Journal of the Korean Association of Geographic Information Studies, 20(1) 2017, pp. 113-126 ISSN 1226-9719(Print)
https://doi.org/10.11108/kagis.2017.20.1.113 ISSN 2287-6952(Online)

텍스트 마이닝 기법을 활용한 환경공간정보 연구 동향 분석*

오관영[1] · 이명진[1※] · 박보영[1] · 이정호[1] · 윤정호[1]

Analysis of the Research Trends by Environmental Spatial-Information Using Text-Mining Technology*

Kwan-Young OH[1] · Moung-Jin LEE[1※] · Bo-Young PARK[1]
Jung-Ho LEE[1] · Jung-Ho YOON[1]

요 약

본 연구의 목적은 빅데이터 분석 기법 중 하나인 텍스트 마이닝 기법을 활용하여 환경 분야의 환경공간정보 활용 연구 동향을 정량적으로 분석하는 것이다. 분석에 활용된 자료는 NDSL (National Digital Science Library)을 통하여 획득한 국내 논문으로 총 869편을 대상으로 하였다. 논문에서 추출된 단어들은 "환경일반", "기후", "대기", 등 환경 분야 10개, "위성영상", "수치지도", "재난재해" 등 환경공간정보 20개로 설정된 분류체계에 따라 재분류 되었다. 재분류된 분류 키워드를 통해, 논문에서 해당 키워드의 출현 빈도 및 시계열 변화를 파악하였으며, 상호 간 연관분석을 수행하였다. 첫째, 빈도 분석 결과 환경 분야에서는 "환경일반"(40.85%)이 환경공간정보에서는 "위성영상" (24.87%)이 가장 높은 활용 빈도를 나타냈다. 둘째, 환경 분야에 대한 시계열 분석 결과 1996년부터 2000년까지는 "기후"에 대한 연구 비중이 높았으나, 2001년부터는 "환경일반"에 대한 연구가 증가하였다. 환경공간정보에서는 "위성영상"에 대한 수요가 전 기간에 걸쳐 가장 높았으며, 활용 비율 또한 점차적으로 증가하고 있었다. 셋째, 환경 분야와 환경공간정보에 대한 연관분석 결과 총 80개의 연관 규칙이 생성되었으며, 환경 분야 중 "환경일반"이 "위성영상", "전자지도" 등 총 17개의 환경공간정보와 가장 많은 수의 연관 규칙을 생성하였다.

주요어 : 텍스트 마이닝, 환경공간정보, 빈도 분석, 시계열 분석, 연관 분석

ABSTRACT

This study aimed to quantitatively analyze the trends in environmental research that utilize environmental geospatial information through text mining, one of the big data

2016년 12월 27일 접수 Received on December 27, 2016 / 2017년 2월 26일 수정 Revised on February 26, 2017 / 2017년 3월 21일 심사완료 Accepted on March 21, 2017

* 본 연구는 한국환경정책 · 평가연구원이 수행하는 국토교통부 국토교통기술촉진연구사업(과제번호 16CTAP-C114629-01-000000) 및 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 기초연구사업(NRF-2014R1A1A1002704)의 지원을 받았음.

1 한국환경정책 · 평가연구원 Korea Environment Institute
* Corresponding Author E-mail : leemj@kei.re.kr

analysis technologies. The analysis was conducted on a total of 869 papers published in the Republic of Korea, which were collected from the National Digital Science Library (NDSL). On the basis of the classification scheme, the keywords extracted from the papers were recategorized into 10 environmental fields including "general environment", "climate", "air quality", and 20 environmental geospatial information fields including "satellite image", "numerical map", and "disaster". With the recategorized keywords, their frequency levels and time series changes in the collected papers were analyzed, as well as the association rules between keywords. First, the results of frequency analysis showed that "general environment"(40.85%) and "satellite image"(24.87%) had the highest frequency levels among environmental fields and environmental geospatial information fields, respectively. Second, the results of the time series analysis on environmental fields showed that the share of "climate" between 1996 and 2000 was high, but since 2001, that of "general environment" has increased. In terms of environmental geospatial information fields, the demand for "satellite image" was highest throughout the period analyzed, and its utilization share has also gradually increased. Third, a total of 80 correlation rules were generated for environmental fields and environmental geospatial information fields. Among environmental fields, "general environment" generated the highest number of correlation rules (17) with environmental geospatial information fields such as "satellite image" and "digital map".

KEYWORDS : Text Mining, Environmental Geospatial Information, Frequency Analysis, Time Series Analysis, Association Analysis

서 론

환경공간정보는 환경현황 및 분석 데이터를 지리 정보와 함께 제공하는 것으로 GIS(Geographic Information System)의 발전과 더불어 그 활용 가치가 꾸준히 상승되고 있다(Cho *et al.*, 1998; Yun *et al.*, 2010; Lim *et al.*, 2014a). 국내 환경공간정보는 환경부 주도로 1998년 구축된 전국 단위 대분류 토지피복지도 제작으로부터 시작되었다. 대분류 토지피복지도는 USGS(United States Geological Survey)의 Landsat 위성영상을 활용하여 남북한 전체를 7개 대분류(시가화 건조지역, 농업지역, 산림지역, 초지, 습지, 나지 및 수역 등)로 구분하였다(Oh *et al.*, 2016). 이후, 환경공간정보의 범위는 점차 확장되어 환경부의 생태자연도, 국토환경성 평가지도, 산림청의 임상도, 농촌진흥청의 토양도, 항공우주연구원의 위성영상, 국토지리정보원의 수치지도 등 그 활용 목적에 따라 다양한 기관에서 생성 및 배포되고 있다.

