Journal of the Korea Academia-Industrial cooperation Society
Vol. 23, No. 2 pp. 203-212, 2022

https://doi.org/10.5762/KAIS.2022.23.2.203
ISSN 1975-4701 / eISSN 2288-4688

국가하천 점용 콘텐츠에 대한 연관어 분석
-전라도 지역을 중심으로-

정성윤*, 김남곤
한국건설기술연구원 미래스마트건설연구본부

Analysis of Associated Words for the Occupancy Contents of National Rivers
-Focusing on the Jeonla-do Region-

Seong-Yun Jeong*, Nam-Gon Kim
Department of Feture & Smart Construction Research,
The Korea Institute of Civil Engineering and Building Technology

요 약 과거에는 하천구역에서의 개발을 통해 공적 또는 사전 이익을 위한 점용허가가 많았다. 최근에서는 하천구역에서의 난 개발을 억제하여 환경 보존을 위한 국민적 관심이 높아지고 있다. 본 연구는 허가관리청에서 관리하는 허가대장 내 점용 콘텐츠에 내재된 유의미한 정보를 유추하여 하천점용의 시사점을 제시하고자 하였다. 정보 유추를 위해 전라도에 위치한 5개의 국가하천 구역에서 허가된 1,768건의 점용 콘텐츠를 구성한 단어들의 출현 빈도를 분석하였다. 계속해서, 군집화와 동시 출현기반의 연관어 및 주제 등을 분석하였다. 분석결과를 토대로 연관어 연결망을 형성하였다. 연결망에서 점용 목적보다는 개발 위주의 행위에 해당하는 단어들이 영향력이 있는 것으로 파악되었다. 이러한 분석결과는 하천 점용의 행위에 대한 인식을 제고할 수 있는 기초자료로 활용될 수 있을 것으로 사료된다. 끝으로, 본 연구에서는 가장 보편적으로 사용되는 텍스트 마이닝 기반의 분석 기법과 절차를 준용하여 연관어의 관계를 분석하였다. 보다 객관적이고, 신뢰성 있는 연구 결과를 얻기 위해서는 본 연구에 적합한 분석 이론과 검증 방법에 대한 연구가 추가적으로 요구된다.

Abstract In the past, there were many occupancy permits for profit-through-development in river areas. In addition, public interest in preserving the environment of river zones has been increasing in recent years. This study suggests implications for river occupancy by inferring meaningful information inherent in the contents for occupancy. The frequency of occurrence of words constituting 1,768 occupancy contents used in five national rivers located in Jeonla-do was analyzed. The clustering, associated words based on co-occurrence, and topics and network of associated words were analyzed. Words corresponding to development-oriented actions rather than occupational purposes were identified as having influence in the network. Hence, it is thought that the results of this analysis can be used as basic data to raise awareness of the behavior of river occupancy. Finally, the relationship between related words was analyzed by applying the most commonly used text mining-based analysis techniques and procedures in this analysis. To obtain more objective and reliable research results, additional research on analysis theory and verification methods suitable for this study is required.

Keywords : National River Occupancy, Text Mining, Associated Words Analysis, Topic Model, Permit Contents

이 논문은 교육과학기술부의 재원으로 한국건설기술연구원 "(22주요_대1_목적)미래 신수요 대응형 스마트건설 융복합 핵심기술 개발 연구(2/2)" 과제와 국토교통부 출연사업인 "22 건설사업정보시스템 운영 및 기능개선"의 지원을 받아 수행되었음.
*Corresponding Author : Seong-Yun Jeong(KICT)
email: syjeong@kict.re.kr
Received October 15, 2021 Revised November 23, 2021
Accepted February 4, 2022 Published February 28, 2022

1. 서론

1.1 연구의 필요성 및 목적

하천법에서는 하천을 지표면에 내린 빗물 등이 모여 흐르는 물길로서 공공의 이해와 밀접한 관계가 있다고 정의하고 있다. 이러한 하천구역에서 점용 행위를 하려는 자는 허가관리청으로부터 허가를 받아야 한다. 과거에서 하천구역의 개발을 위해 점용허가를 내어주었다. 허가대장에 수록되는 항목들 중 점용 콘텐츠는 비정형 텍스트로 기록되고 있다. 점용 콘텐츠에는 유의미 정보를 내포하고 있지만 단순히 기록 수준으로 관리되고 있다. 단순 기록용으로 사용하다보니, 허가대장에서 하천구역의 개발과 점용허가와의 어떤 관계가 있는지를 파악하기 위해 점용허가와 관련한 주제가 무엇이고, 주제에서 유사한 의미를 갖은 주제 단어가 어떤 것이 있는지, 주제별로 주제 단어들 간에 어떤 연관 관계를 가지고 있는지에 대한 연구가 전무한 실정이다. 본 연구의 목적은 텍스트 마이닝을 적용하여 하천점용허가에 내재된 핵심 주제와 주제마다 주제 단어들 간의 연관 관계를 파악하고, 시간의 흐름에 따라 주제 변화 추이를 살펴봄으로써 향후 국가하천 점용허가의 방향성을 모색하고자 한다. 더불어, 점용 콘텐츠에 내포된 유의미 정보를 유추한다면 하천 점용 정책 대안의 이론적 근거를 제시할 수 있는 기초자료로 활용할 수 있을 것으로 판단하였다. 이를 위해 텍스트 마이닝 기법을 사용하여 비정형 형태의 점용 콘텐츠에 내포된 유의미적인 단어들 간의 연관 관계를 분석하였다. 연관어 분석을 통해 어떤 단어들이 점용 콘텐츠에서 핵심적인지를 파악함으로서 현재까지의 하천 점용의 목적 또는 행위에 대한 시사점을 유추할 수 있을 것이다.

1.2 연구 범위 및 구성

대부분의 허가관리청에서는 점용허가정보를 대장으로 기록하고 있다. 허가관리청에 따라 허가대장으로 관리하는 항목과 수준에 차이가 있다. 따라서 전국을 대상으로 점용 콘텐츠를 분석하기보다는 특정 허가관리청에서 관리하는 점용 콘텐츠를 분석하는 것이 보다 현실적이라 판단하였다. 그래서 1975년부터 2021년 8월까지 전라도에 위치한 섬진강, 영산강, 만경강, 동진강 및 탐진강 구역에서 점용된 허가대장을 수집하였다. 허가대장 항목들 중 하천명과 허가년도 및 점용 목적과 개요를 취합한 점용 콘텐츠를 정리하여 분석 자료로 사용하였다.

