Research Paper

Analysis of Research Trends Based on Text Mining in the Journal of Korean Society of Environmental Engineering

Donghyeon Lee¹ · Sojeong Lee² · Dokyeong Lim² · Jongkwan Park^{1,2,*}

¹School of Civil, Environmental and Chemical Engineering, Changwon National University

²Department of Smart Ocean Environmental Energy, Changwon National University

(Received December 28, 2020; Revised January 5, 2021; Accepted January 7, 2021)

Objectives : The purpose of this study is to analyze research trends based on text mining technology from the published papers in the Journal of Korean Society of Environmental Engineering.

Methods : From 2000 to 2019, a total of 2,743 published papers were analyzed using text mining techniques. Term frequency, TF-IDF for document classification, word association analysis were applied to find the characteristics of text data.

Results and Discussion : When confirming the high appearance of the word in the published paper during 2000-2019, 'adsorption', 'heavy metals', 'activated carbon', 'sediment', 'sewage sludge' was found in order. It implies that large number of studies in the journal were focused on the water quality field mainly. TF-IDF analysis classified the studies into five groups; 1) drinking water treatment field, 2) water quality modeling field, 3) heavy metal adsorption field, 4) biological sewage treatment field, 5) environmental catalyst field. These results by TF-IDF show that a large proportion of studies were published in the field of water quality modeling and biological sewage treatment. When we analyzed the term frequency every five years, "adsorption" and "heavy metals" were the highly-frequency occurrence words from 2000 to 2009, but in the last 5 years, new words such as "fine dust", "cesium", and "ecological toxicity" were appeared. It seems that the research was reflected in the recent environmental issues.

Conclusions : A lot of studies has been focused on the field of water quality but in recent years, new research topics are being studied related to atmosphere, toxicity, and radiation. Applying a more sophisticated and diverse text mining technique will be of great help to improve the environmental engineering research field.

Keywords : Text Mining, The Journal of Korean Society of Environmental Engineering, Term Frequency, TF-IDF, Word Association Analysis

The Korean text of this paper can be translated into multiple languages on the website of <http://jksee.or.kr> through Google Translator.

* Corresponding author E-mail: jkpark2019@changwon.ac.kr Tel: 055-213-3742 Fax: 055-213-3749

연구논문

텍스트 마이닝을 이용한 대한환경공학회지 연구 동향 분석

이동현¹ · 이소정² · 임도경² · 박종관^{1,2,*}

¹창원대학교 토목환경공학융합공학부

²창원대학교 스마트해양환경에너지공학 협동과정

목적: 텍스트 마이닝 기법을 적용하여 대한환경공학회지 논문을 분석함으로써 현재까지 환경분야 연구동향을 살펴보고 앞으로의 방향성을 제시하는 것을 목적으로 한다.

방법: 2000년부터 2019년까지 대한환경공학회지에 게재된 총 2,743개의 논문 자료를 텍스트 마이닝 기법을 적용하여 분석하였다. 분석 방법은 빈도분석, TF-IDF 기법을 통한 논문 군집 분류, 20년의 자료를 4개의 구간(2000년-2004년, 2005년-2009년, 2010년-2014년, 2015년-2019년)으로 구분하여 각 기간별 연구분야들의 특성을 분석한 시계열 분석, 동시 빈출 빈도가 높은 단어들의 관계를 분석하는 연관어 분석을 진행하였다.

결과 및 토의: 2000년-2019년 동안 게재된 논문에서 제시한 주제어를 바탕으로 출현 빈도가 높은 단어를 확인하였을 때, ‘흡착’, ‘중금속’, ‘활성탄’, ‘퇴적물’, ‘하수슬러지’ 순서로 분석되었고 이것은 그동안 수질분야에 중점적으로 연구가 집중되었음을 보여주는 결과이다. TF-IDF 분석 결과 총 5개의 군집(정수처리분야, 수질 모델링 분야, 중금속 흡착 분야, 생물학적 하수처리 분야, 환경 촉매 분야)으로 분류되었으며 이 중에서 수질 모델링 분야와 생물학적 하수처리 분야 논문의 숫자가 가장 많았다. 시계열 분석 결과 ‘흡착’, ‘중금속’은 매년 많이 연구되어지는 분야였고, 최근 5년 사이에는 환경이슈를 반영하는 키워드인 ‘미세먼지’, ‘세습’, ‘생태독성’ 등의 새로운 단어들이 확인되었다. 연관어 분석을 통해서도 흡착, 활성탄, 중금속과 연관된 다양한 연구가 활발히 진행되어 왔음을 확인하였다.

결론: 텍스트 마이닝 기법으로 대한환경공학회지를 분석한 결과 많은 연구가 수질 분야에 집중되어 있는 것으로 확인되었고, 최근에는 대기, 독성, 방사능 분야로의 연구들도 진행되는 것으로 보인다. 추후에 더 정교하고 다양한 텍스트 마이닝 기법을 적용하여 분석한다면 환경공학분야의 연구 방향제시에 큰 도움이 될 것이라 판단된다.