환경공간정보는 생성 자체로 끝나는 것이 아니라 기존의 환경 분야 연구에 융합되어 활용될 때 더 큰 의미를 지닌다. 예를 들어, 도심지역의 열섬현상 분석에 있어 열적외선 위성영상, 토지피복지도(산림지역 변화) 등의 환경공간정보 융합 활용은 기존 연구의 신뢰성과 활용성을 크게 향상시킬 수 있다(Oh *et al.*, 2016). 환경공간정보가 다양화되고, 시계열의 자료가 구축됨에 따라 타 기술과의 융합 활용의 가능성이 확대되고, 더 많은 가치를 창출할 수 있을 것으로 기대된다. 환경공간정보의 융합 활용을 위해서는 환경 분야와 환경공간정보 모두에 대한 이해가 필요하다. 그러나 환경 분야에는 전통적인 환경 전문가들은 많음에 비해 환경공간정보의 융합 활용을 위한 총괄적 관점의 전문가는 부족하다. 이러한 관점에서 환경 분야별 환경공간정보의 발전 추이 및 활용 동향을 파악하고, 이를 기반으로 기존 연구와 환경공간정보의 융합 활용 연구의 확장 또는 발전을 위한

체계화 연구가 필요하다.

한편, 빅데이터 분석 기술 중 텍스트 마이닝 기법이 발달됨에 따라 장기간에 걸친 대용량의 텍스트 자료에 대한 분석이 가능해졌다. 텍스트 마이닝은 Feldman and Dagan(1995)에 의해 텍스트 데이터베이스 기반의 지적 발견(Knowledge Discovery in Textual Database)이라는 개념으로 처음 언급되었다. 텍스트 마이닝은 텍스트 기반 데이터베이스로부터 자연어(문자정보) 처리를 통하여 잠재되어 있는 문헌의 패턴을 분석하는 방법이다(Feldman and Hirsh, 1996). 또한 기계적 알고리즘을 활용하여 인간이 관심을 가지는 자연어를 자동으로 추출하는 것을 의미한다(Hotho *et al.*, 2005). 최근 전산 · 통계, 식품, 건설 등 다양한 분야에서 텍스트 마이닝 기법을 활용한 국내외 연구가 증가하고 있다. Shin *et al.*(2013)은 에너지 하비스팅 분야의 핵심 키워드를 추출하고, 핵심 키워드와 서브 키워드들 간의 상호 연관관계와 계측 구조 관계 분석을 수행하였다. Bae *et al.*(2013)은 텍스트 마이닝 기법을 이용하여 기후변화 관련 식품 분야 논문 초록에서 사용된 용어들의 출현 빈도를 분석하고, 이를 통해 관련 연구들의 추세와 관심 주제어를 추출하는 연구를 수행하였다. Kim and Chang (2011)은 축적된 뉴스 기사의 분석을 통해 검색 키워드와 밀접하게 연관된 키워드를 추출하는 연구를 수행하였다. Lim *et al.*(2014b)은 U-city와 Smart City에 대한 논문 분석을 통하여 두 용어에 대한 유사성과 차별성을 분석하고, 특정 키워드와 연관성이 높은 연관 키워드 추출에 대한 연구를 수행하였다. 공간정보와 관련된 텍스트 마이닝 연구도 부분적으로 진행되었다. Sakong and Seo(2007)은 GIS 관련 학회에 게재된 논문을 대상으로 GIS 기술의 연구 동향을 분석하였다. Lim *et al.*(2014a)은 공간정보 관련 논문 및 보고서에서 추출된 키워드의 시기별 출현 빈도 및 변화를 분석하여, 공간정보 분야의 연구 동향을 정리하였다. 해당 연구에서는 공간정보 관련 키워드를 system, data, application, method 등으로 군집화 하여 분석하였다. 한편, 선행된 연구들은 공간정보 및 GIS 기술에 대한 동향 분석으로 환경 분야와의 관련성은 고려되지 않았다. 이에 본 연구에서는 환경공간정보의 환경 분야 활용의 관점에서 특화된 연구를 수행하고자 하였다. 세부적으로는 환경 분야 10개, 환경공간정보 20개로 구분된 분류체계를 중심으로 환경 분야 및 환경공간정보 별 빈도, 추이, 연관성 등을 종합적으로 분석하였다.

본 연구의 목적은 텍스트 마이닝 기법을 활용하여 환경 분야별 활용된 환경공간정보의 빈도, 시계열 경향, 연관성 등을 정량적으로 분석하는 것이다. 이를 위하여 첫째, 최근 20년(1996년-2015년)동안 발간된 환경공간정보가 활용된 환경 분야 국내 논문들을 수집하였다. 둘째, 텍스트 마이닝 분석을 위하여 환경공간정보 및 환경 분야에 대한 분류체계를 구성하였다. 셋째, 전술된 분류체계에 따라 수집된 논문의 초록에서 유의미한 키워드를 추출하고, 추출된 키워드에 대한 빈도 분석, 시계열 분석 및 연관분석을 수행하였다. 이를 통해 환경 연구에서 주요하게 활용된 환경공간정보가 무엇이며, 그 경향성이 무엇인지 점검하였다.

연구방법

1. 연구 자료

본 연구는 텍스트 마이닝 기법을 통해 환경공간정보의 환경 분야 활용에 대한 연구 동향을 파악하는 것이다. 따라서 텍스트 마이닝 기법을 위한 연구 자료는 국내에서 발행된 연구 논문을 대상으로 하는 것이 타당하다. 왜냐 하면, 국내에서 생성된 환경공간정보는 국내 연구에서 제한적으로 사용될 가능성이 높으며, 연구 논문 이외의 신문기사, 블로그 등의 자료 등은 본 연구의 전문성에 부합되지 않기 때문이다. 연구 자료의 수집 시기는 최근 20년(1996년-2015년)간 발행된 국내 논문으로 한정하였다. 왜냐하면 최근 20년 이전에는 국내 환경공간정보 관련 논문의 수가 매우 적기 때문이다. 본 연구에서 사용된 자료는 NDSL (National

TABLE 1. Searching keyword

Division	Searching keyword
Environment field (35)	Environment problems, environment policy, water pollution, soil contaminant, air pollution, climate change, disaster, strong wind, drought, flood, landslide, global warming, sea-level, volcano, heavy rain, heavy snow, yellow dust, earthquake, fine dust, habitat, eco-friendly, agricultural etc.
Environment geospatial information field (21)	Environment map, spatial information, GIS, remote sensing, land cover map, ecological map, environmental conservation value assessment map, digital map, digital topographic map, geological map, land-use map, forest type map, aerial photograph, aerial images, satellite image etc.

