는 질의에 대한 답변을 서면으로 작성하여 모든 입찰자가 정보를 확인할 수 있도록 하여야 한다. 입찰자 질의 프로세스는 단순히 입찰을 위한 형식적인 과정이 아니며, 입찰문서 내 불확실하고 모호한 정보를 명확히 함으로써 입찰자의 견적 정확성을 높이고 추후 발생 가능한 설계변경 및 클레임/분쟁 리스크를 사전에 방지할 수 있는 제도적 장치이기도 하다. 뿐만 아니라 입찰자 질의 자료는 추후에 입찰서에 첨부되어 계약의 일부가 되기 때문에 입찰단계에서 발생하는 중요한 정보 중 하나라 볼 수 있다. 따라서 입찰자 질의서에는 입찰문서에서 발생할 수 있는 다양한 유형의 잠재 리스크 요인들이 포함될 가능성이 높으며, 이러한 문서들을 분석함으로써 입찰문서에서 공통적으로 지적되는 주요 리스크 요인들을 추출할 수 있을 것이라 판단되다.

Tanaka (1988)의 연구에 따르면 미국에서 발생한 건설관련 클레임의 74.4%가 불충분한 입찰문서 정보, 계약조항의 모호함, 입찰문서 내 정보간의 불일치(상충) 등으로 인해 발생한 것으로 나타났다. 이는 입찰문서에 작성된 정보들로 인해 시공과정에서 리스크가 발생할 수 있음을 보여주는 것으로 입찰문서 검토의 중요성을 보여주는 사례이기도 하다. 또한 국내 건설기업의 해외건설공사 입찰단계의 역량 분석을 실시한 Kim et al. (2014)의 연구에서 국내기업들의 입찰준비단계에서 수행하는 프로젝트 리스크 검토 및 계약조건 검토 등의 업무가 중요성 대비 보유 역량의 차이가 상대적으로 큰 것으로 나타나 국내 건설기업의 경쟁력 향상 측면에서도 해외건설공사의 입찰정보 분석은 필요하다고 판단된다.

3. 데이터 수집 및 전처리

3.1 분석 데이터

해외건설공사 입찰 질의정보에 대한 분석을 실시하기 위해 본 연구에서는 미국 캘리포니아 주정부 교통국(California Department of Transportation, Caltrans)에서 최근 3년 내에 발주한 공공 건설프로젝트를 대상으로 입찰 질의서 데이터를 수집하였다. 미국 공공 건설프로젝트를 텍스트 분석의 데이터로 선정한 이유는 국내 기업들이 많이 진출한 아시아나 중동지역의 경우 공공프로젝트의 입찰문서 및 계약사항, 입찰 질의서 등과 같은 상세 정보를 공개하고 있지 않을 뿐만 아니라 미국 캘리포니아 주정부 교통국에서는 매년 다수의 건설 사업을 발주하고 있기 때문에 보다 효과적인 분석이 가능하다고 판단하였기 때문이다.

분석 데이터는 211개의 도로 인프라 공공 프로젝트에서 발생한 총 1,054건의 입찰 질의 문서로서, 텍스트 데이터 전처리 과정을 거쳐 분석을 실시하였다. Table 1은 입찰 질의 문서의 일부로서, 개별 문서들은 시공사의 질의문(inquiry)과 발주자/엔지니어의 답변문(response)로 구성되어 있다.

3.2 비정형 텍스트 데이터 전처리

텍스트 마이닝 분석을 실시하기 위해서는 불필요한 정보를 제거하고, 비정형 데이터를 정형 데이터로 구조화하는 작업이 필요한데 이를 위해 데이터 전처리 과정은 필수적이라 할 수 있다. 데이터 전처리 과정은 텍스트 형태로 작성된 문서를 컴퓨터가 자동으로 인식할 수 있도록 사전 작업을 하는 것으로 텍스트 데이터의 상황에 따라 전처리 방법은 가변적이다.