주제어: 텍스트 마이닝, 대한환경공학회지, 빈도분석, TF-IDF, 연관어 분석

1. 서론

4차 산업 혁명 시대가 도래함에 따라 사회전반에 많은 변화가 일어나고 있으며 연구분야에서도 새로운 형태의 융합 연구들이 진행되고 있다. 최근에는 환경공학 분야에서도 빅데이터와 인공지능 기술을 결합한 융합기술의 연구가 활발하게 진행되고 있다.¹⁻³⁾ 모바일, 사물인터넷(IoT, Internet of Things), 빅데이터(Big Data) 등의 ICT (Information and Communication Technologies) 기술이 실생활 곳곳에 빠른 속도로 적용되고 있고, 정보의 흐름과 생산속도가 점차 가속화되고 있는 상황에서 중요한 자료들을 보다 체계적으로 활용하기 위한 필요성이 제기되고 있다. 학문간의 경계를 넘어서 융합과 이로 인하여 발생하는 폭발적인 데이터 증가는 기존에 진행되던 환경공학의 연구 분야에서 변화를 일으키고 있다. 환경 매체를 중심으로 이루어지던 환경 분석을 넘어서

대용량의 분석 데이터를 바탕으로 다가올 상황을 예측하거나 문제점을 파악하는 데이터 중심 연구가 늘어나고 있다.⁴⁻⁶⁾ 환경분야에서도 ICT 기술로부터 발생하는 대용량의 다양한 데이터 활용이 중요해지고 있는 상황에서 본 논문에서는 텍스트 마이닝(Text Mining)이라는 기술을 활용하여 대한환경공학회지를 분석하고 이를 바탕으로 현재까지의 환경분야의 연구동향을 살펴보고자 한다.

대한환경공학회지는 대한환경공학회에서 발간하는 전문 학술지이며, 1979년부터 발간을 시작하여 매월 1회, 연간 12회의 학회지를 발간하고 있다. 오랜 역사와 전통을 가지고 있는 학회지로서 많은 연구자들의 연구 결과물들이 축적되어 국내 환경공학 연구에 큰 역할을 하고 있다. 수질, 대기, 토양, 지하수, 폐기물, 환경에너지 등 환경공학 기술 전반에 관한 연구 내용을 출판하고 있고, 1979년부터 2019년 4월까지 발행 권호수는 4,058편이고, 수록된 논문은 총 4,266편이다. 최

Fig. 1. Search results for data including “text mining” from 2000 to 2020 on Google scholar.

근에는 친환경 또는 그린 뉴딜의 정책적인 기조와 더불어 환경에 관한 사회적 관심이 나날이 높아지고 있는 상황에서 자연환경의 매체(물, 대기, 토양)를 중심으로 오염원 제거를 목적으로 하는 전통적인 환경공학의 범위에 최근에는 자원 회수, 기후변화, 해수담수화, 에너지 분야까지 포괄하는 현재의 환경공학 연구범위는 나날이 그 확장성을 넓히면서 성장하고 있다.⁷⁻¹⁰⁾ 따라서 지금까지 축적된 다량의 연구결과 자료를 활용하여 기존의 연구 흐름을 분석하는 일은 앞으로의 발전적 연구 방향을 설정하는데 의미 있는 역할을 할 수 있다.

텍스트 마이닝은 컴퓨터를 활용해서 텍스트 데이터로부터 숨겨진 의미를 추출해 내는 과정으로 정의된다.¹¹⁾ 데이터는 크게 정형화된 데이터(structured data)와 비정형화된 데이터(unstructured data)로 구분되는데 정형화된 데이터는 일정한 형식과 조건을 갖춘 형태의 데이터(표)이며, 비정형화된 데이터는 워드프로세서, e-mail, 프레젠테이션, PDF와 같은 복합 문서와 인터넷 페이지 등 비정형 데이터로 구성된 정보이다.¹²⁾ 텍스트 마이닝 기법은 자연어로 구성된 비정형 및 반정형 텍스트 데이터에서 패턴 또는 관계를 추출하여 가치와 의미 있는 정보를 찾아내는 마이닝 기법으로 이는 사람들이 말하는 언어를 이해할 수 있는(비정형화된) 자연어처리(Natural Language Processing)기술에 기반하고 있다.¹³⁾ 최근에 다양한 분야에서 텍스트 마이닝을 활용한 연구결과들이 발표되고 있는데, Google scholar (<https://scholar.google.com/>)를 통해서 연도별 ‘text mining’이라는 키워드를 포함한 자료를 검색한 결과를 Fig. 1에 나타내었다. 2000년도 448건에 불과했던 자료의 숫자는 매년 꾸준히 증가하여 2019년에는 2000년대의 결과값보다 약 47배 큰 23,000건의 자료가 되었다. 이는 ICT 기술이 발달하면서 인터넷 기반의 다양한 정보들이 빅데이터를 형성하며 실시간으로 축적되고 있고, 다양한 텍스트 마이닝 분석법의 개발, 컴퓨팅 파워의 개선, 전자정보 저장기술의 발달로 인한 결과라 할 수 있다.¹⁴⁾