Digital Science Library)에서 환경 분야와 환경공간정보에 대한 검색 키워드를 교차 검색하였을 때 추출되는 국내 논문이다. 이를 위하여 환경 분야 35개, 환경공간정보 분야 21개 검색 키워드를 입력하였다. 검색 키워드의 선정은 빅데이터 분석 도구인 마인즈인사이트(http://www.mindsinsight.co.kr)[1)]에서 "환경 분야"를 입력하였을 때, 또는 "환경공간정보"를 입력하였을 때 도출되는 각각의 연관 키워드를 중심으로 재구성하였다. 이때 선정된 검색 키워드는 다음 표 1과 같다. 전술된 검색 키워드를 NDSL의 검색엔진에 입력하여, 1996년부터 2015년까지 총 869편의 국내 논문을 수집하였다. NDSL에서는 국내에서 발간된 연구논문 전체에 대한 초록 및 서지정보를 제공하기 때문에 기타 학술 검색엔진의 중복검색은 고려하지 않았다.

2. 연구 수행 절차

본 연구의 수행 절차는 크게 자료 전처리 단계와 텍스트 마이닝 분석 단계로 구분된다. 첫째, 자료 전처리 단계는 수집된 논문의 초록을 대상으로 유의미한 단어를 추출하고, 추출된 단어를 환경 분야 및 환경공간정보 분류체계에 따라 재분류하는 과정이다. 단어 추출은 프로그램 언어 R 3.3.1의 한국어 분석 Tool box인 KoNLP (Korean Natural Language Processing)를 활용하였다. 이때, 추출된 단어에는 환경 분야 및 환경공간정보와 부합하지 않는 단어들이 포함되어 있고, 동일한 대상에 대한 유의어가 혼용되어 있다. 예를 들어 추출된 단어에는 환경 분야에서 "기후"로 분류할 수 있는 "온난화", "이상기온", "폭염" 등의 세분화된 단어들이 포함되어 있다. 그러나 본 연구의 범위는 환경공간정보가 활용된 환경 분야 전체를 대상으로 하기 때문에 전술된 세분화된 단어에 대한 분석으로는 유의미한 정보를 도출하기 어렵다. 만약, 연구의 범위를 환경 분야 중 "기후"로 한정하고 자료를 수집했다면 "온난화", "이상기온", "폭염" 등의 세분화된 단어들에 대한 분석이 적합할 것이다. 따라서 본 연구의 범위와 목적에 적합한 분석을 위해서는 불필요한 단어들을 제거하고, 추출된 단어들을 일정한 체계에 따라 군집화하는 과정이 필요하다. 그러나 현재까지 본 연구의 목적에 부합하는 공인된 환경 분야 및 환경공간정보 분류체계에 대한 기존의 연구는 미흡하였다. 이에 본 연구에서는 Lee *et al.* (2014) 및 NSDI(Korea National Spatial Data Infrastructure Portal)에서 제시한 기준을 중심으로 환경분야 및 환경공간정보에 대한 분류체계를 재구성하였다. 재구성된 분류체계에 따른 환경 분야 분류 키워드 10개와 환경공간정보 분류 키워드 20개는 각각 표 2, 3과 같다.

둘째, 텍스트 마이닝 분석 단계에서는 분류체계에 따라 재 군집된 분류 키워드에 대한 빈도분석, 시계열 분석 및 키워드 상호 간 연관 분석을 수행하였다. 빈도 및 시계열 분석은 논문에서 출현한 상대적 중요도를 비교하는 방법이다. 이때, 수집된 각각의 논문에서 분류 키워드가 수차례 반복되었더라도 논문 1편당 1회 언급된 것으로 연산하였으며, 여러 개의 키워드가 포함되어 있더라도 중요도가 큰 대표 키워드를 선정하였다. 즉, 1편의 논문 당 환경 분야 분류 키워드 1개, 환경공간정보 분류 키워드 1개의 대표 키워드 행렬을 추출하였다. 또한, 시계열 분석에서는 1년 동안 발행되는 논문의 수가 적

TABLE 2. Classification keyword of environmental field

Classification keyword	Word
Atmosphere	Air pollution, air quality, fine dust, yellow dust, pollution, exhaust gas, smog, air, smoke, SOx, NOx, nitrogen dioxide, sulfur dioxide, VOC, PM10, PM2.5, atmosphere environment, clear sky, acid rain, visible distance, ozone, dioxin, atmosphere discharge, specific air pollutants, long-range transboundary pollutants, diesel particle, indoor air, flue gas, clean fuel, incinerator, benzene, air pollution measurement network, background concentration, enforcement emission-cap regulation, Seoul Metropolitan air quality, dust
Biodiversity	Biodiversity, nature conservation, natural environment, wildlife, species diversity, invasive alien species, exotic species, endangered animals, native plants, wild birds, forest protection, deforestation, nature destruction, natural land scape, national park, poaching, ecotourism, genetic resources, bio technology, hunting, endemic species, CITES, conservation area, ecosystem service, marine organism, Nagoya protocol, genetically modified organism, GMO, rural forest, marshy land, eco-peace park, genetic diversity, ecological footprint, biocapacity
Climate	Climate, GHG, global warming, carbon dioxide, carbon, heavy snow, heavy rain, flood, drought, typhoon, Intense heat, bitter cold, sea level, heat island, anormal weather, severe rain storm, surge, tsunami, arctic glacier, antarctic ice, glacier of himalayas, greenland ice sheet, desertification, green growth
General environment	Environment, agricultural, damage caused by the wind and food, earthquake, change detection, volcano, ground coverage, submarine groundwater, Mt. Baegdu, forest fire, damage investigation, heat island, environmental planning, environmental education, urbanization
Harmful substance /Health	Environmental hormone, sick house syndrome, carcinogen, environment health, heavy metal pollution, humidifier disinfectants, harmful chemical substance, atopy, environmental diseases, endocrine disrupter, toxic heavy metals, dioxin, asbestos, phthalate, bisphenol A, act on registration and evaluation, etc of chemicals, chemicals control act, radon, defoliant, DDT, formaldehyde, persistent organic pollutants, POPs, cadmium, mercury, hydrofluoric acid, hydrochloric acid, phenol, toxic material, risk
Noise	Noise, sound proof
Ocean	Mud flat, coast, sea level, red tide, surge, tsunami, marine debris, ocean disposal, oil spill, ocean ecology, ocean environment, seascape, reclamation, saemangeum, eutrophication, shoreline erosion, beach erosion, seawater quality, abnormal Waves, rip currents, fish stranding, marine pollution
Soil	Land pollution, soil remediation, soil environment, ground water, soil ecology, abandoned mine, death of livestock, livestock burial areas, heavy metal pollution, oil pollution, contamination sites, soil microbes, soil quality, toxic soil
Waste	Waste, separate garbage collection, recycle, measured rate system, incineration, reclamation, food waste leachate, illegal waste disposal, waste vinyl, ocean disposal, waste resources, waste heat, up-cycle, waste metal, material flow analysis, migration testing, resource circulation, end-of-life vehicles
Water quality	Water-bloom, tap water, water quality, wastewater, BOD, COD, 4 major rivers, river pollution, ecological waters, clear water, clean water, aquatic ecology, river ecology, hydrological environment, buffer strip, total phosphorus, fish mortality, total nitrogen, conduit, excrements, septic tank, water saving, water play, rainwater management, flooding, first class of water, fry discharge, river restoration, restoration project, marshy land, fish stranding, river-water purifying, unauthorized discharge, drinking water, water shortage, water rate, rainwater harvesting, water reuse, seawater desalination