Table 1. Bidder inquiry (partial)

Proj. code	Inquiry	Response
01-0A1004	Specification 14-6.02, Species Protection, the state indicates that it expects nesting at 3 of the 4 bridges. However, the specification indicates that nesting should be prevented between March 1 and August 31. Since this project is bidding on March 22, 2016 is the state currently taking steps for this provision to prevent nesting prior to award of the contract to the contractor?	No measures are currently being taken to prevent nesting of swallows.
01-262054	Caltrans has identified by Bid Item No. 10 the application of 145,000 square yards of tacked straw, however it is not clear in the plans or specifications in what project phase or where the material may be applied. Will Caltrans please provide this information?	Bid Item No. 10 is "Temporary Tacked Straw". The material is to be applied as necessary in compliance with the Contractor's Storm Water Pollution Prevention Plan.
01-0C3504	Can consideration be given to the max. 7 day time from cold plane to overlay as stated in spec. section 15-2.02B(3)(a). Thereisoverfivemilesofdigoutsandcrackfilling. Wewouldbeforcedtocompleteallworkinsegmentsifthisisrequired.	Your inquiry is under review. Please note that due to the current time frame between the inquiry submittal and the bid opening date, a response may not be provided before the bid opening. If a response is not provided before the bid opening addressing your concern, please bid per the current contract documents. Thank you for your patience.
03-0G3704	: Page 19 of the specifications for Section 39-5.02B(3) states, "Asphalt binder used in HMA for BWC-G must be PG 64-28M." If RHMA is being used, you cannot use PG 64-28PM as a base binder for Asphalt Rubber Binder. Only grades to be used as base binders for ARB, are 64-10/16 or 58-22 or 70-10 depending on climatic region to be used in. Will an addendum be issued to correct the Grade of binder to be used?	Please bid per the current contract documents

3.2.1 불필요한 정보 제거

텍스트 문서들에는 명사, 형용사, 동사 등 다양한 형태의 단어들이 존재하며, 구두점, 기호, 공백 등과 같이 분석을 함에 있어 불필요한 단어들도 존재한다. 따라서 분석의 효율을 높이기 위해서는 이러한 불필요한 정보들을 제거할 필요가 있다.

1,054건의 입찰자 질의 문서에 대한 불필요한 정보 제거의 일환으로서 대문자와 소문자의 구분을 없애기 위해 모든 단어를 소문자로 변환하는 작업을 우선 실시하였다.[1] 그런 후 문서 내 모든 구두점(마침표, 콤마, 세미콜론, 콜론 등)을 제거하고, 관사, 전치사, 조사, 접속사 등 문장에서 내용을 설명함에 있어 큰 비중을 차지하지 않는 단어들을 불용어로 정의하여 제거하였다.[2] 또한 불용어는 아니지만 'inquiry', 'response', 'contractor'등과 같이 입찰자 질의서에 반복적으로 등장하는 단어들은 그 자체가 분석에 있어 특정한 의미를 갖지 못하기 때문에 효과적인 분석 및 작업 속도의 향상을 위해 제거하였다.[3] 그 결과 당초 11,874개의 단어로 구성되었던 문서들이 불필요한 정보 제거 이후 5,196개의 단어로 50% 정도 감소한 것을 확인할 수 있었다(Table 2).

Table 2. Pre-processing of text data

Input data information after preprocessing		
number of documents		1054 documents
number of terms	initial	11,874 terms
	after preprocessing	5,196 terms
minimum word length		3 characters
sparsity		97%

3.2.2 텍스트 데이터 구조화

텍스트 형태의 문서 데이터를 컴퓨터가 자동으로 인식할 수 있는 형태로 만들기 위해서는 텍스트 데이터를 구조화된 정형 데이터로 전환하여야 한다. 이를 위해 텍스트 기반의 문서를 벡터 공간 모델(vector space model)로 나타내게 되는데, 이는 텍스트 문서를 색인어(index term)와 같은 식별자들의 벡터로 나타내는 대수적 모델이라고도 할 수 있다. 분석 대상을 벡터 공간 모델로 나타내기 위해서는 불필요한 정보가 제거된 문서들을 대상으로 term-document matrix를 생성할 수 있다. term-document matrix는 문서들의 집합에서 발생하는 단어들의 빈도를 설명하는 수학적인 행렬구조이다. term-document matrix에서 행은 문서, 열은 단어를 의미하며, 각 행렬의 원소는 특성 단어가 특정 문서에서 발생하는 빈도를 의미한다.