국내의 경우는 Park 등¹⁵⁾은 국가과학기술정보센터에서 논문 350여 편을 수집하여 텍스트 마이닝을 활용하여 환경 분야에 ICT기술을 활용하는 연구동향을 분석하였다. 그 결과 ‘환경 일반’ 및 ‘기후’분야의 키워드 출현이 전체 중 77%를 차지했으며, ICT 기술은 ‘공공융합서비스’ 및 ‘산업융합서비스’에 주로 적용되고 있음을 확인하였다. 또한 시계열 분석을 통해 ICT 활용연구가 최근 활발하게 진행되고 있음을 수치적으로 입증하였다. Oh 등¹⁶⁾은 환경공간정보 연구 동향에 텍스트 마이닝 분석을 진행하여 위성영상을 활용한 공간정보연구가 높은 것을 확인하였으며, Jin 등¹⁷⁾은 기후환경 이슈 분석을 위해 텍스트 마이닝 기법 활용방안을 연구하였다. 논문 자료만을 활용한 기존 연구와는 다르게 뉴스, 환경부 보도자료, 환경 백서 등의 다양한 형태의 자료를 활용하였고, 분석결과 ‘미세먼지’, ‘폭염’, ‘친환경’ 등의 키워드가 증가세를 보이는 것을 확인하였다. 이처럼 텍스트 마이닝 기법은 기존의 축적된 자료를 활용하여 현재의 동향 및 다가올 미래를 준비하기 위한 분석기법으로 널리 활용되고 있다.

1.1. 텍스트 마이닝의 분석 기법

1.1.1. 빈도 분석(Term Frequency)

텍스트 마이닝 기법에서 가장 많이 사용되는 분석법으로 하나의 문서 내에 포함된 단어들의 출현 빈도를 수치화한다. 대상으로 하는 문서의 단어는 기본적으로 띄어쓰기를 기준으로 구분이 되지만 한글의 경우 영어와는 다르게 조사가 공백 없이 결합되어 있는 특징으로 인해 전처리 과정이 매우 중요하다. 또한 문서의 특성을 파악하기 위해서는 분리된 단어의 품사를 지정하는 것이 필요한데, 이러한 과정을 품사 태깅(part-of-speech tagging)이라고 하며 형태소 분석기를 사용하여 분리된 단어에 품사를 지정해 줄 수 있다.¹⁴⁾ 품사가 지정된 단어들은 문서 내에서 출현 빈도수를 계산하여 주요 키워드로 지정이 되기도 하고, 서로 다른 문서간의 유사도 검증을 위해 사용되기도 한다. 출현 빈도에 의해 단어의 중요도가 결정되기 때문에 분석에 띄어쓰기로 구분된 단어 중에 숫자나 오타자 또는 문장부호 등의 의미없는 불용어를 제거하는 것이 필수적이다.

1.1.2. TF-IDF 기법과 K-평균 클러스터링(K-means clustering) 기법을 활용한 문서 분류

TF-IDF (Term Frequency - Inverse Document Frequency)는 여러 문서로 이루어진 자료가 존재할 때 특정 단어가 하나의 문서 내에서 얼마나 중요한지를 나타내는 통계적 수치이다.¹⁸⁾ 하나의 문서에서 발견되는 특정 단어의 수를 의미하는 단어 빈도(Term Frequency; TF)는 반복 출현 횟수가 높은 단어일수록 문서에서 중요한 역할을 한다는 점을 의미한다. 하지만 문서에 포함된 데이터가 많은 경우, 즉 문서의 페이지가 많을수록 빈도수가 커지기 때문에 페이지 수가 다른 문서 간의

비교를 위해서는 이러한 점을 보완하기 위한 방법이 필요하다. 또한 여러 개의 문서가 있을 때 특정단어가 문서군에서 자주 사용될 경우, 흔하게 사용되는 일반적인 용어를 의미하게 됨으로써 문서의 특징을 나타내는 단어라고 보기 어렵게 된다. 이러한 이유로 TF만으로는 문서의 분류에 한계가 존재한다.

TF를 보완하기 위한 개념이 역문서 빈도(Inverse Document Frequency; IDF)이다. IDF는 문서빈도(Document Frequency; DF)의 역수로서 DF가 높을수록 그 단어가 문서군에서 흔하게 사용되는 단어라는 의미를 가진다. 따라서 IDF 값이 높을수록 특정 문서에서만 나타나는 핵심어가 될 가능성이 높아진다. TF-IDF는 TF와 IDF 값의 곱으로 결정되는데 이 값이 높을수록 여러 문서의 집합에서 특정 단어가 얼마나 중요한가를 판단하는 기준이 된다. 따라서 이를 사용하면 특정 중요 단어를 공유하는 문서 군집화가 가능해진다.

