한국해안 · 해양공학회논문집/ISSN 1976-8192(Print), ISSN 2288-2227(Online)
Journal of Korean Society of Coastal and Ocean Engineers 30(1), pp. 19~28, Feb. 2018
https://doi.org/10.9765/KSCOE.2018.30.1.19

해안해양공학 연구 분야의 SCOPUS 서지정보 Text Mining 분석
Text Mining Analysis on the Research Field of the Coastal and Ocean Engineering Based on the SCOPUS Bibliographic Information

이기섭* · 조홍연* · 한재림*
Gi Seop Lee*, Hong Yeon Cho* and Jae Rim Han*

요 지 : 서지정보학의 발달 및 전산화로 방대한 양의 연구논문들이 축적되고 있다. 이에 따라 전 세계에서 출판되는 관련 분야 논문들을 모두 검토하기는 실질적으로 어려워졌으며, 연구방향을 잡고 추진하는 것도 어려워졌다. 그러나 자연어 처리기법의 발달로 인해 출판된 연구논문들의 경향 분석이 수월해졌다. 여기서는 해안 · 해양공학 분야의 SCOPUS DB(Data Base) 서지정보 텍스트 마이닝(Text Mining) 분석을 R언어를 이용하여 수행했다. 분석 결과, 예상한 바와 같이 'wave' 용어가 압도적으로 우세하였으며, 'numerical model', 'numerical simulation' 및 experimental study' 용어로부터 여전히 수치해석 및 수리실험의 우세가 확인되었다. 또한 최근 해양에너지와 관련되는 'wave energy' 용어 사용이 부각되고 있는 것으로 파악되었다. 한편, 해안 · 해양공학 분야의 연구주제 용어의 빈도와 연결 관계는 'wave -> height, energy' 우세를 정량적으로 확인할 수 있었으며, 향후 세부분야 및 시기별 고해상도 분석 가능성을 제시하였다.

핵심용어 : 텍스트 마이닝, 해안공학, 서지정보, SCOPUS, R

Abstract : Numerous research papers have been accumulated due to the development and computerization of bibliometrics. This made it difficult to review all of the related papers published worldwide to conduct the study. However, due to the development of Natural language processing techniques, the tendency analysis of published research papers has become easier. In this study, text mining analysis using the statistical computing language R was carried out based on the bibliographic information of SCOPUS DB (Data Base) in the field of coastal and ocean engineering. As expected, the term 'wave' predominates, and it was confirmed that numerical analysis and hydraulic experiments were still dominant from the terms 'numerical model', 'numerical simulation', and 'experimental study'. In addition, recent use of the term 'wave energy' related to marine energy has been recognized. On the other hand, it was quantitatively confirmed that the frequency of connection between 'wave', and 'height' or 'energy' prevailed, and suggested the possibility of high resolution analysis by detailed field and period in the future.

Keywords : text mining, coastal engineering, bibliographic, SCOPUS, R

1. 서 론

유사 이래 인간이 가장 많이 사용한 자료의 형태는 단연 문자일 것이다. 죽간(竹簡)부터 Data Base까지 저장 방법은 달라졌지만 저장되는 내용은 각 언어로 기록된 문자 정보이다. 현대에 들어와서 텍스트 자료의 분석은 기업 내부에 대량으로 축적된 자료들을 경영에 이용하려는 BI(Business Intelligence) 분야에서 처음 정의되었지만 구조화되지 않은 자료의 처리가 어려웠기 때문에 주로 수치형 자료의 분석 기법이 발달하게 된다(Grimes, 2007). 1990년대에 이르러서 하드웨어 기술과 Data mining 기법의 발달과 함께 Text mining이 다시 주목받았다(Witten, 2018). 현재는 분야를 막론하고 활발하게 쓰이는 종합 정보 분석 도구이지만 Text mining을 주제로 하는 논문은 해안 · 해양공학 분야에서는 전무하며, 해양관련 분야에서 이를 활용한 사례는 2018년 1월 17일 기준으로 SCOPUS DB에 단 두 건 뿐이다(Hui, 2017; Wu et al., 2018). 타 분야에서는 Text mining을 이용하여 인간의 표현형 분류 연구와(Van Driel et al., 2006) 25년간의 서지정보를 이용한 비즈니스와 공공분야의 성과관리 경향 분석을 수행한 바 있다(Cuccurullo et al., 2016). 이 사례 외에도 경영, 마케팅, 사회과학, 의학 및 기타 분야에서 다양한 연구가 수행되고 있다. 따라서 본 연구에서는 상대적으로 Text mining 기법의 활

*한국해양과학기술원 해양자료실(Corresponding author: Hong Yeon Cho, Ocean Data Science Section, Korea Institute of Ocean Science & Technology, 385 Haeyang-ro, Yeongdo-gu, Busan 49111, Korea, Tel: +82-51-664-3786, hycho@kiost.ac.kr)

용이 미진한 해양 관련 분야의 다양성 제고를 위하여 SCOPUS 서지정보 DB를 이용한 Text 자료 분석 및 활용 가능성을 제시한다.