$$\text{term-document matrix} = \begin{pmatrix} & T_1 & T_2 & \dots & T_t \\ D_1 & w_{11} & w_{12} & \dots & w_{1t} \\ D_2 & w_{21} & w_{22} & \dots & w_{2t} \\ \dots & \dots & \dots & \dots & \dots \\ D_n & w_{n1} & w_{n2} & \dots & w_{nt} \end{pmatrix}$$

1,054건의 문서와 5,196개의 단어들을 대상으로 term-document matrix를 생성할 경우 상당한 크기의 행렬이 만들어진다. 그러나 행렬 내의 단어들이 전처리 과정을 거쳤다고는 하지만, 모든 단어들이 동등하게 중요한 의미를 가지는 것이 아니다.[4] 즉, 상대적으로 중요한 단어들에 가중치를 부여하여 우선순위가 높은 단어들을 선별할 필요가 있는데, 이 값을 계산하는 데에는 여러 가지 방법이 있으며 그 중 가장 잘 알려진 방법이 tf-idf (term frequency - inverse document frequency) 가중치 산정방법이다. tf-idf는 각 문서 내 단어의 빈도뿐만 아니라 여러 문서들에서 단어가 발생하는 빈도를 함께 고려한 개념으로, 특정 단어가 모든 문서에서 빈번하게 나타날수록 그 단어의 중요도는 오히려 떨어진다는 개념을 기본으로 한다(Lee et al., 2016).[5] 본 연구에서는 tf-idf 가중치 산정방법을 통해 문서 내 존재하는 단어들의 가중치를 적용하였으며, 그 산정 식은 다음과 같다.

$$w_{t,d} = \log(1+tf) \times \log_{10}(N/df)$$

$w_{t,d}$: tf-idf가중치
tf : term frequency
N : total number of documents
df : document frequency

4. 텍스트 마이닝을 통한 입찰 질의서 분석

본 장에서는 앞서 실시한 텍스트 데이터 전처리 과정을 통해 정제되고 구조화된 데이터를 바탕으로 빈출 단어 분석, 단어 간 연관규칙 분석 및 토픽 분석을 실시하였다.

4.1 입찰 질의서 빈출단어 분석

해외건설공사 입찰 질의서에 대한 텍스트 마이닝 결과 Fig. 1과 같이 빈출 상위단어에 대한 빈도수를 구할 수 있었다. 그 결과를 살펴보면 'bid (1,693건)', 'submitted (1,198건)'와 같이 입찰문서의 제출과 관련한 단어들이 가장 빈번히 등

1) R의 text mining 패키지(tm) 중에서 tolower 함수 적용
2) R의 text mining 패키지(tm) 중에서 removePunctuation 함수 적용
3) R의 text mining 패키지(tm) 중에서 stopwords 함수 적용

4) Table 2를 살펴보면 전처리 과정을 거쳤다고는 하지만 여전히 5,000개 이상의 단어들이 문서 내에 존재하고 있으며, 단어의 희박한 수준을 의미하는 sparsity의 값이 97%인 것을 볼 때 term-document matrix에서 대부분의 단어의 빈도수가 0임을 알 수 있다. 따라서 단어들의 가중치를 차등적으로 부여하여 분석의 효율을 높일 필요가 있다.
5) 일반적으로 모든 문서에서 지속적으로 빈번하게 등장하는 단어는 'the', 'is', 'a' 등과 같은 단어들이 많기 때문에 tf-idf에서는 이러한 단어들의 가중치를 낮게 부여하여 분석과정에서 제외될 수 있도록 한다.