다음으로, 본 연구 내용을 다음과 같이 구성하였다. 제1장에서는 서론을 기술하였고, 제2장에서는 국가하천의 점용허가 현황과 선행연구 사례 및 텍스트 마이닝 분석 절차를 살펴보았다. 제3장에서는 점용 콘텐츠의 전처리, 단어출현 빈도, 군집화 및 주제 모델링 등을 통해 서로 연관되는 단어들을 분석하였다. 끝으로, 본 연구 결과를 정리하여 시사점과 연구의 한계 등 결론으로 제시하였다.

2. 이론적 고찰

2.1 국가하천 점용허가 현황

하천법 33조(하천의 점용허가)에 따라 하천구역 안에서 토지의 점용, 하천시설의 점용 등 6개의 조항 중 하나 이상의 행위를 하려는 자는 하천관리청의 허가를 받아야 한다. 하천 점용은 하천의 일부에 대한 권리를 가지고 유형적・고정적으로 점유하여 사용하는 것을 의미한다[1]. 국가하천구역에서의 점용 목적 및 행위에 따라 지방국토관리청과 지방자치단체가 허가관리청으로 지정되어 있다. 지방국토관리청에서는 하천시설의 점용, 하구둑・다목적댐의 공작물의 신축・개축・변경, 토질의 굴착・성토・절토・그 밖의 토질의 형질 변경 등에 대해 점용을 허가한다. 지방자치단체는 토지의 점용 토석・모래・자갈의 채취, 그 밖의 산출물의 채취 등 점용을 허가하고 있다. 2020년 말 기준으로, 지방국토관리청에서는 14,010건 점용 허가 건을 대장으로 관리하고 있다. 2021년 5월 기준으로, 55개의 지방자치단체는 3,070건의 허가를 대장으로 관리하는 것으로 파악되었다. 본 연구에서는 영산강, 섬진강, 만경강, 동진강, 탐진강 구역에서 지방국토관리청이 허가한 1,768건의 점용 콘텐츠를 수집하였다. Table 1은 수집한 허가 건을 하천별로 분류한 것이다.

Table 1. Number of occupancy contents collected by rivers

Division	70's	80's	90's	00's	10's	20's	sum
Yeongsan Riv.	1	21	163	171	355	47	758
Seomjin Riv.	0	11	65	89	158	28	351
Mangyeong Riv.	2	36	108	93	98	26	363
Dongjin Riv.	1	26	47	55	65	3	197
Tamjin Riv.	0	2	22	31	40	4	99
sum	4	96	405	439	716	108	1,768

2.2 선행 연구사례

본 연구의 성격은 점용 콘텐츠를 분석 대상으로 하기 때문에 이론보다는 응용에 가깝다. 또한, 국내에만 적용되는 특성을 고려하여 국회도서관과 네이버의 학술정보를 중심으로 선행 연구사례를 조사하였다. 조사 결과로는 주로 하천점용 허가제도의 개선[1,2]와 하천점용 업무의 효율화 방안[3]에 대한 연구가 있었다. 이들 연구는 점용정보 관리방안에 관한 내용을 포함하고 있지만 점용 콘텐츠 활용에 대한 연구 내용은 전무한 것으로 파악되었다. 더불어, 점용과 관련한 텍스트 마이닝 또는 연관어 분석에 관한 연구사례도 없는 것으로 판단되었다. 다만, 하천부문에서 텍스트 마이닝을 적용한 연구 사례로는 Do[4], Kim[5] 및 Kim[6]이 있는 것으로 조사되었다. Do[4]는 국내 담수외래종에 대한 논문 초록을 대상으로 출현 빈도수, 주제어 간의 연관 관계 분석 및 연결망 시각화를 통해 연구 동향을 살펴보았다. Kim[5]는 수돗물의 사회적 인식과 식수 사용 유형을 파악하기 위해 상수도 수질에 관련된 주제 모델링과 네트워크 관계를 분석하였다. Kim[6]은 부산시 민원에 대해 TF-IDF (Term Frequency-Inverse Document Frequency 이하 TF-IDF)기반의 출현 빈도, 동시 출현 단어 및 단어 연관성 등의 분석을 통해 연관어 관계를 유추하였다. 이들 연구는 새로운 텍스트 마이닝 이론 또는 분석 알고리즘을 제시하기보다는 콘텐츠에 내재된 데이터 간의 연관 관계를 분석하기 위해 텍스트 마이닝을 적용한 것으로 생각된다.

2.3 텍스트 마이닝 개념

콘텐츠는 비구조화, 비정형된 텍스트로 구성되어 있다. 비정형된 텍스트는 단어와 단어 간에 연관 관계를 가지고, 연관 관계 속에서의 맥락 속에는 의미 있는 정보를 내포할 수 있다. 유의미한 정보를 파악하기 위한 기법들 중 대표적으로 텍스트 마이닝(text mining)을 사용할 수 있다. 비정형 텍스트로부터 의미 있는 연관된 정보를 추출하는 분석 기법이다. 유의미 정보를 찾기 위해 텍스트를 전처리(pre- processing)하고, 출현 단어들을 구조화, 색인화, 수치화한 후에 정보 분류, 요약 등 일련의 분석 과정을 거친다[7]. 텍스트 마이닝은 머신러닝, 자연어 처리, 챗봇 등의 기술로 사용되면서 동향 분석[4], 예측 분석[8], 정보 분석[9] 등 여러 분야에서 응용되었다.

2.4 연관어 분석 절차

본 연구는 점용 콘텐츠에 내재된 유의미한 정보 분석을 위해 선행된 텍스트 마이닝 분석 절차[4,10,11]을 준용하여 다음과 같은 흐름으로 연구를 진행하였다.

2.4.1 전처리

점용 콘텐츠를 구성하는 단어들을 토큰화(Tokenize)하였다. 토큰 단위로 공백, 특수문자, 문장부호, 숫자, 영문자, 1음절 및 불용어 등을 제거하였다. 제거하여 남은 토큰을 대상으로 KoNLPy 라이브러리를 사용하여 명사 단어만을 추출하였다.