2. 사용 자료 및 전처리 방법

2.1 사용 자료

분석에 사용한 자료의 출처는 Elsevier 출판사가 운영하는 SCOPUS의 서지정보 Data Base(https://www.scopus.com/)이며, 한국해양과학기술원(KIOST) 해양과학전자도서관을 경유해 다운로드하였다. 본 자료는 SCOPUS와 계약이 되어 있는 기관을 통해 유료로 제공되기 때문에 자료 접근 권한이 있는지 확인이 필요하며, 모든 정보를 포함한 자료의 경우, SCOPUS 회사의 자료제공 정책으로 인해 1회에 2,000개 자료까지만 다운로드가 가능하다. 자료는 2017년 11월 30일을 기준으로 1966년부터 2018년도까지 발간된 논문을 검색한 결과를 사용하였으며 검색 기준년도인 2017에 게재가 확정되어 2018년도에 출간 예정인 논문까지 검색 대상으로 하였다. 검색은 논문의 분야를 나타내는 Source Title 항목에서 “Coastal engineering” 또는 “Ocean engineering” 조건으로 하였으며, 총 13360편의 논문이 검색되었다. 자료의 구조는 저자, 제목, 연도, 초록 등 주요 정보를 비롯하여 총 42개 항목으로 이루어져있으며, 다운로드 시 필요한 항목을 설정할 수 있다. 여기서는 논문 제목인 Title 항목을 이용하여 분석을 수행했다.

SCOPUS에서 제공하는 파일의 형식은 csv, BibTex, text, RIS 등 다양한 형식으로 검색 결과를 제공하나 여기서는 가장 범용성이 높은 파일 형식이고, 엑셀과 호환이 바로 되는 csv(comma separated values) 형식의 파일로 기본 분석을 수행했다.

2.2 분석 도구 및 환경

분석에 사용한 도구는 R 언어이며, 다음 링크(www.r-project.org)에서 무료로 다운로드하여 사용할 수 있다(R Core Team , 2017). R 언어는 오픈소스 소프트웨어로 다양한 기능을 제공하는 라이브러리들을 이용할 수 있다. 본 분석에서는 텍스트 자료의 분석도구로 data.table, dplyr, tidytext, tidyr, pluralize (github, 오픈소스 소프트웨어 공동개발 플랫폼) 패키지를 사용하였으며 시각화 도구로 wordcloud, igraph, ggraph 패키지를 사용하였다(Csardi and Nepusz, 2006; Dowle and Srinivasan, 2017; Fellows, 2014; Pedersen, 2017; Rudis and Embrey, 2016; Silge and Robinson, 2016; Silge and Robinson, 2017; Wickham, Francois, Henry and Muller, 2017; Wickham and Henry, 2017). 추가로 BibTex 형식의 파일을 R 언어의 서지정보 분석 패키지인 ‘Bibliometrics’ 라이브러리를 이용하여 분석했다(Aria and Cuccurullo, 2017). 운영체제는 Windows 7, R 프로그램은 3.4.3 버전의 환경에서 분석을 수행했다.

2.3 문자 자료의 전처리

모든 자료분석 과정에서 전처리(Preprocessing)는 중요하지만 특히 텍스트 마이닝에서 의미 있는 결과를 도출하기 위해서는 세심한 문자 자료의 전처리 과정이 요구된다. 그러나 문자 자료는 수치형 자료와 달리 문자 자료에 직접적인 가공을 가하는 전처리과정부터 분석과정에 포함되는 것이 큰 차이점이기 때문에 구체적인 자료 가공 과정은 본론에서 다룬다. 따라서 여기서는 용어 사용 시의 혼동을 피하기 위해 자료구조 변경 및 병합, 기타 프로그래밍 과정에서 발생하는 기술적인 문제들이 처리된 원시자료를 일반적인 개념의 전처리 자료로 간주한다. 패키지 설치와 자료를 불러들이는 준비과정 및 전처리 과정에서 발생한 이슈 등은 부록에 수록했다.

3. 텍스트 자료의 가공 및 시각화

토큰(Token)이란, 단어, 구 등과 같이 문장에서 의미를 갖는 문자의 집합을 나타내며, 토큰화(Tokenization)는 문장을 여러 개의 Token으로 분할하는 것을 말한다. 이를 통해 각 단어의 빈도, 관계 등을 분석할 수 있다. 개별 단어, 합성어, 구, 절, 등을 임의로 Token화 할 수 있으며, 개별단어는 라틴 계열 접두어로 하나의 의미를 가진 uni-를 붙여 Unigram으로 표현한다. 이와 같은 원리로, 합성어와 같이 2개 이상의 단어들을 연결하여 Token화 한 것을 n-gram이라 한다. 다음은 Unigram을 포함한 n-gram 분석을 수행한 결과이다. 본문의 설명 뒤에는 부록에 첨부한 전체 코드의 재현을 돕기 위해 해당 부분의 코드를 제시하였다.

3.1 Unigram의 가공

먼저 정제된 원시자료를 토큰화하고 불필요한 단어들을 제거한다. 여기서 말하는 불필요한 단어들이란, 관사나 전치사 및 일반적으로 빈번하게 사용되는 부사와 같이 의미 파악에 불필요한 요소들을 말한다. 자료가 단어들로 토큰화 되었다면, 동의어 및 유사어, 영어의 경우 단수형과 복수형 단어들의 통합이 필요하다. 그렇지 않으면 wave와 waves 단어처럼 실제로는 같은 의미를 가지나 단어의 형태에 따라 빈도가 분산되어 분석결과가 왜곡되기 때문이다. 여기서는 Git Hub에 공유된 ‘hrbrmstr’ 유저가 개발 중인 pluralize package의 singularize() 함수를 이용하여 복수형 단어를 단수형으로 변환 하였다.