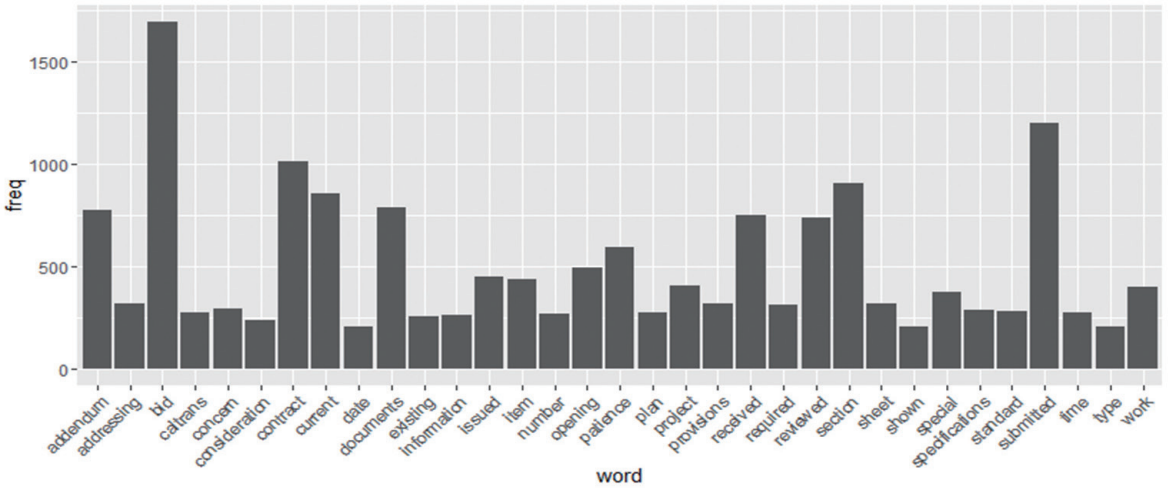

Fig. 1. Word frequency

장한 것을 알 수 있으며, 'contract (1,014건)', 'section (904건)'과 같이 계약문서 및 문서 내 세부 절을 의미하는 단어들도 많이 발견되었다. 또한 'special (374건)', 'provisions (318건)'와 같은 단어들을 통해 다수의 입찰 질의서에서 특수 계약조건에 대한 질의가 이루어졌음을 알 수 있다. 뿐만 아니라 'specification (283건)', 'plan (272건)', 'sheet (318건)'와 같이 설계도서에서 문제가 발생하여 질의서를 작성하였음을 추측할 수 있는 단어들도 다수 등장하고 있는 것을 확인하였다.

4.2 입찰 질의서 빈출 단어 간 연관규칙 분석

빈출 단어 분석을 통해 입찰 질의서에서 반복적으로 사용되고 있는 단어들에 대한 전반적인 파악은 가능하지만, 각 단어들이 문장 내에서 어떠한 의미로 사용되었는지를 알기 위해서는 빈도수 분석만으로는 이해가 어렵다. 따라서 단어들 간의 공통된 관계를 파악하기 위해 빈출 상위단어들에 대한 클러스터링을 실시하였으며, 그 결과 Fig. 2와 같은 덴드로그램을 얻을 수 있었다. Fig. 2의 결과를 살펴보면, 단일 단어로 군집화된 경우를 제외한 나머지 세 가지 군집의 단어들의 조합에서 계약서에 언급된 공사 관련 기한(공기, 서류 제출 기한 등)에 대한 군집과 설계도서에 대한 군집, 시방서 및 특수계약조건과 관련된 군집으로 단어들이 조합된 것을 확인할 수 있었다.