2.4.2 단어출현 빈도 분석

전 처리된 단어들이 콘텐츠에서 얼마나 자주 출현하는지를 나타내는 단순 단어출현 빈도(TF)를 분석하였다. TF는 단순히 단어출현 빈도만을 측정하였기 때문에 중요하지 않은 단어가 빈번하게 출현할 경우에 의도되지 않는 결과가 나올 수 있다. 이를 보완하기 위해 TF와 역문서 빈도(IDF)를 곱한 TF-IDF를 산출하였다[12]. TF-IDF 분석을 통해 점용 콘텐츠에서 특정 단어가 얼마나 중요한지를 살펴보았다.

2.4.3 군집 분석

TF-IDF 분석만으로는 콘텐츠에 내재된 의미 있는 단어들을 찾는 것은 쉽지 않다. 이를 위해서는 서로 비슷한 의미를 갖는 단어들을 묶은 군집화(clustering)를 분석할 수 있다. 군집 분석은 유사한 특성을 가진 단어들을 분류하여 묶어주는 기법이다. 유사한 특성을 파악하기 위해 동시에 출현한 단어의 빈도를 산출하여 두 단어 간의 거리를 기반으로 유사도 값을 측정하였다. 유사도는 콘텐츠를 구성하는 단어들 간의 유사한 정도를 정량적으로 측정하는 것을 말한다. 유사도가 클수록 단어들 간의 거리가 가까워지고, 반대로 유사도가 낮을수록 거리는 멀어진다. 군집화를 위해 k-평균 군집 모델을 사용하였다. 이 군집 모델은 구현이 간단하고 해석이 쉽기 때문에 군집화에 많이 사용하는 방식이다. 사전에 군집화 할 개수(k)를 설정하였고, 지정한 k개의 군집 개수에 따라 임시 지점을 잡았다. 가장 가까운 군집 중심선 간의 거리평균 지점에 대해 반복적으로 군집의 중심을 계산하였다. 하지만 이 방법은 주관적으로 k를 설정해야 하고, 중심점을 어디로 지정하느냐에 따라 군집 위치가 달라질 수 있다[13].

2.4.4 동시 출현기반 빈도 분석

출현 단어의 순서와 단어들 간의 상호 의존성을 측정하여 단어들 간의 연관 관계를 파악하기 위해 동시 출현 기반 빈도를 분석하였다. 이 분석은 콘텐츠에서 특정 단어와 비교대상 단어를 연속으로 출현한 횟수를 계산한 후에 출현 빈도가 높은 단어들이 연결 강도가 높다고 평가하였다. 이 분석을 통해 연관어 연결망을 형성할 수 있다[14]. 연결망 내에 구성되는 단어들 간의 유의미한 관계성을 분석할 수 있다.

2.4.5 주제 모델 분석

앞에서 진행한 군집 분석은 하나의 콘텐츠가 여러 주제(topic)를 포함하는 것에 대해 충분히 고려하지 못한다. 좀 더 면밀한 분석을 위해서는 모든 콘텐츠에 포함된 다수의 주제를 파악하는 방법으로 주제 모델을 사용할 수 있다[15]. 주제 모델링을 통해 점용 콘텐츠가 어떤 주제를 다루고 있는지를 살펴보았다. 주제 모델링을 위해 잠재 다리클레 할당(Latent Dirichlet Allocation, 이하 LDA) 확률모형을 많이 사용한다[4]. LDA를 적용하여 모든 콘텐츠를 취합하여 말뭉치(corpus)를 구성하였고, 말뭉치에 잠재된 의미 있는 주제를 찾았다. LDA는 사전에 주제 개수(k)를 정해야 한다. 이 때, 적합한 주제 개수를 추정하기 위해 혼란도(perplexity)와 주제일관성(topic coherence)을 사용하였다. 혼란도 계산을 위해 콘텐츠에서 주제 출현 확률과 주제별로 단어출현 확률을 평가하였다. 주제 개수를 점차 증가시키면서 어느 시점에서 더 이상 혼란도가 감소하지 않는 지점을 최적의 주제 개수로 보았다. 하지만 혼란도가 낮다고 해서 반드시 적절한 주제 개수라고 평가할 수는 없다[16]. 이를 보완하기 위해 주제일관성을 측정할 수 있다. 주제일관성은 출현 빈도가 높은 n개의 단어들 간의 유사도와 유사도의 평균을 계산하여 해당 주제가 유의미한 단어들로 묶었는지를 평가하였다. 주제일관성이 높았다가 낮아지는 지점을 적정한 주제 개수로 결정하였다.

2.4.6 연관어 연결망 분석

주제별로 도출한 단어들 간의 영향력을 설명하기 위해 연관어 연결망을 사용하였다. 이 연결망은 전체 단어들 간의 연결 관계를 분석하기 위해 특정 단어가 가지는 영향력을 측정하기 위해 동시 출현기반의 연관어 분석, 유사도, 중심성 계수를 사용한다[17,18]. 이들 중 중심성 계수는 특정 단어가 중심에 위치하는 정도를 보여주는 척도이다. 중심성 계수는 대표적으로 연결 중심성(degree centrality), 근접 중심성(closeness centrality), 매개 중심성(between centrality), 고유벡터 중심성(Eigenvector centrality)을 사용한다[19]. 이들 중심성 계수는 분석 대상에 따라 적합한 계수를 선택하여 사용해야 한다. 연결선이 많은 단어가 중심성 계수가 높게 측정되고, 그만큼 해당 단어는 영향력이 높다고 해석할 수 있다.

3. 분석 결과

3.1 분석 절차

본 연구는 선행연구[4,10,11]에서 진행하였던 텍스트 마이닝을 적용한 분석 절차를 준용하여 Fig. 1과 같은 절차로 허가대장의 점용 콘텐츠에 대한 연관어를 분석하였다.