텍스트 마이닝의 주제에 따라 다르지만 본 내용에서 다루는 학술논문의 Title이나 Keyword에 관한 분석은 Token화 시킨 모든 어절에 대하여 수행하는 것은 불필요하거나 부적절할 수 있다. Title의 Unigram 분석에서는 명사와 형용사 이외의 품사에서 의미 있는 정보를 추출할 기댓값이 낮다는 정성적인 판단 하에 제거를 결정했다. 특히 학술논문의 Keyword에 기입하는 단어들은 주로 명사이며 복합어일 경우 분사형태의 형용사를 사용하기도 하지만 n-gram 이상의 Token에 해

당하는 이야기이므로 Unigram Token화에서는 명사와 형용사만 추출했다.

3.2 n-gram의 가공

2음절 이상의 n-gram(Bigram) 처리과정에서도 Unigram과 동일한 방법으로 불필요한 단어들을 제거하고, 단수형 변환 과정을 거쳤다. 그러나 전문 분야에서 사용하는 복합명사의 정보손실을 고려하여 품사별 추출은 하지 않았다.

3.3 Unigram 시각화

다음은 3.1에서 가공된 Unigram 결과물을 wordcloud() 함수로 시각화한 결과이다(Fig. 1).

해안 · 해양공학 분야의 논문 제목에서 추출한 단어들의 빈도를 비교한 결과 동 분야의 정체성을 나타내는 핵심어는 'wave'로 나타났다. Fig. 2는 빈도수 600 이상의 단어들을 나타낸 것으로 파랑의 빈도가 압도적으로 높다.

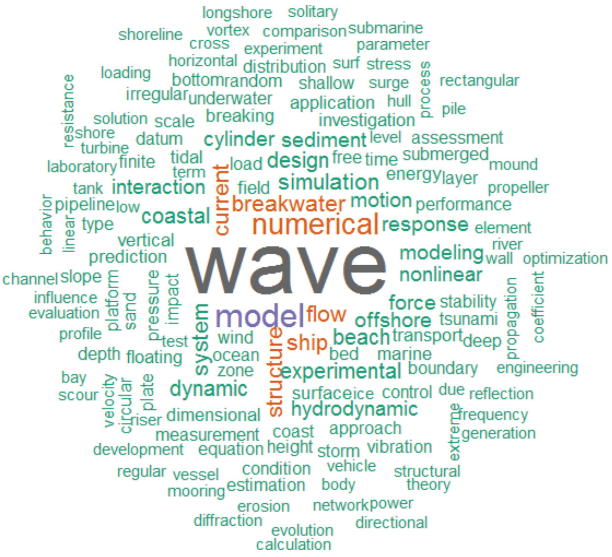

Fig. 1. Word cloud of Unigram in Coastal & Ocean Engineering articles.

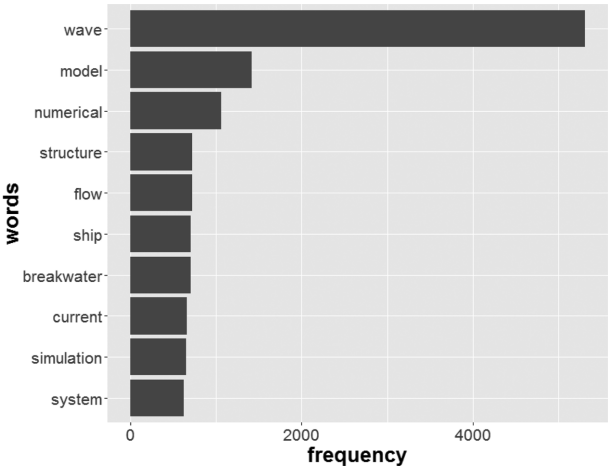

Fig. 2. Word frequency in Coastal & Ocean Engineering articles.

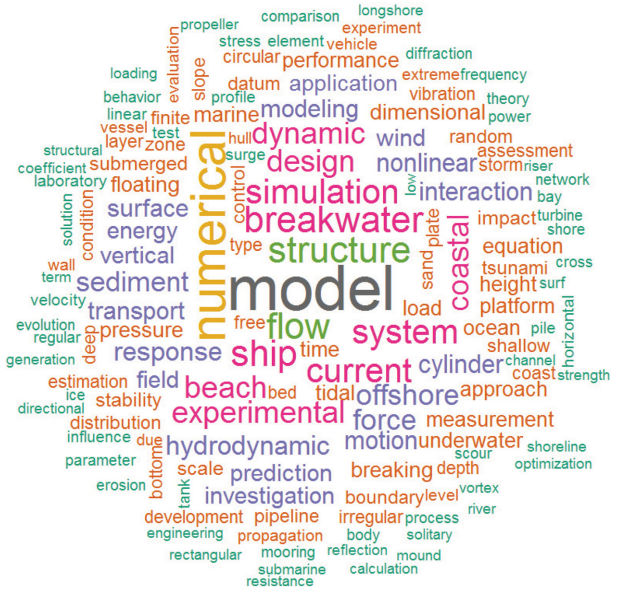

Fig. 3. Word cloud of Unigram in Coastal & Ocean Engineering articles except 'wave'.