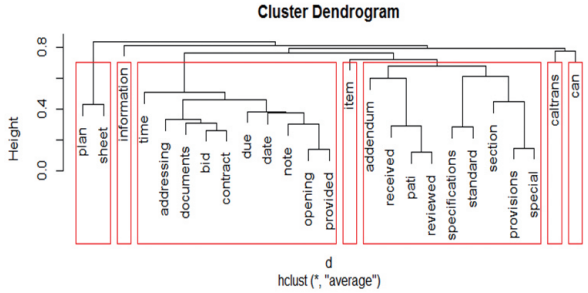

Fig. 2. Text clustering

단어들 간의 보다 유기적 관계를 살펴보기 위해 연관규칙(association rules) 분석을 함께 실시하였다. 연관규칙 분석은 장바구니 분석 사례로도 알려져 있는 데이터 마이닝 기법으로 어떤 사건이 얼마나 자주 동시에 발생하는가를 표현하는 규칙 또는 조건을 의미한다(Yu and Hong, 2015). 입찰 질의서에 대한 연관규칙 분석은 동일한 입찰 질의서 내에 특정 단어가 등장할 때 공통적으로 발견되는 단어들이 무엇인지를 파악하기 위한 목적으로 실시하였다. 예를 들어, 'specification'이라는 단어가 문서 내에서 어떠한 의미로 쓰였는지 이해하기 위해 연관규칙 분석을 실시함으로써 문장 내 의미를 보다 정확히 파악하고자 하였다.

연관규칙 분석을 위한 평가지표는 지지율(support), 신뢰도(confidence), 향상도(lift)로 정의할 수 있는데 본 절의 분석에서 각각의 개념은 다음과 같이 설명할 수 있다.

- 지지도(support) : 전체 발생사건 중 단어 A와 단어 B가 동시에 발생하는 비율
- 신뢰도(confidence) : 단어 A가 발생한 사건 중 단어 B가 포함된 사건의 비율 (조건부 확률)
- 향상도(lift) : 단어 A가 발생한 사건 중 단어 B가 포함된 사건과 단어 B가 발생한 사건과의 비율

분석 데이터를 토대로 연관규칙 분석을 실시한 결과 총 5,167개의 연관규칙이 생성되었다. Fig. 4는 도출된 총 5,167개 연관규칙의 지지도, 신뢰도, 향상도의 분포를 전체적으로 보여주는 산포도로, 대부분 규칙의 신뢰도가 0.1 이상으로 상대적으로 높은 신뢰 수준의 결과가 제시된 것을 알 수 있다.

그러나 신뢰도가 높다고 해서 모든 연관규칙이 의미 있는 규칙은 아니며, 설명이 가능한 규칙도 아니다. 따라서 본 연구에서는 설명력이 높은 몇 가지 규칙들만을 일부 추출하였고, 그 일부를 Table 3에 정리하였다. 분석 결과 제시된 규칙들을 살펴보면 문서 내에서 단어들이 등장할 때 어떤 단어들과 함께 사용되고 있는지에 대한 전반적인 파악이 가능하다.

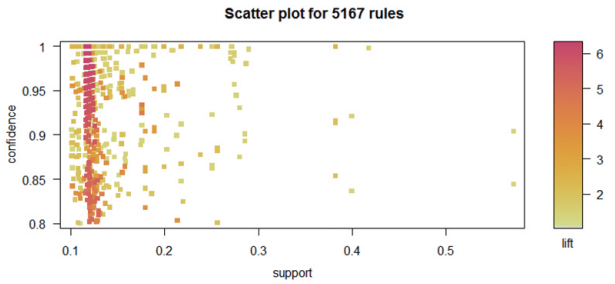

Fig. 3. Scatter plot of text association rules

또한 Fig. 4는 연관규칙 분석에 대한 grouped matrix 차트로서, 열의 요소는 선행, 행의 요소는 결과를 의미한다. 해석하자면, 선행인 LHS (left-hand-side)의 좌측 첫 번째 내용인 '{note, +7 items} – 116 rules'는 'note'를 포함한 최대 7개의 다른 단어와 결과인 RHS (right-hand-side)의 조합으로 이루어진 규칙이 116개 있음을 의미하는 것이다.