TABLE 3. Classification field of environmental geospatial information field

Classification keyword	Word
Aerial photograph	-
Agricultural	Agricultural infrastructures, farmland comprehensive information, general agricultural and farm villages
Atmosphere / Climate	Climate and atmosphere information
Cadastral map	-
Cultural properties	Cultural assets, theaters, cultural facilities, foreign cultural centers, general culture and cultural assets
Digital map	-

TABLE 3. Continued

Classification keyword	Word
Disaster prevention	119 fire protection, GAS accident information, control measures interest information, collapse information, accident present conditions, forest fire hazard information, landslide hazard information, fire station region, disaster information, electrical fire information, earthquake shelters information, disease information, coastal inundation prediction information, coastal disaster vulnerability assessment
Environment	Environmental conservation value assessment, life environment information, soil environment information, environmental geographic information, information of hazardous chemicals storage facilities, water supply, sewerage and water quality information, water environment information, general environment
Forestry	Trail, forest, forest attractions, forest location information, forest classification, forest land information, forest road network maps, forest type maps, forest cover type maps, soil map, general forestry and mountain eco-villages
Local city	Development restriction area, construction status. old building information, urban planning, commercial district, land use zoning, underground facilities, residential land information, land-use, existing land use, land characteristics, administration border, general regions and citys
Ocean/ Marine product	Mud flat information, landscape area, public waters, national fishing harbors, uninhabited island, fisheries resources protection area, wetlands protection areas, aquafarm information, coastal management zone planning information, coastal information, coastal port, natural coast information, special management marine area, maritime boundary territorial waters straight base-line, beach, sea environmental conservation, shoreline, general ocean, fishery and fishing village
Nature	Degree of green naturality, natural environment conservation area, ecological zoning maps, vegetation maps, land cover classification maps, general nature
Other map	-
Physical map	-
Rail	-
Road	LPG filling stations, traffic CCTV, traffic accident, national traffic information, road transportation, road networks, bus traffic, pedestrian-priority zone, precise networks for pedestrian, filling stations, parking facilities site, taxi stands, general road
Satellite image	KOMPSAT, GEOKOMPSAT, LANDSAT, MODIS, SPOT, Cosmo-SkyMed, ALOS, IRS
Science technology	General science technology
Thematic map	-
Water resource	Water supply facilities, hydrogeolobycal maps, water resources integrated information, boring development information, groundwater information, sewerages, river information, general water resources

음으로, 전체 수집 기간(1996년-2015년)을 5년 단위로 구분하여 수행하였다. 마지막으로 키워드 상호 간 연관 분석은 연관 규칙을 기반으로 빈발 출현 키워드의 쌍을 추출하는 방법이다. 이때, 목표 값에 대한 연관 규칙은 신뢰도(Confidence) 0.005, 지지도(Support) 0.005의 임계 값 환경에서 연산하였다.

텍스트 마이닝 분석 결과

본 연구에서는 전술된 연구 방법에 따라 환경 분야 활용의 관점에서 국내 환경공간정보에 대한 텍스트 마이닝 분석 결과를 제시한다. 이에 수집된 연구 자료의 전처리 과정으로부터 추출된 각 환경 분야와 환경공간정보 분류 키워드를 대상으로 빈도 분석, 시계열 분석, 연관 분석을 수행하였다.

1. 빈도 분석

환경 분야 분류 키워드에 대한 빈도 분석 결과는 그림 1과 같다. 총 10개의 환경 분야로 구분된 분류 키워드의 상위 발생 빈도는 "환경일반" 40.85%(279회), "기후"32.8% (224회)로 전체의 약 73%를 나타냈다. 그 외에는 "생물다양성"(8.93%), "토질"(6.3%), "수질"(4.25%), "대기"(3.37%), "해양"(2.34%), "폐기물"(0.59%),

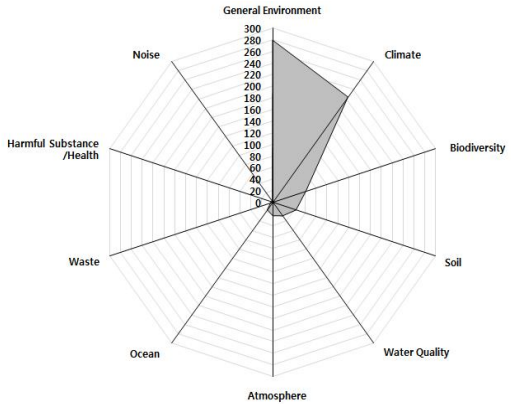

FIGURE 1. Frequency of classification keyword appearance in environmental field

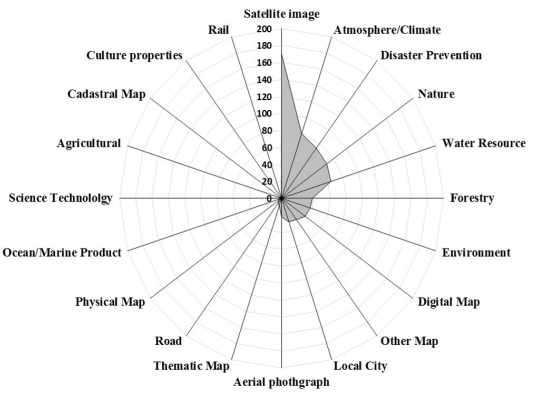

FIGURE 2. Frequency of classification keyword appearance in environmental geospatial information field