다음으로는 연구방법을 나타내는 model과 numerical이 뒤를 이었고, 상대적으로 experimental, observation, measurement와 같은 단어들은 그 빈도가 많이 떨어졌다. 이를 통해 1966년부터 현재까지 실험 및 관측 연구보다 모델링 연구가 더 활발히 이루어졌음을 알 수 있다.

Fig. 1에서 'wave'의 압도적인 빈도 때문에 상대적으로 작아보였던 단어들을 보기 위해 'wave'를 제외시키고 시각화 결과를 검토했다(Fig. 3). 해양-해양공학 분야의 기본 단어에 해당하는 'wave' 단어를 제외하는 경우, 세부 관련 분야의 용어빈도를 보다 자세하게 파악할 수 있다. Fig. 3에서 볼 수 있는 바와 같이, 'model', 'numerical' 용어 빈도가 부각되고, 'simulation', 'breakwater', 'experimental', 'system', 'dynamic', 'design' 등의 용어 사용빈도도 부각되고 있다. 본 그림은 하나의 단어 빈도만을 제시하고 있으나, 단어의 가능한 조합을 감안한다면, 'numerical simulation', 'breakwater-design', 'experimental design' 등의 조합으로부터 수치실험, 수리실험 등의 기법에 근거한 해안구조물 설계 등의 주제가 부각되는 것으로 판단할 수 있다.

3.4 Bigram 시각화 - word cloud

다음은 기존의 단어 하나만의 빈도만으로는 구체적인 주제 파악이 한계가 있기 때문에, 두 개의 단어조합을 이용하여 사용빈도를 분석하였다. 기존의 단어를 Token화하여 전처리를 한 Bigram을 시각화한 결과이다(Fig. 4).

Unigram에서 나타난 것처럼 연구 방법인 수치해석('numerical simulation')이 여러 가지 형태로 높은 빈도를 보이고 있고, 'wave'와 관련된 단어들이 많이 보이며 그 중 wave height 관련 연구가 가장 활발한 것으로 보인다. 또한 Fig. 1~3에서는 크게 두드러지지 않았던 sediment는 sediment transport라는

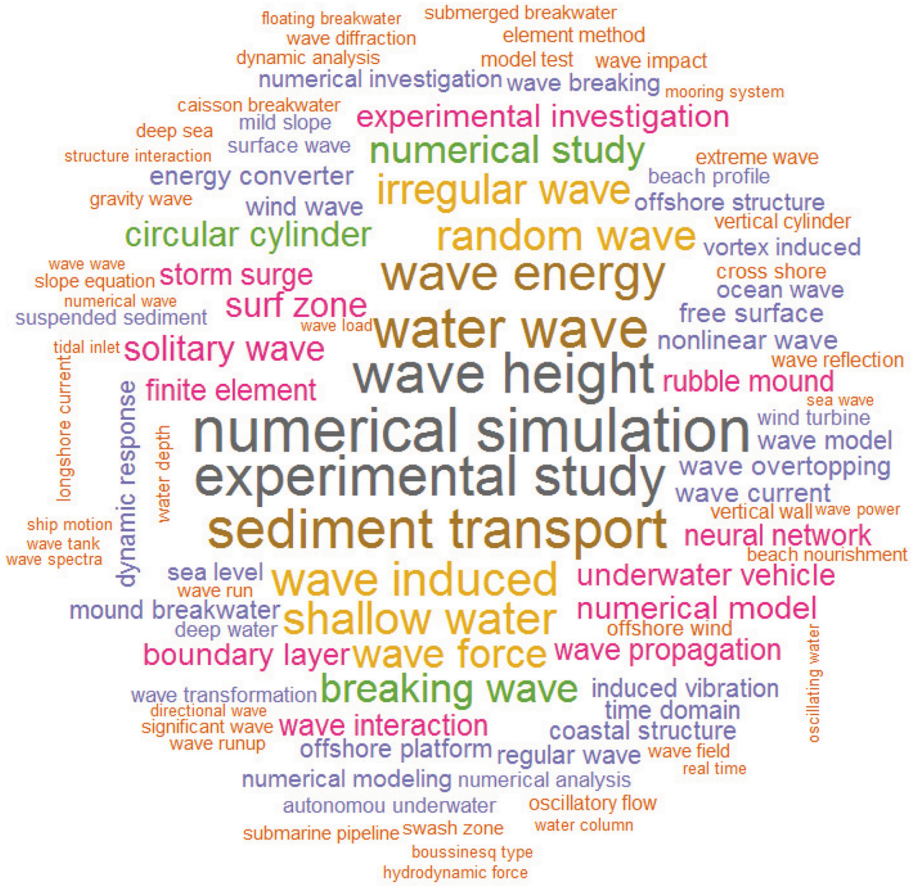

Fig. 4. Word cloud of Bigram in Coastal & Ocean Engineering articles.