Table 3. Results of text association rules

A ⇒ B		support	confidence	lift
term A	term B			
specifications	standard	0.1204	0.6939	4.0412
standard	section	0.1546	0.9005	1.9733
specifications	section	0.1375	0.7923	1.7362
contract, specifications	bid	0.1176	0.8857	1.3111
item, section	bid	0.1138	0.8888	1.3158
date, note, provided	due	0.1176	1	6.3113
contract, date, documents, note	due	0.1157	1	6.3113

4.3 입찰 질의서 토픽 모델링 분석

단어 사이의 관계분석에 대한 이해를 바탕으로 본 절에서는 입찰 질의서에 대한 텍스트 마이닝을 통해 각 문서들이 어떤 주제(토픽)들로 묶일 수 있는지, 문서들에 대한 토픽 모델링을 실시하였다. 앞 절에서 실시한 빈출단어 분석, 단어 간 군집분석, 연관규칙 분석이 개별 단어들 간의 관계를 파악하기 위한 작업이었다면 토픽 모델링 분석은 문서 내에 어떠한 주제의 내용들이 포함되어 있는지를 파악할 수 있는 방법으로, 하나의 문서에 두 개 이상의 주제가 포함될 수 있음을 전제로 한다는 점에서 텍스트 클러스터링과는 차이가 있다.

R 프로그래밍을 통해 토픽 모델링을 수행한 결과 Table 4와 같이 총 5가지의 토픽을 찾을 수 있었다. 1,054건의 입찰질의 문서에 대한 5가지의 토픽을 선정하기 위해 여러 번의 시행착오(trial and error)를 거쳤으며, 가장 유사한 단어들이 하나의 토픽으로 선정된 경우를 최종적으로 선정하였다. 각 토픽에 명시된 단어는 해당 토픽에 대한 설명력이 높은 것으로 선정된 단어들이다.

먼저, Topic 1의 경우 입찰문서 중 시방서와 관련된 단어들이 해당 토픽에 대한 설명력이 높은 단어로 선정된 것을 볼 수 있으며, 따라서 시방서와 관련된 문제들을 질의한 토픽 그룹으로 볼 수 있다. 또한 'material', 'type', 'requirement'와 같은 단어들을 통해 시방서에 기술된 자재 타입 및 요구사항에 대한 문제(오류, 상충, 누락 등)들이 입찰질의서에 작성되었음을 파악할 수 있다. Topic 2는 공사 일반 정보 및 현장 운영과 관련된 토픽으로 이해할 수 있는데, 'water', 'site', 'traffic'과 같은 단어들이 등장한 것을 볼 때 현장운영을 위한 제반사항 및 현장상황에 대한 질의사항이 포함되어 있음