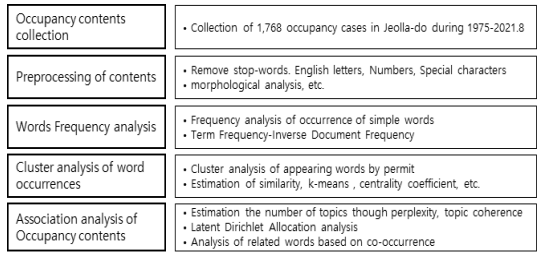


Fig. 1. Analysis procedure based on text mining

가정 먼저, 허가관청에서 관리하는 1975년부터 2021년 8월까지의 허가대장 자료를 수집하였고, 점용 콘텐츠를 정리하였다. 두 번째로 점용 콘텐츠를 구성하는 단어들에 대한 전처리 과정을 거쳤다. 계속해서, 단어의 출현 빈도를 분석하였고, 코사인 유사도가 비슷한 단어들을 하나로 묶는 군집화를 분석하였다. 다음으로, 주제일관성과 혼란도 척도를 측정하여 점용 콘텐츠에 적합한 주제 개수를 파악하고, LDA 모형을 이용하여 3개의 주제와 주제별로 주제 단어를 분석하였다. 끝으로, 점용허가 시기별로 하천별 주제 변화 추이를 살펴보았다. 이러한 분석에는 텍스트 마이닝 분석에 널리 사용되는 Python 3.9.8 버전의 오픈소스 프로그램을 사용하였다.

3.2 점용 텍스트의 전처리

점용 콘텐츠에서 사용된 단어의 출현 빈도를 분석하기에 앞서 점용 콘텐츠 중 703건이 중복된 것으로 파악되었다. 중복된 콘텐츠를 제거한 1,065건을 대상으로 공백, 특수문자, 문장부호, 숫자, 영문자, 1음절, 불용어 등

을 제거하였다. 형태소 분석과정을 통해 명사에 해당하는 3,776개의 단어를 추출하였다. 추출한 단어들 중 중복된 단어를 제거하여 964개의 고유 단어를 얻었다. 개별 점용 콘텐츠에서 가장 많이 사용된 단어의 수가 15개이고, 콘텐츠 당 평균 3.54개의 단어가 사용되었다.

3.3 단어출현 빈도 분석

3,776개의 단어를 가지고서 단어출현 빈도수(TF)를 계산하였다. Table 2는 TF와 TF-IDF 분석을 통해 상위 6위의 단어출현의 빈도와 비율을 나타낸 것이다.

Table 2. TF-IDF calculation result

Division	Top 6 words of TF/TF-IDF(%)
Total Riv. (TF)	Installation(13.1), Facility(3.2), Pumping station(2.7), Buried(2.5), Pipeline(2.3)
Total Riv. (TF-IDF)	Installation(5.2), Facility(2.0), Pumping station(1.8), Buried(1.7), Pipeline(1.6), Drainage(1.5)
Yeongsan Riv. (TF-IDF)	Installation(5.7), Pumping station(2.5), Facility(2.1), Pipeline(2.0), Buried(2.0), Drainage(1.7)
Seomjin Riv. (TF-IDF)	Installation(5.8), Build(2.1), Facility(1.9), Pumping station(1.9), Pipeline(1.8), Buri-ed(1.7)
Mangyeong Riv. (TF-IDF)	Installation(5.5), Facility(3.0), Buried(2.9), Drainage(2.2), Pipeline(1.9), Road(1.9)
Dongjin Riv. (TF-IDF)	Installation(7.1), Drainage(3.2), Buried(3.1), Facility(2.9), Pavement(2.4), Pipeline(2.2)
Tamjin Riv. (TF-IDF)	Installation(8.0), Buried(3.8), Pipeline(2.4), Facility(2.4), Pumping station(2.4), Sewer pipeline(2.1)

Installation이 가장 출현 횟수가 많았고, 전체 단어들 중 약 5.2%를 차지하였다. 한편, 특정 단어가 점용 콘텐츠에서 얼마나 중요한지를 나타내는 TF-IDF 분석 결과로도 Installation이 가장 큰 수치로 나왔으며, 전체에서 약 14.6%를 차지하였다. 나머지 단어들도 순위의 차이가 있으나 거의 유사한 빈도 비율을 가졌다. TF-IDF 값이 크다는 것은 특정 단어가 점용 콘텐츠에서 중요하다고 해석할 수 있다. 다음으로, Fig. 2와 같이 전체 점용 콘텐츠에서 출현단어들의 빈도 누적 비율을 살펴보았다. 상위 10%를 차지하는 단어들에 대한 출현 빈도의 누적 비율이 전체에서 약 50%를 차지하였다. 이처럼 일부 단어들은 출현 빈도가 높았고, 나머지 단어들은 출현 빈도가 낮은 긴 꼬리 형태로 분포되는 지프의 법칙(Zif's law)[20]을 갖는 것으로 알 수 있었다.

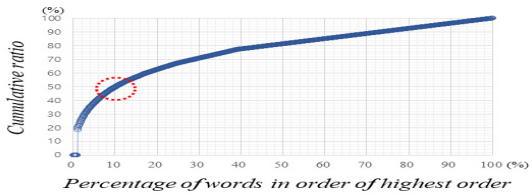


Fig. 2. Cumulative ratio of word occurrences

3.4 군집 분석

3.4.1 군집 개수 추정

TF-IDF와 동시 출현 빈도를 측정한 결과를 토대로 연관되는 단어들을 하나로 묶은 군집화를 수행하였다. 점용 콘텐츠에 적합한 군집화의 개수를 선택하기 위해 Fig. 3과 같이 군집 개수를 2부터 10개까지 부여하면서 왜곡된 정도의 변화를 추정하였다. 2에서 4까지 왜곡 값이 변화하기 시작하였다. 이를 근간하여 군집 개수(k)를 3개로 설정하였다.

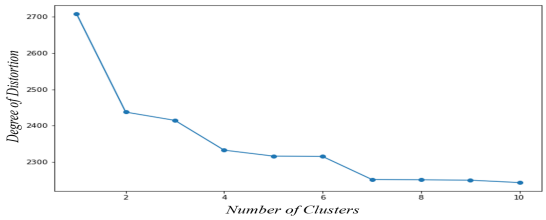


Fig. 3. Estimating the number of clusters

3.4.2 군집화

Fig. 4는 k를 3개로 설정했을 때 K-평균 군집 모델링을 통해 얻은 군집화 결과이다. Fig. 4에서 보면 1번 군집과 3번 군집은 잘 구분되었으나, 2번 군집은 1번 군집과 혼재되었다. 이러한 현상은 점용 콘텐츠가 점용 목적과 행위에 해당하는 단어들이 혼재되어 나타난 것이다. 게다가, 콘텐츠가 단문이면서 출현 단어의 종류가 많지 않아 나타낸 결과라고 생각된다. 이러한 요인으로 모든 군집이 명확하게 구분되기보다는 2번 군집은 1번 군집에

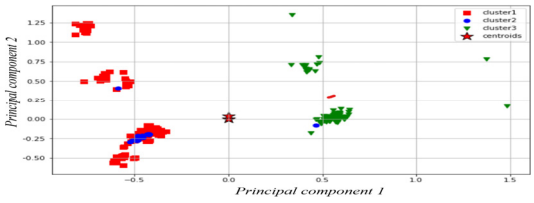


Fig. 4. Visualization of 3 cluster analysis results

포함되는 형태를 가진 것으로 보인다. 또한, 중심점을 기준으로 1번 군집과 3번 군집이 양분화 되는 형상을 갖는 것으로 파악되었다.