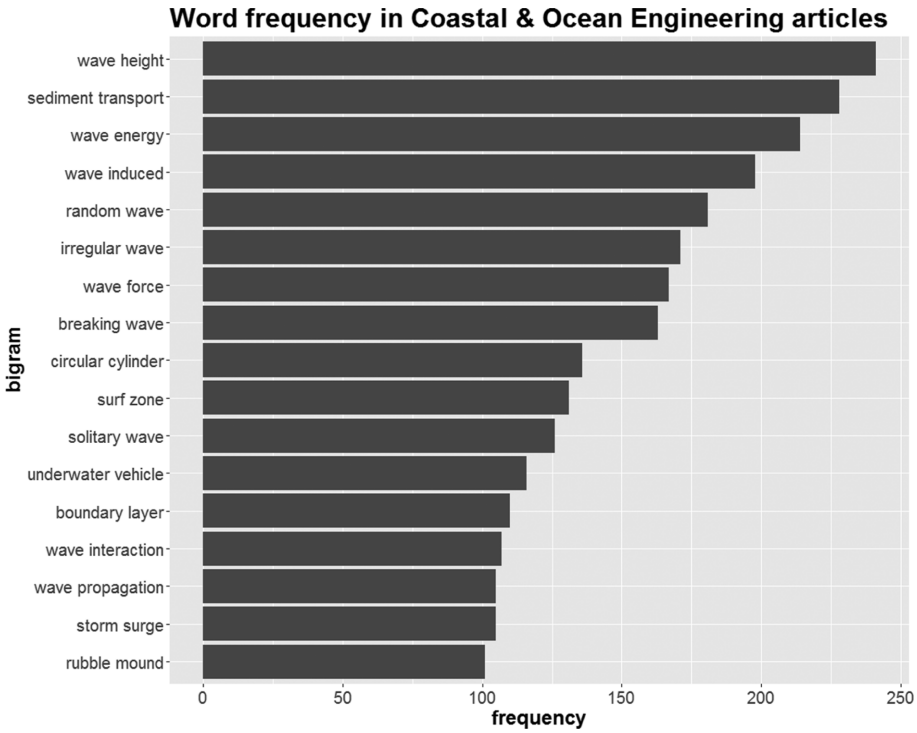

Fig. 5. Bigram frequency in Coastal and Ocean Engineering articles.

합성어의 빈도가 높게 나타났다. 더불어 wave energy에 관한 관심도 높은 것으로 나타났다. Fig. 5는 numerical simulation과 같은 연구방법들을 제외한 연구 주제에 집중한 단어들의 빈도를 나타낸 것이다. wave height와 sediment transport를

필두로 wave 관련 연구들이 뒤를 잇고 있다. Bi-gram 분석에서는 단어의 조합을 통하여 보다 구체적인 연구 주제를 파악할 수 있는 장점이 있다.

3.5 Bigram 시각화 - word network

단어가 두 개 이상인 경우, 각각의 단어와 묶여서 출현하는 단어들은 무엇인지 생각해볼 수 있는데, 이 연결의 빈도를 시각화 할 수 있다. Fig. 6은 Bigram으로 Token화 한 단어들로 network plot을 그린 것이다.

Fig. 4의 결과처럼 Fig. 6에서도 'wave'를 중심으로 단어들이 연결되어 있다. 특히 wave → height와 wave → energy의 연결 빈도가 높은 것으로 나타났고 wave이외의 단어들에서는 앞서 언급한 수치모델 연구 방법과 표사이동(sediment transport)이 두드러진다. 특이한 점으로는 Fig. 6의 좌측 중간에 보이는 인공신경망(artificial neural network) 단어인데, 최근 기계학습 분야의 약진이 해안 · 해양공학 분야까지 영향을 미치고 있다고 해석해볼 수 있으며 그 사례로 Kim and Suh(2011)의 연구를 들 수 있다. 더불어 'extreme wave' 단어 연결빈도는 연안재해-재난 주제가 부각되는 것으로 판단할 수 있다.

4. 토 의

텍스트 자료의 분석 방법은 자료의 길이, 형식, 주제에 따라 다양하다. 앞선 분석과정에서는 연구내용이 압축되어 정성적인 처리가 어느 정도 되었다고 간주할 수 있는 Title 항목을 이용하여 분석하였으나, 극단적인 예로 Keyword 항목은 이미 그 단어의 빈도 자체가 중요도라고 할 수 있다. 그러나 Abstract처럼 문단 이상의 분량을 가진 문자자료에서 반복되는 단어들은 가중치를 보정해주어야 한다. 예를 들어, 과학 분야 또는 본 분석에서 조명한 해안공학 분야의 논문 초록에서 자주 등장할 것으로 예상되는 Analysis 등과 같은 일반 단어는 Text Mining에서 중요한 단어들을 가리는 장애요소가 될 수 있다. 따라서 단어들의 가중치를 보정해주는 TF-IDF(Term Frequency-Inverse Document Frequency) 등의 이론적 기반도 참고해야한다.

문자 자료는 전형적인 비정형 자료이다. 따라서 원시자료를 처리하는 과정에서는 항상 예기치 못한 상황을 염두에 두어야 하며 각 분석 분야의 전문지식이 필요할 경우도 발생한다. 앞서 관사, 전치사 등과 같이 따로 의미를 갖지 않는 단어들은 stop_words 리스트를 통해 제거하였다. 경우에 따라