K-평균(K-mean) 클러스터링 기법은 여러 개의 문서가 존재할 경우 k개의 군집으로 나누어 분류하는 방식이다. 초기에는 임의로 k개의 중심점을 지정하고 여기서부터 주변의 데이터들 사이의 거리를 계산해서 가장 가까운 데이터들로 둘러싸인 군집을 결정하게 된다. 그 후 결정된 그룹의 데이터 간의 거리를 재계산하여 새로운 중심점을 지정한 후 다시 주변 데이터와 거리 계산을 진행한다. 이렇게 여러번 반복해서 진행한 후에 이전 단계의 군집과 새로 계산된 군집 간의 차이가 없을 경우 알고리즘이 마무리된다.¹⁹⁾

따라서 TF-IDF를 이용해서 문서를 대표하는 주요 핵심단어들의 집합을 만든 후에 K-평균 클러스터링 기법을 이용하면 유사성을 가진 문서끼리 군집을 만들 수 있다.

1.1.3. 연관어 분석(Word Association Analysis)

연관어는 주어진 문서 또는 문장에서 두 개의 단어가 서로 얼마나 연관되어 있는지를 뜻한다. 두 단어가 같은 문서에서 함께 출현하는 횟수를 세는 방법을 동시 출현 기반 연관어 분석(co-word analysis)이라 하는데, 즉 같은 문서에서 함께 나타나는 빈도가 높을수록 강한 연관관계가 성립한다는 가정을 토대로 한다. 단어 두 개의 연관성을 구하고 난 후에는 두 개의 단어 중에서 상대적으로 중요한 단어를 찾기 위해 중심성 계수(Centrality)를 계산한다. 연관성과 중심성 계수를 이용해서 연관어 관계를 시각적으로 구조화하게 되면 데이터의 전체적인 해석이 가능하다.^{20,21)}

본 연구에서는 이러한 텍스트 마이닝 기법을 활용하여 대한환경공학회지에 2000-2019년 사이에 출판된 2,743개의 논문 초록(abstract)과 주제어(keywords)를 대상으로 빈도 분석, 군집분석, 시계열 분석, 연관어 분석 등을 수행하여 최근 20년간의 연구 동향을 분석해 보았다. 연구의 수행 절차는 첫째, 분석 대상 자료 수집, 둘째, 자료의 전처리, 셋째, 빈도분석, 군집분석, 시계열분석, 연관분석 및 시각화 단계로 진행하였다.

2. 실험방법

2.1. 연구 자료

연구의 자료는 최근 20년(2000년-2019년)간의 대한환경공학회지에서 발행된 2,743개의 국내연구논문의 초록과 주제어를 대상으로 한다. 환경공학회 홈페이지(<https://www.kosenv.or.kr/>)에서 연구 논문을 다운로드하였으나, 이 중에서 대부분의 논문은 텍스트가 이미지화되어 있어 텍스트 마이닝을 진행하기 어려운 형태로 되어 있었다. 따라서 OCR (optical character recognition)을 실행하여, 이미지화되어 있는 문자들을 모두 편집이 가능한 형태로 변환하였다. 전처리 진행 후, python을 사용해서 제시된 주제어 중심으로 빈도 분석, 시계열 분석을 하였고 논문 분야별 분류를 위한 군집 분석과 연관어 분석은 논문초록을 사용하였다.

2.2. 텍스트 마이닝 분석

본 연구는 주제어를 사용하여 빈도분석을 실시하였고, 결과를 시각화하기 위하여 단어구름(word cloud)을 사용하였다. 단어구름은 문서에 사용된 키워드의 빈도를 계산하여 시각적으로 표현하는 것을 말하며 빈도수에 따라 글자의 굵기가 변하며, 빈도가 높고 핵심어일수록 그 값에 비례하는 크기로 표현된다.

군집 분석은 논문 초록을 데이터 활용하여 진행하였으며, TF-IDF 기법으로 문서간에 일반적으로 사용되는 단어들을 제거하고, 각 문서군을 특정지을 수 있는 단어들을 선별하였다.²²⁾ 그 후 K-평균 클러스터링 기법을 활용하며 문서를 분류하였다.

시계열분석은 특정 대상의 시간적 변동을 계속적으로 관측하여 얻은 자료를 바탕으로 분석한 것으로, 빈도분석과 같이 논문에 출현한 상대적 중요도를 비교하는 방법이다. 논문 주제어를 바탕으로 20년의 자료를 4개의 구간(2000년-2004년, 2005년-2009년, 2010년-2014년, 2015년-2019년)으로 구분하여 각 기간별 연구분야들의 특성을 분석하고 그래프를 이용해 시각화하였다.

연관어 분석은 논문 주제어를 바탕으로 두 단어가 동시에 출현한 횟수를 분석하였다. 연관분석은 문맥의 연관성을 알아내기 위한 것이기 때문에 빈도분석과 시계열분석과는 다르게 주제어를 사용하지 않고 초록을 사용하여 진행하였다. 초록에서 유의미한 단어들을 추출하기 위해 KoNLPy를 사용하여 명사만 추출하였으며, 명사 중에서도 불필요한 명사들을 제거하기 위해 불용어 리스트를 만들었다. 불용어 리스트는 Ranks NL 웹사이트(<https://www.ranks.nl/stopwords/korean>)에서 제공하는 한국어 불용어 리스트를 기반으로 만들었으며, 본 연구자료에서 발생하는 숫자, 오타자, 문장 부호 등의 불용어를 추가하여 사용하였다. 이후 시각화를 위한 프로그램은 Gephi를 사용하였다.