3.4.3 유사 질의어 분석

1,065건의 콘텐츠와 3,776개의 단어를 행렬로 배열한 후에 동시 출현기반의 빈도 분석을 통해 특정 단어를 질의어로 지정하였을 때 유사성이 높은 단어를 측정하였다. Table 3은 코사인 유사도를 계산하여 말뭉치와 질의어 간의 거리를 측정한 결과를 나타낸 것이다. 코사인 유사도가 1에 가까울수록 두 단어는 유사도가 높다고 간주할 수 있다. Drainage는 Bridge과 연관성이 낮으나, Buried와는 연관성이 있는 것을 알 수 있다. Drainage, Bridge, Buried는 Installation과 Facility와 연계된 것으로 보인다. 이는 Drainage, Bridge, Buried는 시설을 설치하는 의미와 연관 관계를 갖는다고 해석할 수 있다.

Table 3. Similar query words analysis

Search word	Similar words(cosine similarity)
Drainage	Pipeline(0.38), Buried(0.29), Installation(0.28), Facility(0.25), Road(0.16)
Bridge	Facility(0.24), Installation(0.19), Project(0.11), Build(0.09), Road(0.07)
Buried	Pipeline(0.56), Installation(0.55), Facility(0.34), Build(0.33), Drainage(0.29)

3.5 연관어 분석

3.5.1 주제 개수 설정

점용 콘텐츠에 적합한 최적의 주제 개수를 얻기 위해 주제 개수를 2~9까지 바꿔어가면서 혼란도와 주제일관성의 변화 정도를 살펴보았다. Fig. 5는 혼란도와 주제일관성의 척도 변화를 나타낸 것이다. 주제 개수를 늘릴수록 혼란도는 감소하는 것을 알 수 있다. 또한, 주제 개수를 늘릴수록 주제일관성은 증가와 감소를 반복하였다.

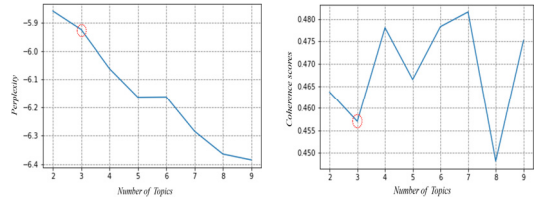


Fig. 5. Estimating the number of topics

이때 주제일관성이 낮을수록 신뢰성이 있다는 의미를 갖는다. 따라서 주제 개수를 3개로 하였을 때 주제일관성 척도가 급격히 낮아졌다. 한편, 혼란도는 3개의 주제 개수를 주었을 때 급격히 감소하기 시작하였다. 혼란도는 혼란도의 값이 작으면 작을수록 해당 주제 모델은 실제로 잘 반영하였다고 해석할 수 있다. 따라서 변화의 시작점인 3개를 주제 개수로 사용하는 것이 가장 합당한 주제 개수라고 해석하였다.

3.5.2 중심성 측정

주제에서 단어들 간의 상호 영향력을 파악하기 위해 중심성 계수를 측정하였다. Table 4는 전체 하천을대상으로 중심성 계수별로 측정하여 얻은 상위 6위 단어들을 나타낸 것이다.

Table 4. Top 10 words by type of centrality coefficient

Centrality	Top 6 words(distribution probability)
Degree	Installation(0.562), Buried(0.188), Facility(0.125), Drainage(0.125), Pavement(0.125), Pipeline(0.063)
Betweenness	Installation(0.292), Buried(0.025), Pavement(0.008), Facility(0.0), Drainage(0.0), Pipeline(0.0)
Closeness	Installation(0.563), Facility(0.316), Drainage(0.316), Bridge(0.298), Power pole(0.298), Facility equipment(0.298)
Eigenvector	Installation(0.684), Facility(0.318), Drainage(0.318), Bridge(0.217), Power pole(0.217), Facility equipment(0.217)
Page rank	Installation(0.235), Buried(0.113), Facility(0.102), Drainage(0.088), Pavement(0.086), Pipeline(0.066)

Installation은 모든 중심선 계수에서 가장 영향력을 갖는 단어로 파악되었다. 이외에도 Facility, Bridge, Drainage, Pipeline 등의 단어는 중심성 순위에서는 다소 차이가 있지만 연관어들 간에 주도적인 위치를 차지하는 것으로 알 수 있었다. 연결 중심성에서는 Buried와 Facility가 높게 나타났다. 매개 중심성에서는 Pavement가 상위 6위 안에 포함되었다. 게다가, 근접 중심성과 고유벡터 중심성에서는 Power pole, Facility equipment는 상위 6위 안에 포함되었다. 페이지 랭크에서는 새로운 단어를 포함하기 보다는 다른 중심선 계수의 상위 6위에 해당하는 단어들을 포함하고 있다. Fig. 6은 전체 하천의 점용 콘텐츠를 말뭉치로 취합한 후에

말뭉치에 대해 페이지 랭크를 적용하였을 경우에 단어들 간에 연관 관계를 나타낸 것이다. Buried와 Installation과 같이 점용 행위에 관한 단어들은 네트워크 내에서 중심적 역할을 하였다. Drainage, Facility, Pipeline와 같은 공작물을 의미하는 단어들이 유기적으로 연관되어 있는 것을 볼 수 있다. Multipurpose square와 Soccer field는 서로 연관 관계를 갖지만 연결망 내에는 포함되지 않고 네트워크 내의 중심에서 많이 떨어진 독립적인 위치에 있는 것을 알 수 있다.