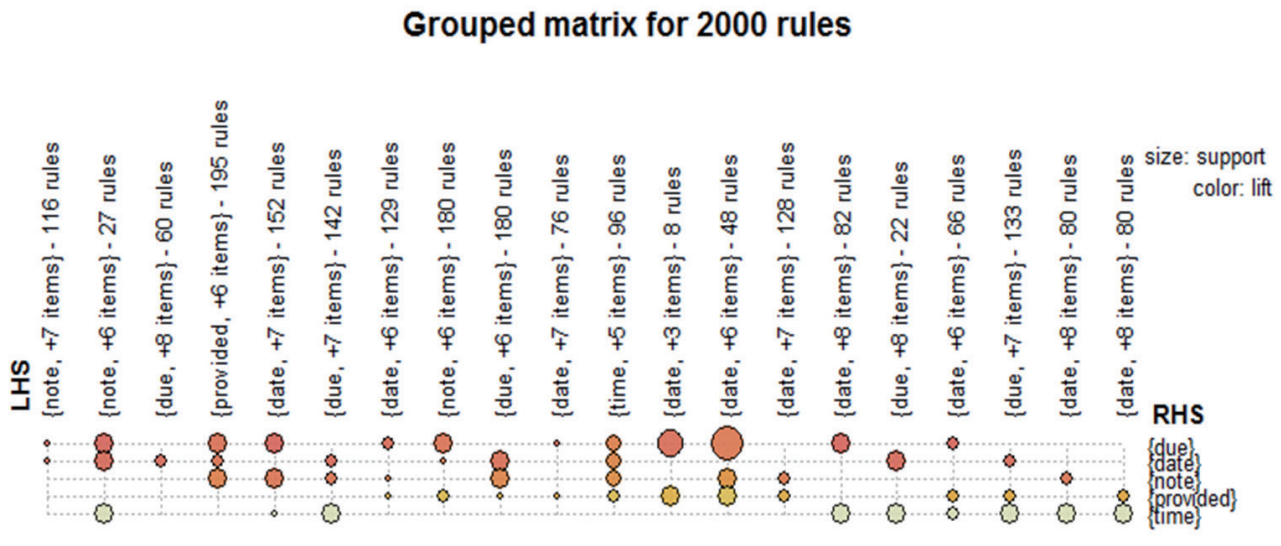

Fig. 4. Grouped matrix for text association rules

을 알 수 있다. Topic 3은 입찰문서에 포함되어 있는 내역서의 Bid item에 대한 토픽으로, 'quantities'라는 단어가 포함된 것으로 볼 때 내역서 물량에 문제가 있어 발주자에게 관련 내용을 질의하였음을 추측할 수 있다. Topic 4는 설계 도면에 있는 정보와 관련된 토픽그룹으로, 설계도면에 명시된 정보를 확인하기 위해 작성된 질의서 문서가 대다수를 이룰 것으로 파악된다. 마지막으로 Topic 5는 계약서 조항과 관련된 토픽 그룹으로, 계약서 내 특수조건 및 관련 정보에 대해 발주자에게 확인을 요하는 내용의 질의서 정보가 포함되어 있는 것으로 판단된다.

Table 4. Topic modeling

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	section	project	bid	sheet	special
2	specifications	due	item	plan	provisions
3	existing	submitted	removal	submitted	standard
4	standard	traffic	included	reviewed	refer
5	material	water	quantities	bridge	information
6	type	information	temporary	shown	page
7	requirement	site	area	plans	contract
8	pavement	question	control	work	current

이상의 내용을 종합하여 각 토픽 그룹의 입찰 질의서 내용의 유형을 정리하면 다음과 같다.

- Topic 1 : 시방서 내 기술된 공사자재 및 타입 등 요구사항에 대한 질의
- Topic 2 : 공사 일반 정보 및 현장 운영, 현장 상황에 대한 질의
- Topic 3 : 내역서 물량 정보에 대한 질의
- Topic 4 : 설계도면 정보에 대한 질의
- Topic 5 : 계약 특수조건 및 조항에 대한 질의

5. 연구의 결론 및 향후 활용 방안

본 연구는 건설 분야의 비정형 텍스트 데이터 분석을 위하여 해외건설공사의 입찰자 질의 정보를 대상으로 텍스트 마이닝을 실시하였으며, 그 결과 빈출단어 유형, 단어들 간의 연관관계, 문서의 주제 유형 등을 파악할 수 있었다. 다시 말해, 과거에 수행되었던 건설공사에서 입찰 참여자들이 입찰문서를 검토할 때 어떤 부분을 중점적으로 검토하고 질의하였는지를 이해할 수 있었다.

본 연구는 텍스트 마이닝을 활용한 해외건설공사 입찰 정보 분석을 통해 직접 개별 문서의 내용을 확인하지 않고도 1,054건이라는 많은 양의 문서들을 종합적으로 파악할 수 있는 방안을 제시했다는 점에서 의미를 찾을 수 있으며, 향후 관련 분야 연구를 확장시킬 수 있는 기반을 마련할 수 있을 것으로 판단된다. 물론, 아직까지 컴퓨터의 자연어 처리 기술이 완벽하지 못하기 때문에 텍스트 분석으로 문맥상의 미세한 의미까지 파악하지는 못하지만 짧은 시간 내에 많은 양의 정보를 효과적으로 분석할 수 있다는 점에서 향후 적용분야가 보다 확대될 수 있을 것이라 생각한다. 또한 관련 분야 연구가 보다 확장된다면 과거에 수행되었던 프로젝트들의 텍스트 데이터를 확보하여 비정형 텍스트 분석을 실시함으로써 과거 수행 프로젝트로부터 중요한 노하우를 획득할 수 있을 것이며, 실패사례에 대한 학습도 가능할 것이라 판단된다.