Table 1. Term frequency analysis using the keywords in the papers from 2000 to 2019.

Rank	Keyword	Freq.	Rank	Keyword	Freq.
1	흡착	97	7	이산화탄소	29
2	중금속	63		탈질	29
3	활성탄	43	8	휘발성유기화합물	28
4	퇴적물	34		응집	28
5	하수슬러지	33	9	오존	26
6	질산화	32		pH	26
			10	생물막	24

3. 결과 및 고찰

3.1. 빈도 분석 결과

Table 1은 대한환경공학회지에 게재되어 있는 2000년-2019년까지의 ‘환경’ 관련 연구 2,743편의 논문 주제어를 기반으로 빈도 분석한 결과이다. 도출된 키워드 13,333개 중 빈도가 높은 순으로 10개를 나타내었으며, **Table 1**에서 알 수 있듯이 ‘환경’ 관련해서는 ‘흡착’이 97의 빈도로 가장 많이 언급되었으며, 그 다음으로 ‘중금속(63)’, ‘활성탄(43)’, ‘퇴적물(34)’, ‘하수슬러지(33)’순으로 나타났다. 이러한 결과는 대한환경공학회지에 수록된 환경관련 연구 분야 중 수질 분야 연구가 많다는 것을 보여주는 결과이다.

다음은 전체 키워드 분석을 기반으로 최근 많이 사용되는 시각화 방법인 워드 클라우드를 사용해 데이터를 시각화해 보았다. 워드 클라우드는 Python을 통해 구현하였으며, 키워드 빈도수 상위 100개를 시각화한 결과는 **Fig. 2**와 같다. **Table 1**에서 제시된 상위 10개의 키워드 외에도 나머지 환경과 관련하여 검출된 키워드들이 모두 표기되어 있다. ‘흡착’이란 단어가 가장 높은 시각화를 보이고, 그 다음으로 ‘중금속’, ‘활성탄’, ‘퇴적물’과 같은 수질분야와 관련된 단어들이 높게 나타났음을 보여준다. 이를 통해서 환경관련 연구 분야 중 수처리 공정에 대한 관심과 중요도가 타 분야에 비해 월등히 높은 것을 직관적으로 판단할 수 있다.

3.2. TF-IDF를 이용한 논문 분류 결과

K-평균 클러스터링에 의해 분류된 각 군집별 논문의 수는 총 2,743개 논문 중에서 군집 1이 369개, 군집 2가 794개, 군집 3이 583개, 군집 4가 812개, 군집 5가 185개의 순서로 군집

Fig. 2. Word cloud by top 100 keywords.

4가 가장 많은 비율로 게재되었음을 나타내고 있다(**Table 2**). 군집에 포함된 핵심 단어 목록을 활용하여 각 군집의 성격을 추론해 보면, 군집 1은 ‘원수’, ‘오존’, ‘응집제’, ‘활성탄’, ‘염소’, ‘응집’, ‘탁도’, ‘정수장’으로 정수처리 분야에 대한 논문으로 구성되어 있음을 확인할 수 있다. 군집 2의 경우 ‘모델’, ‘수질’, ‘예측’, ‘토양’, ‘유입’, ‘오염물질’ 등의 내용으로 특징지어지는데, 이들을 종합해 보면 수질 모델링 분야에 관한 연구 논문 그룹으로 유추할 수 있다. 군집 3의 경우 ‘중금속’, ‘이온’, ‘흡착’, ‘흡착제’, ‘구리’, ‘활성탄’, ‘카드뮴’과 같은 단어들로 구분되어지는 그룹으로 흡착을 이용한 중금속 처리에

Table 2. Document clustering by TF-IDF method.

군집	논문 수	핵심 단어 목록
군집1	369	원수, 오존, 응집제, 활성탄, 염소, 응집, 탁도, 정수장
군집2	794	모델, 수질, 예측, 토양, 유입, 오염물질
군집3	583	중금속, 이온, 흡착, 흡착제, 구리, 활성탄, 카드뮴
군집4	812	질소, 미생물, 슬러지, 유기물, 제거효율, 질산화, 생물학적, 혐기성
군집5	185	촉매, 온도, 산화, 광촉매, 활성, TiO ₂ , CO ₂ , 과산화수소

Fig. 3. Tree map of the percentage by classified research area.

관한 것으로 해석되어 진다. 군집 4의 경우는 ‘질소’, ‘미생물’, ‘슬러지’, ‘유기물’, ‘질산화’, ‘생물학적’, ‘혐기성’으로 특징을 보이며 이 단어들이 생물학적 하수처리 시스템과 연결되어 있음을 파악할 수 있다. 마지막으로 군집 5는 ‘측매’, ‘온도’, ‘산화’, ‘광촉매’, ‘TiO₂’, ‘CO₂’, ‘과산화수소’ 등으로 특징지어지는데, 이 그룹은 촉매를 활용하는 환경공학 분야로 특징지을 수 있다. 20년간 대한환경공학회지 논문에서 각 군집이 차지하는 비율을 그래프로 정리하여 Fig. 3에 나타내었다.