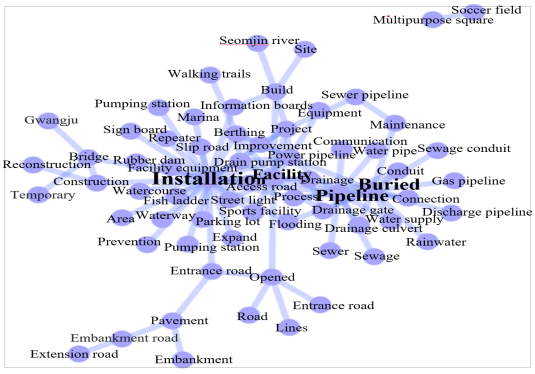


Fig. 6. When applying page rank, relevance of words

3.5.3 주제 모델 분석

3개의 주제를 설정한 후에 의미적으로 일관된 주제 단어들을 주제별로 분류하기 위해 LDA기반으로 주제 모델링을 하였다. 주제 모델링을 통해 말뭉치의 단어들이 주제별로 얼마나 잘 할당하였는지를 평가하기 위해 Gensim의 Word2Vec 라이브러리를 사용하였다. Word2Vec을 사용하여 점용 콘텐츠와 점용 콘텐츠별로 출현하는 단어를 벡터행렬로 환원한 후에 코사인 유사도를 계산하여 벡터 간의 유사도와 중심성 계수를 측정하였다. 이렇게 측정한 값은 단어들 간에 유의미한 정도를 해석하는데 사용하였다. 다음으로, pyvis.network 라이브러리를 사용하여 주체별로 차지하는 비율을 살펴보았다. Table 5는 전체 하천에 대한 주제별로 상위 10위에 해당하는 주제 단어를 도출한 것이다.

pyvis.network를 사용하였을 때, 첫 번째 주제(T1)는 전체 주제 비율 중 46.1%로, 거의 과반수를 차지하였고, 상대적으로 셋 번째 주제(T3)가 20.5%로 가장 낮은 것으로 분석되었다. 또한, 도로 포장과 교량과 같은 시설물의 설치와 관련된 주제어가 영향력이 높은 것으로 판단된다. 두 번째 주제(T2)는 Power pole, Embankment와 같이 하천유역의 지상 설치물과 관련한 주제어가 상

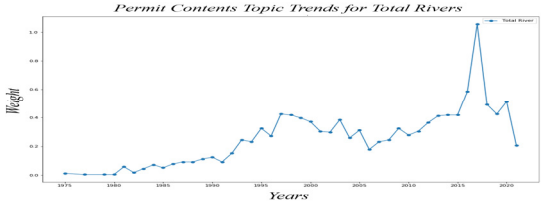
Table 5. Result of topic modeling

Topic #	Top 10 words(distribution probability)
T1 (46.1%)	Installation(0.065), Facility(0.032), Bridge(0.022), Pavement(0.021), Temporary(0.021), Debouchment (0.020), Road(0.018), Drainage(0.014), Entry(0.013), Buried(0.012)
T2 (33.4%)	Installation(0.189), Facility(0.042), Embankment(0.033), Buried(0.031), Pavement(0.031), Drainage(0.027), Pipeline(0.026), Bridge(0.024), Road(0.024), Power pole(0.020)
T3 (20.5%)	Buried(0.079), Installation(0.060), Drainage(0.047), Pipeline(0.030), Facility(0.028), Road(0.021), Pump-ing station(0.020), Sewer(0.016), Project(0.016), Intercept(0.016)

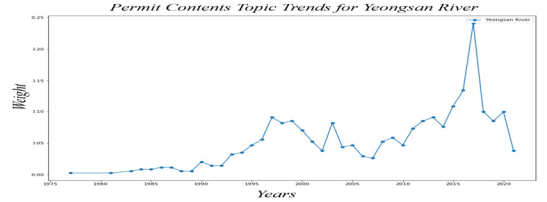
위의 영향력에 포함하는 것으로 분석되었다. 세 번째 주제(T3)는 Buried, Installation, Drainage, Pumping station, Sewer와 같이 지하매설물과 관련한 주제어가 영향력이 있는 것으로 측정되었다. 하지만, 전반적으로 Installation, Facility, Pipeline, Pavement, Pipeline은 모든 주제에 공통으로 포함되었고, Installation은 점용 콘텐츠에서 가장 유의미한 단어라고 평가할 수 있다. 이러한 평가는 TF-IDF, 군집 분석, 모든 중심성 계수에서도 가장 높게 측정된 것과 일맥상통한 것을 알 수 있다. 다만, 주제별로 중복된 단어들이 존재하였다. 이는 점용 콘텐츠가 하천시설의 점용, 하구둑・다목적댐의 공작물의 신축・개축・변경, 토질의 굴착・성토・절토・그 밖의 토질의 형질 변경 등에 대해 점용허가에 국한되었다. 이로 인해 중복성이 높은 것으로 판단되었다. 특히, 콘텐츠 당 평균 3.54개의 단어로 이루어진 단문이면서 개조식으로 작성되었고, 출현단어의 종류도 한정되어 나타난 현상이라 사료된다. 한편, Fig. 7은 전체 하천, 허가 건이 가장 많은 영산강 및 허가 건이 가장 적은 탐진강을 대상으로 연도별로 주제의 변화추이를 나타낸 것이다.

Fig. 7의 (a)~(c) 모두는 주제 단어들이 2010년대부터 급격히 증가하였다가 2020년에 다시 감소하였다. 이 시기에 주제 단어가 가장 많은 것은 그 만큼 점용허가가 많은 것으로 유추할 수 있다. 이러한 현상은 4대강 정비사업이 2009년 7월에 영산강 유역을 시작으로 본격적으로 착공하면서 영산강을 비롯하여 섬진강 유역과 지천에 제방보강, 호안 보호공, 자전거도로, 산책로, 수변공원, 생태습지, 초지원, 수문, 하수처리시설 등의 신설 또는 증설에 따른 점용 허가가 급격히 증가한 영향이라 해석된다. 한편, 전체 하천과 영산강은 거의 유사하게 주제

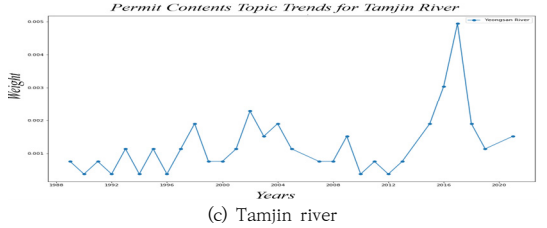
변화하였고, 탐진강은 다른 하천에 비해 1995년에서 2000년까지 증가폭이 큰 것으로 보인다. 이는 전남 장성과 영암지역까지 남해고속도로가 연장되면서 탐진강의 수변지역에 정비사업과 탐방로와 같은 생태공원 등을 조성됨에 따라 점용 신청이 증가하여 나타낸 결과로 보인다. 이처럼 전남 지역의 국가하천에서 2010년대부터 급격히 증가한 것은 그 만큼 하천구역에서의 개발 행위가 많았다고 유추할 수 있다.