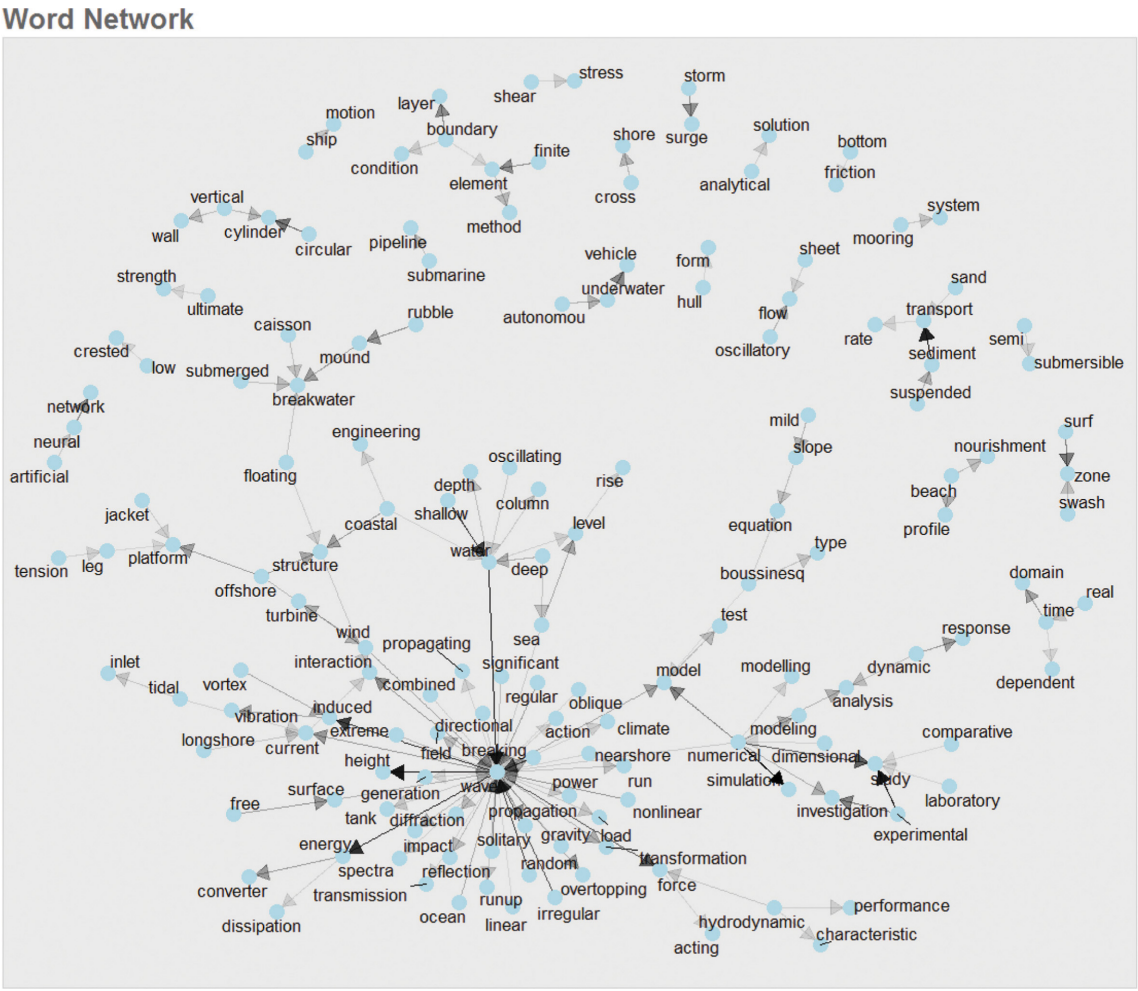

Fig. 6. The word network of bigram in Coastal & Ocean Engineering articles.

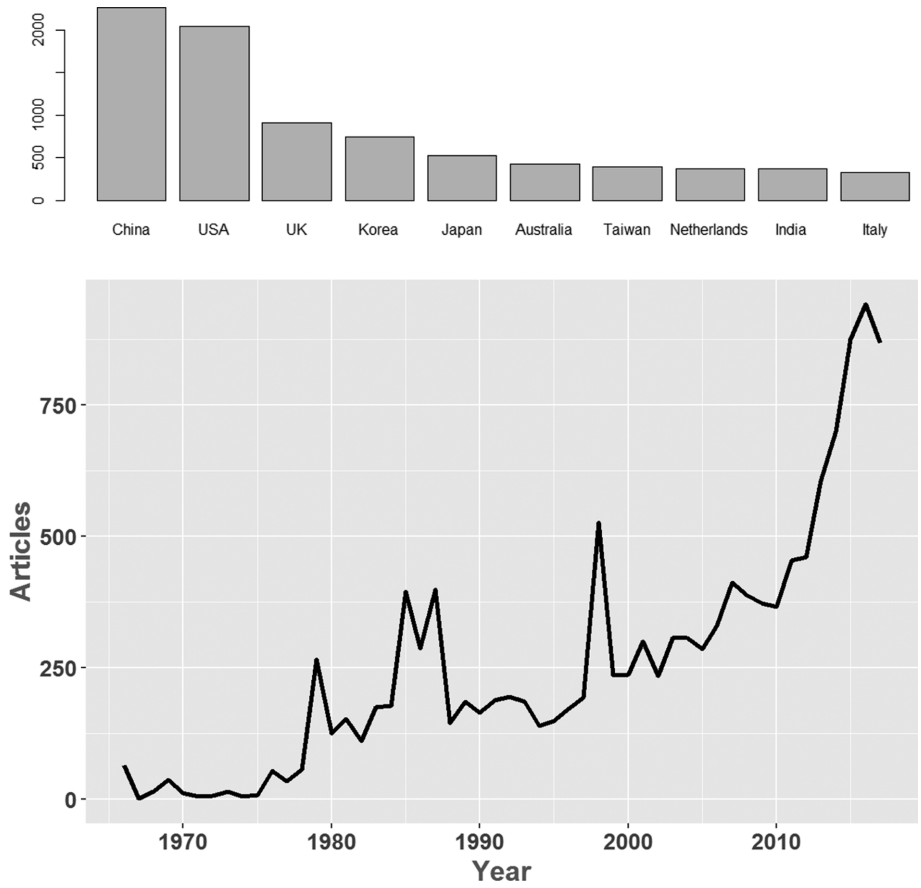

Fig. 7. Total article production by country (top) and annual production of articles (bottom) in the Coastal & Ocean Engineering field (1966~).