Fig. 3을 보면 지난 20년간 대한환경공학회지 논문이 생물학

적 처리 또는 하수처리 분야와 수질모델링 분야에 많이 집중되어 있음을 알 수 있다. 그 뒤로 중금속 흡착 분야, 정수처리 분야, 환경 촉매 분야로 구성되어 있다. 또한 1, 2, 3위를 차지한 분야를 보면 모두 수질 분야인 것을 확인할 수 있기에 대한환경공학회지가 다양한 환경 전 분야를 아우르는 연구 범위를 가지고 있지만 2000년에서 2019년 동안에는 수질 관련 연구가 집중된 것으로 분석된다.

3.3. 시계열 분석

조사기간(2000년-2019년)을 5년을 주기로 4개 구간으로 나누는 후 가장 빈번하게 검출되는 상위 키워드 7개를 추출하여 시계열을 분석하였다(Fig. 4). 2000년-2004년에는 ‘흡착’, ‘탈질’, ‘중금속’, ‘활성탄’, ‘이산화탄소’, ‘질산화’, ‘생분해’의 키워드 빈도가 높았다. ‘흡착’은 ‘활성탄’과 연관성이 높은 단어이고, ‘탈질’, ‘질산화’, ‘생분해’의 경우는 하수처리와 연관성이 높은 단어이다. 2005년-2009년의 상위 7개 키워드는 ‘흡착’, ‘중금속’, ‘질산화’, ‘하수슬러지’, ‘활성탄’, ‘산화’, ‘퇴적물’로써 이전 5년간의 키워드와 큰 차이는 없었다. 2010년-2014년 사이에는 ‘중금속’, ‘흡착’, ‘퇴적물’, ‘활성탄’, ‘휘발성 유기화합물’, ‘응집’, ‘미생물연료전지’가 높은 비율을 차지했으며, 이때는 이전 자료와는 다르게 ‘중금속’과 연관된 연구의 양이 ‘흡착’관련 보다 더 많았다. 마지막으로 2015년-2019년에는 ‘흡착’, ‘생태독성’, ‘응집’, ‘세균’, ‘중금속’, ‘미세먼지’, ‘제올라이트’가 상위 7개 키워드로 분석되었고, ‘생태독성’, ‘세균’, ‘미세먼지’, ‘제올라이트’ 키워드가 새롭게 나타났다.

Fig. 4에 기간별 반복해서 상위 빈출 빈도를 가진 키워드를

Fig. 4. Top 7 keywords by period. (a) 2000-2004 years, (b) 2005-2009 years, (c) 2010-2014 years, (d) 2015-2019 years.

다른 색상으로 표시하였다. ‘흡착’의 경우 20년 동안 꾸준히 빈번하게 분석되는 단어이고 이를 통해 볼 때 흡착과 관련된 다수의 연구가 꾸준히 진행되어 왔음을 파악할 수 있다. ‘흡착’과 비슷한 패턴으로 ‘중금속’ 역시 많이 검출되는 단어로서 이는 ‘중금속-흡착’ 또는 ‘활성탄-흡착’과 같은 연관성을 가지고 연구되어진 것으로 판단된다. 각 단어들의 연관성에 관한 추가 분석은 다음 섹션 3.4 연관어 분석에서 진행되었기에 자세한 추가 설명은 3.4부터 서술하도록 하겠다. ‘질산화’의 경우는 2000년부터 2010년까지 높은 빈도로 검출되었지만 최근 10년(2011년-2019년)에는 빈도수가 많이 줄어들었다. 반면에 ‘응집’이 최근 10년간 상위 빈출 빈도에 포함되었다. 마지막으로 가장 최근 5년간 분석 결과에는 기존에 존재하지 않던 키워드들이 많이 검출되었는데 ‘미세먼지’, ‘세슘’, ‘제올라이트’, ‘생태독성’ 등으로 최근 사회문제로 대두

되고 있는 환경 이슈들을 반영한 결과라 판단된다.

3.4. 연관 분석

단어 출현 빈도만을 사용하여 결과를 해석할 때의 한계점은 하나의 단어만으로는 게재된 논문이 주로 어떤 분야로 연구가 진행되었는지 알 수 없다는 점이다. 이러한 단점을 보완하기 위해서 연관어 분석을 통해서 각 단어들이 어떠한 관점으로 연구되었는지 확인해 보았다. 연관어 분석에 중심어로 사용된 단어는 총 3가지(흡착, 활성탄, 중금속)로써 빈도 분석 및 시계열 분석에서 빈번하게 검출된 키워드를 선정하였다. 네트워크 분석을 통한 시각화 결과는 **Fig. 5**에 나타내었다. **Fig. 5(a)**는 연관어 분석으로 확인된 모든 단어들을 표기한 것으로 다양한 단어들이 복잡한 형태로 연결되어 있는 것을 파악할 수 있다. 이처럼 연관어 분석을 진행하면 각 문서에서

Fig. 5. Associated term network with (a) total keywords, (b) ‘adsorption’, (c) ‘activated carbon’, and (d) ‘heavy metals’.