(a) Total rivers



(b) Yeongsan river



(c) Tamjin river

Fig. 7. Topic trends by years

3.5.4 연관어 분석

주제 모델링을 통해 얻은 주제에 포함된 단어들의 의미적 연관 관계를 쉽게 살펴보기 위해 그래프로 도식화한 연관어 연결망을 형성할 수 있다. Fig. 7은 전체 하천의 점용 콘텐츠에 대한 연관어 연결망을 나타낸 것이다. 연결망 내에서 주제 단어를 의미하는 노드와 두 개 이상의 노드들 간의 연관 관계를 연결선으로 표시하였다. 중요도가 높은 노드는 크고, 진하게 나타내었다. Fig. 8에서 군집 모델과 주제 모델에서 측정한 것과 같이 Installation과 Facility가 연결망에서도 연관되는 노드의 수가 가장 많아서 가장 유의미한 단어로 평가되었다.

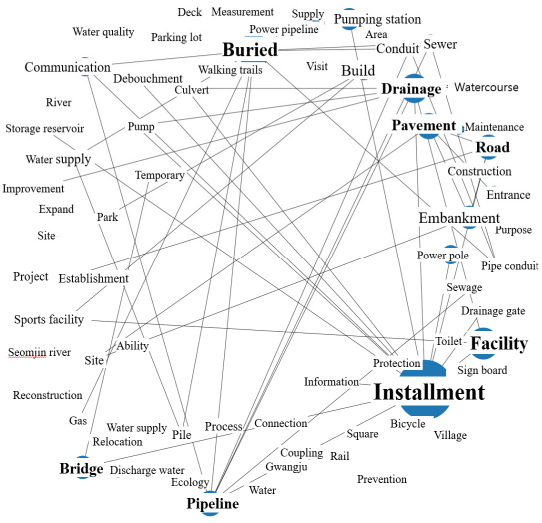


Fig. 8. Associated words network

나머지 상위 단어들도 비슷한 것으로 확인되었다. 다만, buried와 Pipeline은 군집 모델과 주제 모델에 비해 연결망에서 연관성이 높은 것으로 보였다. 점용 행위를 의미하는 Installment, Buried가 중심적 위치에 있었고, 도로시설물인 Road, Pavement, Bridge, Facility 및 하천수 관련 공작물인 Drainage, Pipeline, Embankment가 상대어들과 연관 관계가 많은 것으로 판단되었다. 한편, 하천 보존을 의미하는 Protection은 설치와 직접적으로 연관되었고, Prevention은 독립적인 위치에 있었다.

4. 결론

과거에는 하천구역의 개발을 공적 또는 사전 이익을 얻기 위해 점용 신청이 많았다. 최근에는 하천의 오염과 하천구역의 친환경적 보존이 강조되면서 점용허가를 최소화하는 경향이다. 한편, 허가관리청은 점용 건을 허가대장으로 기록하고 있으나, 허가대장에 내재된 유용한 정보를 제대로 활용하지 못하고 있다. 본 연구는 허가대장에 수록된 점용 콘텐츠에 숨어있는 정보를 유출하여 하천구역의 난 개발 억제에 필요한 기초자료를 생성하고자 하였다. 이를 위해 전라도 지역에 위치한 5개의 국가하천에서 점용된 허가대장을 수집, 분류, 정리한 후에 단어출현의 빈도, 군집화, 주제 모델링 및 연관어 연결망 등을 분석하였다. 이러한 분석 과정을 통해 주로 하천구역에서 도로 시설물과 하천수와 관련된 공작물의 설치와 관련한 점용 행위 위주로 점용이 허가된 것으로 파악되

었다. 다만, 환경 보존과 관련될 수 있는 Protection은 Installation과 연관되지만 Prevention은 연결망 밖에 위치하는 것으로 나타났다. 이처럼 아직까지는 대부분의 점용허가가 개발 위주로 진행된 것으로 해석할 수 있다. 한편, 본 연구의 분석결과는 다음과 같은 의미를 가진다. 국내 처음으로 하천 점용 콘텐츠를 대상으로 연관어의 관계 분석을 통해 점용의 시사점을 유추하였다. 이러한 시사점은 하천구역에서의 개발에 대한 인식 제고를 위한 기초자료로 활용될 수 있을 것으로 사료된다. 다음으로, 점용 콘텐츠는 주제 범위가 한정되고 중복 단어가 많으며 개조식의 단문으로 구성되어 있다. 이러한 특성을 갖는 텍스트를 연관어 분석하는데 본 연구내용이 도움을 줄 수 있을 것으로 생각된다. 하지만, 본 연구결과는 몇 가지 한계와 추가적인 연구가 필요하다. 첫째로, 전라도에 위치한 5개 국가하천의 점용 콘텐츠만을 분석 대상으로 정하였다. 본 연구 결과를 국가 차원의 하천 점용 정책 대안의 기초자료로 활용하기 위해서는 나머지 지역과 토지의 점용까지 분석 대상을 확대할 필요가 있다. 둘째로, 점용 콘텐츠만을 분석 대상으로 하였다. 내실 있는 연구결과를 얻기 위해서는 점용 신청 자료와 허가 조건 등의 콘텐츠를 확충하여 단문의 콘텐츠가 가질 수 있는 제약을 보완할 필요가 있다. 셋째로, 보다 정세한 연관어를 분류하기 위해서는 Facility, Facility equipment와 같이 파생된 유사 의미를 갖는 단어들을 어근 동일화(stemming)할 필요가 있다. 끝으로, 본 연구에서는 가장 보편적으로 사용되는 텍스트 마이닝기반의 분석 기법과 절차를 준용하여 연관어의 관계를 분석하였다. 보다 객관적이고, 신뢰성 있는 연구 결과를 얻기 위해서는 본 연구에 적합한 분석 이론과 검증 방법에 대한 연구가 추가적으로 요구된다.