"유해물질 및 보건"(0.44%), "소음"(0.15%)의 순위를 보여주었다. "환경일반"에 대한 출현 빈도가 높은 것은 해당 키워드가 포함하는 분야가 농업, 환경교육, 도시화 등 관련 분야의 폭이 넓기 때문인 것으로 판단할 수 있다(표 2). 한편, "기후" 분야는 전체의 32.8%로 "환경일반" 이외의 단일 분류체계에서는 가장 높은 비율을 나타냈다. "기후"가 포함하는 분야는 폭염, 산사태, 이상 기온 등으로 "환경일반"보다는 일관된 주제를 지님으로, 환경 분야에서 가장 많은 관심과 연구가 진행된 분야는 "기후" 분야로 분석하는 것이 바람직하다.

환경공간정보 분류 키워드에 대한 빈도 분석 결과는 그림 2와 같다. 총 20개의 환경공간정보로 구분된 분류 키워드의 상위 발생 빈도는 "위성영상"24.87%(170회), "기후대기"11.71%(80회), "재난방재"10.69%(73회), "자연" 10.1%(69회)로 전체의 약 58%를 나타냈다. 그 외에는 "수자원"(9.37%), "임업"(5.56%), "환경"(5.42%), "전자지도"(5.27%), "기타지도"(4.54%),"지역도시"(4.25%), "항공사진"(3.22%), "주제도"(1.32%), "도로"(1.17%), "지형도"(0.59%), "해양수산"(0.59%), "과학기술"(0.44%), "농업"(0.29%), "지적도"(0.29%), "문화 및 문화재"(0.15%), "철도"(0.15%)의 순위를 보여주었다. 전체적인 활용 빈도는 "위성영상"이 가장 높은 수치를 나타냈는데, 이는 환경 연구의 특성상 공간적 개념의 광의적 분석 및 시계열 분석이 매우 중요하며, "위성영상"이 이러한 요구를 가장 잘 충족하는 연구 자료이기 때문인 것으로 판단 할 수 있다.

그림 3는 각각의 환경 분야별 환경공간정보의 활용 빈도를 나타낸다. "환경일반", "기후", "해양" 분야의 경우 공통적으로 "위성영상"이 가장 높은 활용 빈도를 나타냈다. "위성영상" 다음으로 "환경일반"에서는 "재난방재" 자료가, "기후" 분야에서는 "기후대기" 자료가, "해양" 분야에서는 "해양수산" 자료의 활용이 높게 나타났다. 그림 1,2,3의 결과를 종합해 볼 때, 환경공간정보는 "환경일반", "기후", "생물다양성", "토양", "수질", "대기", "해양"의 환경 분야에서 유의미한 활용을 나타냈다. 또한, "환경일반", "기후", "해양"을 제외한 환경 분야에서 가장 많이 활용된 환경공간정보는 해당 분야의 특수성이 반영된 것으로 타 분야에서의 교차 활용의 빈도가 낮았다. 그러나 "위성영상"은 "환경일반", "기후", "해양" 분야를 비롯한 "생물다양성", "토양", "수질", "대기", 분야에서의 활용 빈도가 상대적으로 높았다. 이는 환경공간정보 생성의 관점에서 다양한 환경 분야에서의 활용을 높이