존재하는 연결성이 높은 단어들이 우선적으로 파악되고, 파악되어진 한 쌍의 단어, 즉 동시출현 빈도가 높은 두 개의 단어는 다른 문서에서 나온 한 쌍의 단어와 비교되면서 공통으로 사용되는 단어에 연결선이 추가되게 되는 형식이다. **Fig. 5(b)**는 ‘흡착’을 중심으로 하는 연관어 네트워크 분석 결과로써 ‘흡착’과 연관성이 높은 단어들과의 연결성을 보여준다. ‘흡착’은 ‘활성탄’, ‘중금속’, ‘유기물질’, ‘미생물’, ‘폐수’ 등과 연결이 높게 분석되었으며, 이를 확인할 때 3.3에서 언급했던 ‘흡착-중금속’의 연관성이 증명된다. 또한 다양한 단어와의 연결성은 ‘흡착’이라는 단어의 연구 영역이 매우 광범위하다는 것을 알 수 있게 한다. **Fig. 5(c)**의 경우는 ‘활성탄’을 중심으로 하는 연관성 분석으로 ‘활성탄-흡착’과의 연관성과 ‘유기물질’, ‘미생물’ 등을 제거하는 연구 내용과 ‘온도’, ‘활성’, ‘흡착특성’ 등의 단어를 볼 때 ‘활성탄’ 효율 증가를 위한 연구들이 많이 진행되고 있음이 확인되었다. 마지막으로 ‘중금속’을 중심으로 하는 연관어 분석(**Fig. 5(d)**)에는 ‘흡착’, ‘토양’, ‘폐수’, ‘퇴적물’ 등의 단어와 연관성을 찾을 수 있었으며 다양한 환경매체에서 중금속 제거를 위한 흡착을 연구 주제로 하는 논문이 게재되었음을 확인하였다.

4. 결론

2000년부터 2019년까지 대한환경공학회지에 게재된 논문에 대해서 텍스트 마이닝 기법을 통한 연구동향을 파악하였다. 자료에 대한 분석으로 추출된 주요 키워드를 활용하는 빈도 분석과 워드클라우드 분석, TF-IDF 분석을 통한 연구분야별 논문 분류, 시계열 분석을 통한 기간별 주요 키워드 확인, 마지막으로 연관어 분석을 통한 키워드별 연구분야 확인을 진행하였고, 다음과 같은 결과를 얻었다.

- 1) 2000년부터 2019년 동안 게재된 논문에서 출현빈도가 높은 단어를 확인하였을 때, ‘흡착’, ‘중금속’, ‘활성탄’, ‘퇴적물’, ‘하수슬러지’ 순서로 분석되었고, 이를 확인하였을 때 수질분야에 중점적으로 연구가 집중되었음을 알 수 있다.
- 2) TF-IDF를 통해 자료들의 분류를 진행하였고, 그 결과 총 5개의 군집(정수처리분야, 수질 모델링 분야, 중금속 흡착 분야, 생물학적 하수처리 분야, 환경 촉매 분야)으로 분류되었다.
- 3) 자료를 5년씩 총 4개의 구간(2000년-2004년, 2005년-2009년, 2010년-2014년, 2015년-2019년)으로 나누어 출현빈도가 높은 단어들을 추출한 결과, ‘흡착’과 ‘중금속’은 모든 구간에서 상위권으로 검출되는 단어였고, 최근 5년(2015년-2019년)에서는 사회문제를 반영하는 키워드인 ‘미세먼지’, ‘세습’, ‘생태독성’이 확인되었다.
- 4) 연관어 분석을 통해 네트워크 시각화를 진행한 결과, ‘흡착’은 ‘탈질’, ‘활성탄’, ‘중금속’, ‘유기물질’, ‘미생물’ 등과 매우 연관성이 높은 단어였고, ‘활성탄’은 ‘흡착’, ‘유기물질’,

‘흡착특성’ 분야와 연결성이 높았다. 마지막으로 ‘중금속’과 연관성이 높은 단어는 ‘흡착’, ‘토양’, ‘폐수’, ‘퇴적물’ 등으로 확인되었다.

텍스트 마이닝 기법은 대용량의 문서 데이터를 바탕으로 이루어지는 분석으로 우선적으로 체계적으로 정리된 문서 자료를 전제로 한다. 지금까지 국내에 축적되어 있는 다양한 형태의 제안서, 보고서, 논문 자료 등을 활용하여 텍스트 마이닝 기법을 확대 적용한다면 의미있는 결과들을 얻을 수 있을 것이라 전망한다.

Acknowledgement

이 논문은 2019~2020년도 창원대학교 자율연구과제 연구비 지원으로 수행된 연구결과임.