이 과정에서 분석 분야의 지식을 기반으로 불필요한 단어를 제거해야 하는데, 그렇기 때문에 일반 단어를 대상으로 하는 함수를 통해 일괄 처리하기가 어려워 해당 분야 전문가의 수작업이 필요하다. 이 외에도 동의어(synonym)의 처리나 어간추출(stemming)등의 과정이 필요할 가능성이 높다. 거의 모든 자료가 그렇겠지만 문자자료의 경우 특히 이러한 전처리 과정들이 분석과정의 대부분을 차지하기 때문에 여기에서 다루지 못한 다양한 문제가 발생할 수 있다.

논문의 서지정보는 문자 자료 자체가 내포하고 있는 정보가 많고, 다양한 항목이 존재하며 시간에 따른 분류가 가능하기 때문에 다변량 및 시계열자료의 특성을 공유하고 있다. 따라서 전체 기간에 대하여 상대적으로 낮은 해상도로 살펴본 본 연구 결과에서 더 나아가 국가별, 시기별, 세부 분야별로 고해상도의 추가분석이 가능하다(Fig. 7).

5. 결 론

SCOPUS 서지정보에 저장된 문자 자료를 이용하여 개별 빈도 및 단어 연관 빈도를 분석하였다. 분석결과 해안 · 해양공학 분야에서 수행된 다양한 연구방법 및 주제들에 대한 경향을 정량적으로 확인해볼 수 있었다. 1966년부터 현재까지 해안 · 해양공학 분야는 파랑과 관련된 연구들이 주를 이루었으며 단어들의 연결 빈도 역시 파랑을 중심으로 파생되는 형태를 보였다. 파랑관련 연구 이외에는 다양한 연구 주제들이 전반적으로 고르게 분포하였다. 또한 관측 및 실험연구보다는 모델링연구가 더 활발한 것으로 나타났고, 근래에 다양한 영역에서 활용하고 있는 인공신경망도 분석 결과에 나타났다. 현재 동 분야에서 연구된 Text mining 사례가 전무하기 때문에 본 연구를 통해 세부 연구분야 및 특정 시기에 대한 구체적인 분석 가능성을 확인하였다.

감사의 글

본 연구는 한국해안해양공학회와 적조피해 최소화를 위한 적조탐지 · 예측시스템 구축 및 실증화(PM60650) 과제의 지원을 받아 수행되었습니다. SCOPUS 자료 이용에 도움을 주신 한국해양과학기술원의 해양과학도서관에도 감사를 드립니다.

References

Aria, M. and Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11(4), 959-975.

Csardi, G. and Nepusz, T. (2006). The igraph software package for

complex network research. InterJournal, Complex Systems, 1695(5), 1-9.

Cuccurullo, C., Aria, M. and Sarto, F. (2016). Foundations and trends in performance management. A twenty-five years bibliometric analysis in business and public administration domains. Scientometrics, 108(2), 595-611.

Dowle, M. and Srinivasan, A. (2017). data.table: Extension of `data.frame`. R package version 1.10.4-3. https://CRAN.R-project.org/package=data.table.

Fellows, I. (2014). wordcloud: Word Clouds. R package version 2.5. https://CRAN.R-project.org/package=wordcloud.

Grimes, S. (2007). Brief history of text analytics. http://www.b-eye-network.com/view/6311. [Google Scholar].

Hui, I. (2017). Shaping the Coast with Permits: Making the State Regulatory Permitting Process Transparent with Text Mining. Coastal Management, 45(3), 179-198.

Kim, S.W. and Suh, K.D. (2011). Prediction of Stability Number for Tetrapod Armour Block Using Artificial Neural Network and M5Model Tree. Journal of Korean Society of Coastal and Ocean Engineers, 23(1), 109-117 (in Korean).

Pedersen, T.L. (2017). ggraph: An implementation of grammar of graphics for graphs and networks. R package version 0.1, 1.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rudis, B. and Embrey, B. (2016). pluralize: Pluralize and Singularize Any (English) Word. R package version 0.1.0. http://github.com/hrbrmstr/pluralize.

Silge, J. and Robinson, D. (2017). Text mining with R: A tidy approach. O'Reilly Media, Inc., Sebastopol, C.A.

Silge, J. and Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in r. The Journal of Open Source Software, 1(3).

Van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G. and Leunissen, J.A. (2006). A text-mining analysis of the human phenome. European Journal of Human Genetics, 14(5), 535-542.

Wickham, H., Francois, R., Henry, L. and Muller, K. (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr.

Wickham, H. and Henry, L. (2017). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.7.2. https://CRAN.R-project.org/package=tidyr.

Witten, IH. (2018). Text mining. https://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf, [accessed 18.01.12].

Wu, Y., Xie, L., Huang, S.L., Li, P., Yuan, Z. and Liu, W. (2018). Using social media to strengthen public awareness of wildlife conservation. Ocean & Coastal Management, 153, 76-83.