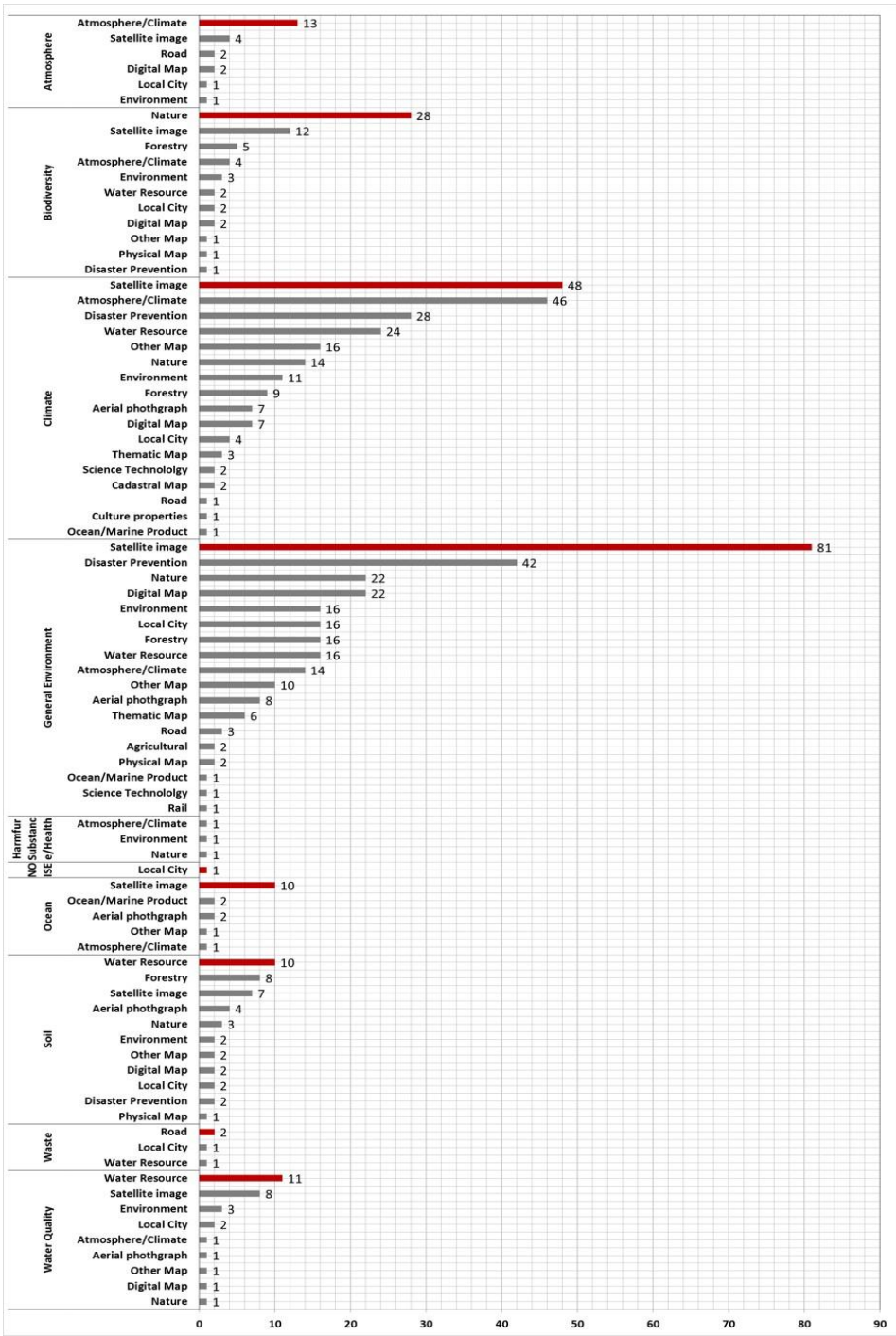

FIGURE 3. Frequency of classification keyword in environmental geospatial information according to environmental field

기 위해서는 목적에 따라 가공되지 않은 위성영상 등의 원시자료가 효과적임을 시사한다.

2. 시계열 분석

환경 분야에 시계열 분석 결과, 1996년부터 2000년까지는 "기후"에 대한 연구 비중이 높았으나, 2001년부터는 "환경일반"에 대한 연구가 증가한 것을 확인할 수 있다(그림 4). 이는 2001년 이후 "농업", "환경 교육" 등 다양한 환경 분야에 대한 관심이 높아진 것과 연관 지을 수 있다. 비록, 2001년 이후 "기후" 분야의 연구 비중이 "환경일반"보다는 다소 낮았으나 전 기간에 걸쳐 상대적으로 높은 비율을 나타냈다. 특히, 2001년-2005년(35회)에서 2006년-2010년(91회)로 전 기간에 걸쳐 가장 급격한 상승을 나타냈다. 이러한 이유는 2005년 2월에 발효된 파리 기후협약과 더불어 "기후" 분야에 대한 세분화된 연구가 증가했기 때문으로 분석 가능하다.

환경 분야에 대한 시계열 분석 결과, "위성영상"에 대한 수요가 전 기간에 걸쳐 가장 높으며, 활용 비율 또한 점차적으로 증가하고 있음을 살펴볼 수 있다. 이는 국내 아리랑 시리즈 및 천리안 시리즈 위성의 발사 및 전 세계적인 지구관측위성의 증가와 관련된 것으로 분석할 수 있다. 더욱이 2018년 천리안 2A호(기상),

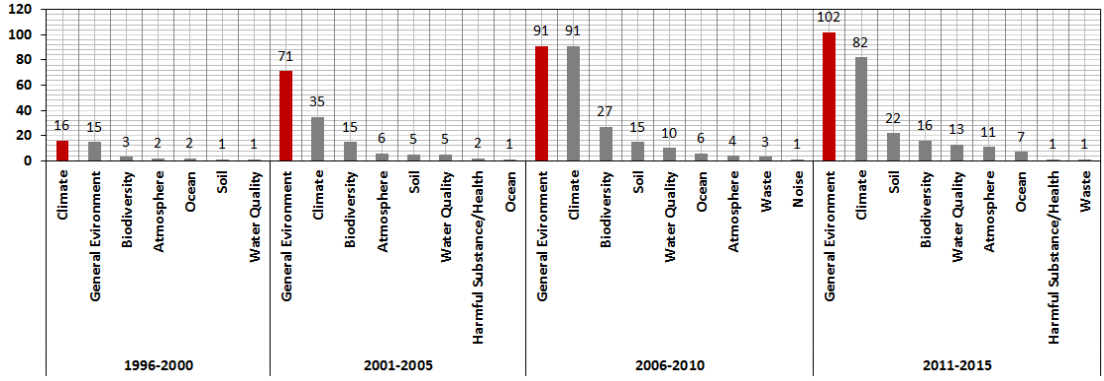

FIGURE 4. Variation of number of classification keyword appearance in environmental field

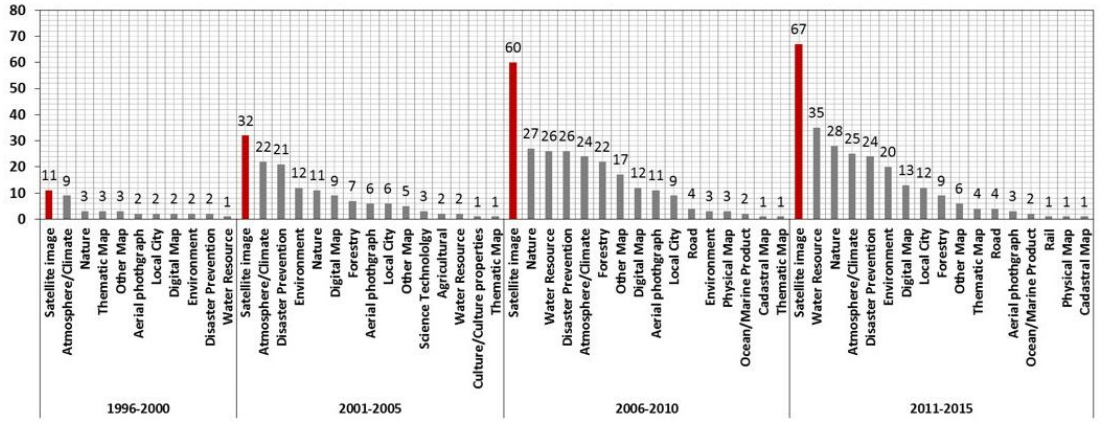

FIGURE. 5 Variation of number of classification keyword appearance in environmental geospatial information field

2019년 천리안 2B(해양, 환경)의 정지궤도 위성의 발사는 "수질", "대기" 분야 등에서의 위성영상 활용 빈도를 더욱 높일 것으로 예상된다. 즉, 환경공간정보의 융합 활용의 관점에서는 "위성영상"에 대한 투자와 관심이 우선적으로 요구되며 각기 다른 센서 특성을 지니고 있는 국내 지구 관측 위성에 대한 다각적인 융합 연구가 필요할 것으로 사료된다(그림 5).