그러나 본 연구에서 실시한 텍스트 마이닝의 결과는 데이터 전처리 및 정제과정에 크게 영향을 받기 때문에 연구자의 주관적인 판단이 개입될 여지가 있다는 부분에서 한계를 갖고 있으며, 향후 이를 보완할 수 있는 전문가 검토 등의 추가 연구가 이루어질 수 있을 것이라 판단된다. 또한 정보 수집의 어려움으로 인해 미국 건설시장에서 발생한 해외건설공사의 입찰 정보만을 대상으로 분석을 실시한 것이기 때문에 추후 국내 기업들의 진출한 지역의 사례 데이터가 확보된다면 보다 다양한 분석 결과를 제시할 수 있을 것이라 판단된다.

감사의 글

본 연구는 교육부/한국연구재단 이공분야기초연구사업의 지원으로 수행되었습니다. (NRF-2015R1D1A1A02061864)

References

Caldas, C., Soibelman, L., and Han, J. (2002). "Automated Classification of Construction Project Documents." *Journal of Computing in Civil Engineering*, pp. 234-243. 〈10.1061/(ASCE)0887-3801(2002)16:4(234).〉

Caltrans (2016). "Caltrans Bidders Inquiries." http://www.dot.ca.gov/hq/esc/oe/inquiry/bid_inquiries.php.

Kim, J. H., and Kim, Y. S. (2014). "An Analysis of Concentrate Competency in Bidding Process for Overseas Project of Domestic Construction Companies." *Korean Journal of Construction Engineering and Management*, KICEM, 15(3), pp. 23-30.

Lee, J. H., Yi, J. S, and Son, J. W. (2016). "Unstructured Construction Data Analytics Using R Programming – Focused on Overseas Construction Adjudication Cases." *Journal of the Architectural Institute of Korea*, AIK, 32(5), pp. 37-44.

Mao, W., Zhu, Y., and Ahmad, I. (2007). "Applying metadata models to unstructured content of construction documents: A view-based approach. Automation in Construction." 16(2), pp. 242-252. 〈DOI: http://dx.doi.org/10.1016/j.autcon.2006.05.005.〉

Seo, J. P., Ryu, H. G., Son, B. S., and Choi, Y. K. (2016). "The Development of Risk Management Process Model during Bidding Phase for Success of Oversea." *Korean Journal of Construction Engineering and Management,* KICEM, 17(4), pp. 76-86.

Simoff, S. J., and Maher, M. L. (1998). "Ontology-based multimedia data mining for design information retrieval." Computing in Civil Engineering, K. C. P. Wang, T. Adams, M. L. Maher, and A. Songer, eds., ASCE, Reston, Va., pp. 212 - 223.

Tanaka, T. (1988). "Analysis of claims in U.S. construction projects." Master thesis, Massachusetts Institute of Technology, Boston.

Yim, D. (2015). Big data analysis using R, Free academy, pp. 21-50.

Yu, C. H., and Hong, S. H. (2015). R Visualization, Insight.

요약 : 건설 프로젝트에서 생산되는 대부분의 데이터는 텍스트 기반의 비정형 데이터이다. 계약서, 시방서, RFi 등 수많은 텍스트 문서들을 효과적으로 분석하기 위해서는 텍스트 마이닝과 같은 비정형 텍스트 데이터 분석 방법이 필요하다. 이에 본 연구에서는 과거에 수행되었던 해외건설공사 프로젝트의 입찰 관련 문서들을 대상으로 텍스트 마이닝을 실시하였으며, 그 결과 빈출단어의 유형, 단어들 간의 연관관계, 문서들의 토픽 유형들에 대한 파악이 가능하였다. 본 연구는 텍스트 마이닝을 활용한 해외건설공사 입찰 정보 분석을 통해 비정형 텍스트 데이터를 효과적으로 분석할 수 있는 방안을 제시하였다는 점에서 의의가 있으며, 향후 관련 분야 연구를 확장시킬 수 있는 기반을 마련할 수 있을 것이라 기대한다.

키워드 : 텍스트 마이닝, 비정형 텍스트 데이터, 입찰질의서