References

1. S.-S. Baek, Y. Choi, J. Jeon, J. C. Pyo, J. Park, K. H. Cho, Replacing the internal standard to estimate micropollutants using deep and machine learning, *Water Res.*, 188, 116535 (2021).
2. K. Yetilmezsoy, B. Ozkaya, M. Cakmakci, Artificial intelligence-based prediction models for environmental engineering, *Neural Network World*, 21(3), 193-218(2011).
3. S. H. Chen, A. J. Jakeman, J. P. Norton, Artificial intelligence techniques: an introduction to their use for modelling environmental systems, *Math. Comput. Simul.*, 78(2-3), 379-400(2008).
4. H. Kim, J. Lee, Y. Kim, Development of state diagnosis algorithms using statistical method for online sensors in A²/O process, *J. Korean Soc. Environ. Eng.*, 41(11), 627-636(2019).
5. J. Kim, N.-S. Park, S. Yun, S.-H. Chae, S. Yoon, Application of isolation forest technique for outlier detection in water quality data, *J. Korean Soc. Environ. Eng.*, 40(12), 473-480(2018).
6. J. Lee, H. Ha, M. Lee, M. Lee, T. Kim, Y. Cha, J. Koo, Assessment of water quality of major tributaries in Seoul using water quality index and cluster analysis, *J. Korean Soc. Environ. Eng.*, 42(10), 452-462(2020).
7. J. Park, H. Mun, M.-J. Park, H. Jang, D.-W. Jeong, Bioethanol production using microalgae, *J. Korean Soc. Environ. Eng.*, 42(3), 164-176(2020).
8. J. I. Dong, Greenhouse gas recycling & utilizing technology for the district heating business in the era of climate change, *J. Korean Soc. Environ. Eng.*, 30(12), 1203-1206(2008).
9. W.-J. Choi, K.-M. Park, B.-G. Yoon, M.-C. Kim, K.-J. Oh, Recovery of presource from sewage sludge by a struvite-forming method, *J. Korean Soc. Environ. Eng.*, 31(7), 557-564(2009).
10. C. Choi, C.-M. Kim, J. Lim, D. Kim, I. S. Kim, Economic assessment based on energy consumption on the capacities

- in seawater reverse osmosis (SWRO) plant in Korea, J. Korean Soc. Environ. Eng., 41(7), 389-398(2019).
11. M. A. Hearst, Untangling text data mining, in Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp. 3-10(1999).
 12. J.-H. Ryu, Y.-Y. You, The fourth industrial revolution core technology association analysis using text mining, J. Digit. Converg., 16(8), 129-136(2018).
 13. A. Tan, Text mining: the state of the art and the challenges, in Proceedings of the PAKDD1999 Workshop on Knowledge Discovery from Advanced Databases, Citeseer, pp. 65-70(1999).
 14. H. Moon, W. Son, J. H. Won, Classification of documents related to economic psychology using text mining techniques, Quarterly National Account Review, 4(2017).
 15. B. Park, K. Oh, J. Lee, J. Yoon, S. K. Lee, M. Lee, A study on environmental research trends by information and communications technologies using text-mining technology, Korean J. Remote. Sens., 33(2), 189-199(2017).
 16. K. Oh, M. Lee, B. Park, J. Lee, J. Yoon, Analysis of the research trends by environmental spatial-information using text-mining technology, J. of the Korean Assoc. of Geogr. Inf. Stud., 20(1), 113-126(2017).
 17. D. Jin, S. Kang, H. Choi, K. Han, D. Kim, A study on the utilization of text mining for climate environmental issues analysis, Climate Environment Policy Research, 2018, 1-152(2018).
 18. M. Lee, H. Bae, Design of keyword extraction system using TFIDF, Korean Journal of Cognitive Science, 13(1), 1-11 (2002).
 19. D. H. Shin, K. K. An, S. C. Choi, H. Choi, Malicious traffic detection using K-means, J. Korean Inst. Commun. Inf. Sci., 41(2), 277-284(2016).
 20. H. Li, H. An, Y. Wang, J. Huang, X. Gao, Evolutionary features of academic articles co-keyword network and keywords co-occurrence network: based on two-mode affiliation network, Phys. A, 450, 657-669(2016).
 21. B. Kim, M. Kim, Language network analysis of the cruise

tourism issues, Tourism Research, 40(4), 1-20(2015).

22. J. Lee, M. Lee, J. Kim, A study on Korean language processing using TF-IDF, The Journal of Information Systems, 28(3), 105-121(2019).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors and Contribution Statement

Donghyeon Lee

School of Civil, Environmental and Chemical Engineering, Changwon National University, Bachelor-Master Combined Course Student, ORCID[®] 0000-0002-1821-4619: Writing - original draft, Data analysis, Data curation.

Sojeong Lee

Department of Smart Ocean Environmental Energy, Changwon National University, Master Student, ORCID[®] 0000-0002-8287-2034: Data analysis, Data curation.

Dokyeong Lim

Department of Smart Ocean Environmental Energy, Changwon National University, Master Student, ORCID[®] 0000-0003-3951-8491: Data analysis, Data curation.

Jongkwan Park

School of Civil, Environmental and Chemical Engineering, Changwon National University, Department of Smart Ocean Environmental Energy, Changwon National University, Professor, ORCID[®] 0000-0002-8308-7644: Conceptualization, Data analysis, Methodology, Supervision, Writing - review and editing.