3. 연관분석

본 연구에서는 환경 분야와 환경공간정보에 대한 연관분석을 수행하였다. 연관분석의 척도는 지지도(support), 신뢰도(confidence), 향상도(lift)가 사용되었다. 지지도는 전체 문서에서 키워드 A와 B가 함께 나오는 확률($P(A \cap B)$), 신뢰도는 키워드 A가 나오는 관계 속에서 A와 B가 동시에 나온 비율을 의미하며($P(B|A)$), 향상도는 두 키워드의 독립 관계를 판단한다 ($P(A \cap B)/(P(A)\,P(B))$). 이때, 향상도가 클수록 키워드가 연관성이 큰 것으로 분석할 수 있다. 즉, 향상도가 클수록 하나의 데이터 셋(예: 논문 1편)에 A(예: 환경일반)라는 키워드가 출현할 때, B(예: 농업)라는 키워드가 동시에 출현할 확률이 커진다는 것을 의미한다. 본 연구에서는 적절한 수의 연관 규칙을 도출하기 위하여 연관관계 규칙은 지지도 0.005, 신뢰도 0.005의 임계값 환경에서 연산하였다. 표 4는 각 연관 규칙에 대한 지지도, 신뢰도, 향상도를 나타낸다. 지지도의 경우, 전체 연관 규칙에서 0.012의 거의 유사한 수치를 나타냈으며, 신뢰도는 최소 0.056에서 최대 1의 분포 범위를 나타냈다. 지지도는 최소 0.618에서 최대 10.5의 수치를 나타냈다. 분석 결과, 총 80개의 연관 규칙이 생성되었다. 환경 분야에서 가장 많은 연관 규칙을 나타낸 것은 "환경일반"으로 17개의 규칙을 보였으며, 그 외에 "기후" 16개, "생물다양성" 11개, "토양" 11개, "수질" 9개, "대기" 5개, "해양" 5개, "유해물질 및 보건" 3개, "폐기물" 2개, "소음" 1개의 연관 규칙이 생성되었다. 연관 규칙이 많이 생성되었다는 것은 해당 환경 분야에서 환경공간정보가 중요한 의미를 지니고 있다고 분석할 수 있으며, 이는 앞