References

[1] C. W. Kim, Measures to improve the efficiency of river occupation permit business, Presentation Report, Han River Flood Control Offices, Korea.

[2] S, E. Lee, et al., Research on river occupancy system improvement and standard manual development, Research Report, Korea Research Institute, Ministry of Land, Infrastructure and Transport, Korea, pp.19-27.

[3] D, H. Seo, Catch Text Mining with Python, BJ Publishers, 2019, pp.126-128.

[4] Y. O. Do, E. J. Ko, Y. M. Kim, H. G. Kim, G. J. Joo, J. Y. Kim, H. W. Kim, "Using Text-mining Method to Identify Research Trends of Freshwater Exotic Species in Korea", *Korean Journal of Ecology and Environment,* Vol.48, No.3, pp.195-202, 2015. DOI: http://dx.doi.org/10.11614/KSL.2015.48.3.195

[5] J. Y. Kim, Y. N. DO, G. J. Joo, E. H. Kim, E. Y. Park, S. H. Lee, M S. Baek, "The Research Trend and Social Perceptions Related with the Tap Water in South Korea", *Korean Journal of Ecology and Environment*, Vol.49, No.3, pp.208-214, 2016. https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002154425

[6] H. J. Kim, T. H. Lee, S. E. Ryu, N. R. KLim, "A Study on Text Mining Methods to Analyze Civil Complaints: Structured Association Analysis", *Journal of the Korea Industrial Information Systems Research*, Vol.23, No.3, pp.13-24, 2018. DOI: https://doi.org/10.9723/jksiis.2018.23.3.013

[7] T. H, Jo, "Concepts and Applications of Text Mining", *Journal of Scientific & Technological Knowledge Infrastructure*, Vol.5, pp.76-85, 2001.

[8] S. S. Choi, *A Study on the Analysis of Trend and Prediction Models in the Aviation Industry Using Text Mining*, Master's thesis, Korea Aerospace Univ, Korea, 2021.

[9] H. Y. Lee, S. J. Kwak, "Relation Analysis Among Academic Research Areas Using Subject Terms of Domestic Journal Papers", *Journal of the Korean Biblia Society for Library and Information Science*, Vol.22, No.3, pp.353-371, 2011.

[10] M. J. Kim, C, J. Kim, "Trend Analysis of News Articles Regarding Sungnyemun Gate using Text Mining", *Journal of contents*, Vol.17, No.3, pp.474-485, 2017. DOI: https://doi.org/10.5392/JKCA.2017.17.03.474

[11] J. W. Seo, Text Mining for Practice, https://github.com/fingeredman/text-mining-for-practice accessed Oct. 1, 2021)

[12] J. H. Lee, M. B. Lee, J. O. Kim. "A study on Korean language processing using TF-IDF", *The Journal of Information Systems*, Vol.28, No.3, pp.105-121, 2019. DOI: http://dx.doi.org/10.5859/KAIS.2019.28.3.105

[13] J. S. Heo, H. K. Lim, K. H. Kim, Y. H. Han, "Finding Meaningful Chronological Pattern of Key Words in Computer Science Bibliography", *Annual Conference of KIPS 2016*, Korea, Vol.23, No.2, pp.542-545, Nov. 2016. DOI: http://dx.doi.org/10.1007/978-981-10-3023-9_131

[14] B. M. Kang, "Constructing Networks of Related Concepts Based on Co-occurring Nouns", *Korean Semantics*. Vol.32, No.32, pp.1-28, 2010.

[15] S. K. Seo, E. K. Chung, "Domain Analysis on the Field of Open Access by Co-Word Analysis", *Journal of the Korean BIBLIA Society for library and Information Science*, Vol.24, No.1, pp.207-228, 2013.

DOI: https://doi.org/10.14699/kbiblia.2013.24.1.207

[16] J. Chang, J. S. Wang, S. Gerrish, S. Wang, D. M. Blei, "Reading tea leaves: How humans interpret topic model", *Advances in Neural Information Processing System*, pp.1-9, 2009.

[17] S. K. Seo, E. K. Chung, "Domain Analysis on the Field of Open Access by Co-Word Analysis", *Journal of the Korean Biblia Society for Library and Information Science*, Vol.24, No.1, pp.207-228, 2013.
DOI: https://doi.org/10.14699/kbiblia.2013.24.1.207

[18] H. J. Ko, *Evaluation of ecosystem services of urban green spaces using text mining techniques*, Ph.D dissertation, Seoul University of Interdisciplinary Program in Landscape Architecture, Seoul, Korea, pp.67-69, 2019.

[19] S. G. Cho, S. B Kim, "Finding Meaningful Pattern of Key Words in IIE Transactions Using Text Mining", *Journal of the Korean Institute of Industrial Engineers*, Vol.38, No.1, pp.67-73, 2012.
DOI: https://doi.org/10.7232/JKIIE.2012.38.1.067

[20] G. K. Zipf, Zipf' Law [Internet] Wikipedia, Available From: https://ko.wikipedia.org/wiki/%EC%A7%80%ED%94%84%EC%9D%98_%EB%B2%95%EC%B9%99 (accessed Oct. 1, 2021)

정 성 윤(Seong-Yun Jeong) [정회원]

- 1992년 2월 : 한양대학교 전자계산학과 (공학학사)
- 1994년 2월 : 숭실대학교 컴퓨터공학과 (공학석사)
- 2018년 2월 : 서울과학기술대학교 산업정보시스템전공 (공학박사)
- 1994년 3월 ~ 현재 : 한국건설기술연구원 미래스마트건설연구본부 연구위원

〈관심분야〉
건설정보화, 텍스트 마이닝, 투자공학

김 남 곤(Nam-Gon Kim) [정회원]

- 1989년 2월 : 울산대학교 전자계산학과 (공학학사)
- 2001년 2월 : 공주대학교 컴퓨터공학과 (공학석사)
- 1991년 10월 ~ 현재 : 한국건설기술연구원 미래스마트건설연구본부 연구위원

〈관심분야〉
건설정보화, 텍스트 마이닝, 빅데이터