Received 31 January, 2018
Accepted 18 February, 2018

부록 : 각각의 분석 단계에 대한 R text mining source code.

package loading

```
if(!require(pacman)) {install.packages("pacman")}

pacman::p_load(dplyr, data.table, tidytext, tidyr,
               ggplot2, wordcloud, wordcloud2, tm,
               openNLP, bibliometrix, igraph, ggraph)

pacman::p_load_gh('hrbrmstr/pluralize')
path_csv <- paste0(dir("csv"))
```

원시자료 전처리

```
data_list_csv <- lapply(paste0("csv/",path_csv), fread)
data_list_csv[[5]][,ISBN:=as.character(ISBN)]
data_list_csv[[6]][,ISBN:=as.character(ISBN)]
data_list_csv[[7]][,ISBN:=as.character(ISBN)]
raw_csv <- rbindlist(data_list_csv)
raw_csv$Abstract[raw_csv$Abstract=="[No abstract available]"] <- NA
rm(data_list_csv)
```

자료 구조 확인

```
str(raw_csv)
```

Unigram Token화

```
tut <- raw_csv %>%
  unnest_tokens(word, Title) %>%
  filter(!word %in% stop_words$word)
```

단수화(Singularizing) 전의 단어 빈도

```
tut %>%
  count(word, sort = T)
```

```
## # A tibble: 8,176 x 2
##           word     n
##          <chr> <int>
##  1        wave  3542
##  2       waves  1771
##  3       model  1132
##  4   numerical  1060
##  5    analysis  1012
##  6       study   887
##  7       water   881
##  8      method   621
##  9        flow   604
## 10 experimental   564
## # ... with 8,166 more rows
```

단수화(Singularizing)

```
tut$word <- singularize(tut$word)
```

품사부여(POS tagging, Part of Speech tagging) 및 추출

```
unigram_title <- tut %>%
  count(word, sort = T) %>%
  merge(., parts_of_speech, by = "word") %>%
  arrange(desc(n)) %>%
  filter(pos == "Noun"| pos == "Adjective", n > 100)
```

Bigram Token화 및 가공

```
tbt_count <- raw_csv %>% select(Title) %>%
  unnest_tokens(bigram, Title, token = "ngrams", n = 2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  count(word1, word2, sort = T)

tbt_count$word1 <- singularize(tbt_count$word1)
tbt_count$word2 <- singularize(tbt_count$word2)

tbt <- tbt_count %>%
  unite(bigram, word1, word2, sep = " ")
```

Unigram word cloud(Fig. 1)

```
unigram_title %>%
  filter(!word %in% c("study", "analysis", "water",
                      "effect", "method", "sea")) %>%
  with(wordcloud(word , n, max.words = 150, random.order = F,
                 scale = c(5,.8),
                 colors = brewer.pal(8, "Dark2")))
```

빈도 600 이상의 Unigram 시각화(Fig. 2)

```
unigram_title %>%
  filter(n > 600,
         !word %in% c("study", "analysis", "water",
                      "effect", "method", "sea")) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  ggtitle("Word frequency in Coastal & Ocean Engineering articles") +
  xlab("words") +
  ylab("frequency") +
  coord_flip()+
  theme(title = element_text(size = 20, face = "bold"),
        axis.text=element_text(size=12),
        axis.title=element_text(size=16,face="bold"))
```

Unigram word cloud(Fig. 3)

```
unigram_title %>%
  filter(!word %in% c("wave", "study", "analysis",
                      "water", "effect", "method", "sea")) %>%
  with(wordcloud(word , n, max.words = 150, random.order = F,
                 scale = c(5,.8),
                 colors = brewer.pal(8, "Dark2")))
```

Bigram word cloud(Fig. 4)

```
tbt %>%
  with(wordcloud(bigram , n, max.words = 50, random.order = F,
                 scale = c(3,.4),
                 colors = brewer.pal(8, "Dark2")))
```

빈도 100 이상의 Bigram 시각화(Fig. 5)

```
tbt %>%
  filter(n > 100,
         !bigram %in% c("water wave", "numerical simulation",
                        "experimental study", "numerical study",
                        "numerical model", "experimental investigation",
                        "finite element", "shallow water")) %>%
  mutate(bigram = reorder(bigram, n)) %>%
  ggplot(aes(bigram, n)) +
  geom_col() +
  ggtitle("Word frequency in Coastal & Ocean Engineering articles") +
  xlab("bigram") +
  ylab("frequency") +
  coord_flip()+
  theme(title = element_text(size = 20, face = "bold"),
        axis.text=element_text(size=15),
        axis.title=element_text(size=18,face="bold"))
```

Bigram word network(Fig. 6)

```
bigram_graph <- tbt_count %>%
  filter(n > 50) %>%
  graph_from_data_frame()

a <- grid::arrow(type = "closed", length = unit(.15, "inches"))

ggraph(bigram_graph, layout = "auto") +
  geom_edge_link(aes(edge_alpha = n*100), show.legend = F,
                 arrow = a, end_cap = circle(.07, "inches")) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 0.5) +
  theme(axis.line=element_blank(),
        axis.text.x=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks=element_blank(),
        axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        legend.position="none",
        #panel.background=element_blank(),
        panel.background=element_rect(colour = "lightgray"),
        panel.border=element_blank(),
        panel.grid.major=element_blank(),
        panel.grid.minor=element_blank(),
        plot.background=element_blank())
```