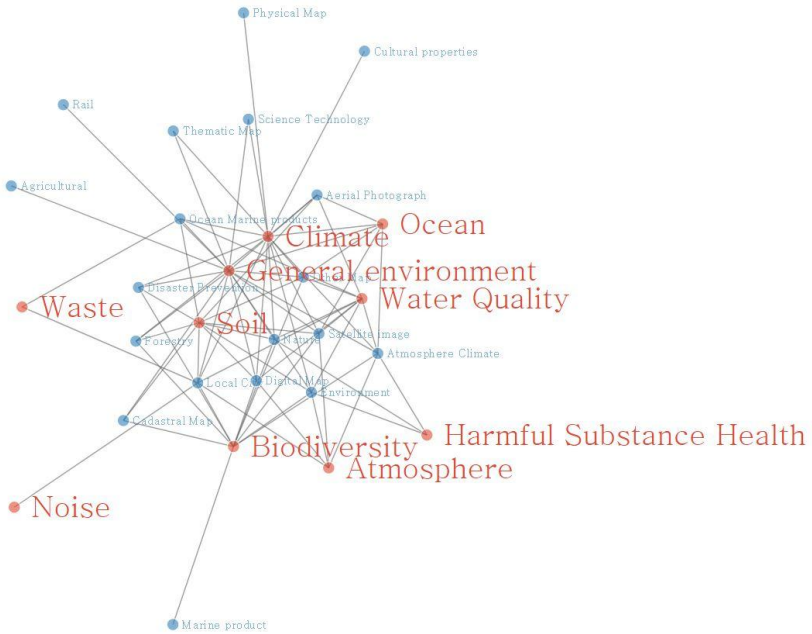

FIGURE 6. Association relation between each classification keyword

TABLE 4. Comparison of association relation between each classification keyword

Keyword	RHS	Support	Confidence	Lift
General environment (17)	Aerial photograph	0.012	0.056	0.933
	Agricultural	0.012	0.056	4.667
	Atmosphere climate	0.012	0.056	0.667
	Digital map	0.012	0.056	0.778
	Disaster prevention	0.012	0.056	1.167
	Environment	0.012	0.056	0.667
	Forestry	0.012	0.056	1.167
	Local city	0.012	0.056	0.583
	Nature	0.012	0.056	0.778
	Ocean/Marine products	0.012	0.056	1.556
	Other map	0.012	0.056	0.778
	Physical map	0.012	0.056	1.556
	Rail	0.012	0.056	4.667
	Satellite photograph	0.012	0.056	0.667
	Science Technology	0.012	0.056	2.333
	Thematic map	0.012	0.056	2.333
	Water resources	0.012	0.056	0.778
Climate (16)	Aerial photograph	0.012	0.059	0.988
	Atmosphere Climate	0.012	0.059	0.706
	Cultural properties	0.012	0.059	4.941
	Cadastral map	0.012	0.059	4.941
	Digital map	0.012	0.059	0.824
	Disaster prevention	0.012	0.059	1.235
	Environment	0.012	0.059	0.706
	Forestry	0.012	0.059	1.235
	Local city	0.012	0.059	0.618
	Ocean/Marine products	0.012	0.059	1.647
	Nature	0.012	0.059	0.824
	Other map	0.012	0.059	0.824
	Satellite photograph	0.012	0.059	0.706
	Science Technology	0.012	0.059	2.471
	Thematic map	0.012	0.059	2.471
	Water Resources	0.012	0.059	0.824
Biodiversity (11)	Atmosphere Climate	0.012	0.091	1.091
	Digital map	0.012	0.091	1.273
	Disaster prevention	0.012	0.091	1.909
	Environment	0.012	0.091	1.091
	Forestry	0.012	0.091	1.909
	Local city	0.012	0.091	0.955
	Nature	0.012	0.091	1.273
	Other map	0.012	0.091	1.273
	Physical map	0.012	0.091	2.545
	Satellite photograph	0.012	0.091	1.091
	Water resources	0.012	0.091	1.273
Soil (11)	Aerial photograph	0.012	0.091	1.527
	Digital map	0.012	0.091	1.273
	Disaster prevention	0.012	0.091	1.909
	Environment	0.012	0.091	1.091
	Forestry	0.012	0.091	1.909
	Local city	0.012	0.091	0.955
	Nature	0.012	0.091	1.273
	Other map	0.012	0.091	1.273
	Physical map	0.012	0.091	2.545
	Satellite photograph	0.012	0.091	1.091
	Water resources	0.012	0.091	1.273
Water Quality (9)	Aerial photograph	0.012	0.111	1.867
	Atmosphere Climate	0.012	0.111	1.333
	Digital map	0.012	0.111	1.556
	Environment	0.012	0.111	1.333
	Local city	0.012	0.111	1.167
	Nature	0.012	0.111	1.556
	Other map	0.012	0.111	1.556
	Satellite photograph	0.012	0.111	1.333
	Water resources	0.012	0.111	1.556
Atmosphere (5)	Atmosphere Climate	0.012	0.167	2.000
	Digital map	0.012	0.167	2.333
	Environment	0.012	0.167	2.000
	Local city	0.012	0.167	1.750
	Satellite photograph	0.012	0.167	2.000
Ocean (5)	Aerial photograph	0.012	0.200	3.360
	Atmosphere Climate	0.012	0.200	2.400
	Ocean/Marine products	0.012	0.200	5.600
	Other map	0.012	0.200	2.800
	Satellite photograph	0.012	0.200	2.400
Harmful Substance /Health (3)	Atmosphere Climate	0.012	0.333	4.000
	Environment	0.012	0.333	4.000
	Nature	0.012	0.333	4.667
Waste (2)	Local city	0.012	0.333	3.500
	Water resources	0.012	0.333	4.667
Noise (1)	Local city	0.012	1.000	10.500

서 제시된 그림 1의 결과와 일치되는 것이다. 그림 6은 프로그램 언어 R로 구현된 d3Network 패키지를 활용하여 도식화한 것이다. 이때, 붉은색으로 표현된 것은 환경 분야를 나타내며, 파란색으로 표현된 것은 환경공간정보를 나타낸다. 그림 6은 환경 분야와 환경공간정보 간에 생성된 연관 규칙을 시각화한 것으로 각각의 키워드 사이의 거리는 분석의 의미가 없다. 다만, 하나의 환경공간정보를 중심으로 한 환경 분야에 대한 연결 관계, 또는 하나의 환경 분야를 중심으로 한 환경공간정보의 연결 관계를 보여주는데 의미가 있을 것이다. 예를 들어, 환경 분야 중 "소음"에 대한 연구를 수행하는 기관이 타 환경 분야와 통합된 연구를 수행하고자 할 때, 환경공간정보 중 "지역도시"를 중심으로 "대기", "수질", "폐기물", "생물다양성", "토양" 분야와의 연계를 고려할 수 있다. 마찬가지로 환경공간정보 중, "대기/기후"를 생성하는 기관이 해당 자료의 질적 향상 및 고도화를 진행하고자 할 때, 환경 분야 중 "유해물질 및 보건", "대기", "수질", "생물다양성", "환경일반", "기후", "해양" 분야에서의 연구 수요를 조사할 수 있을 것이다. 즉, 표 4와 그림 6의 결과는 본 연구의 서론에서 언급한 바와 같이 환경공간정보의 융합 활용을 높이기 위한 가이드라인 제시로서의 의미가 있을 것이다.

결론 및 향후과제

본 연구에서는 텍스트 마이닝 기법을 활용하여 환경 분야별 활용된 환경공간정보의 빈도 분석, 시계열 분석, 연관분석을 수행하였다.

첫째, 빈도 분석 결과 환경 분야에서는 농업, 환경교육, 토지피복 등 환경 분야 일반적 내용을 포괄적으로 포함하는 "환경일반" 40.85%로 가장 높았지만, "기후" 분야는 전체의 32.8%로 "환경일반" 이외의 단일 분류체계에서는 가장 높은 비율을 나타냈다. "기후"가 포함하는 분야는 폭염, 산사태, 이상 기온 등 "환경일반" 보다는 일관된 주제를 지님으로, 환경 분야에서 가장 많은 관심과 연구가 진행된 분야는 "기후" 분야로 분석하는 것이 타당하다. 환경공간정보에서는 "위성영상" (24.87%)이 가장 높은 활용 빈도 나타냈는데 이는 환경 연구의 특성상 공간적 개념의 광의적 분석 및 시계열 분석이 매우 중요하며, "위성영상"이 이러한 요구를 가장 잘 충족하는 연구 자료이기 때문인 것으로 판단할 수 있다. 둘째, 환경 분야에 대한 시계열 분석에서 1996년부터 2000년까지는 "기후"에 대한 연구 비중이 높았으나, 2001년부터는 "환경일반"에 대한 연구가 증가한 것을 확인할 수 있다. 이는 2001년 이후 "농업", "환경 교육" 등 다양한 환경 분야에 대한 관심이 높아진 것과 연관지을 수 있다. 환경공간정보에서는 "위성영상"에 대한 수요가 전 기간에 걸쳐 가장 높았으며 활용 비율 또한 점차적으로 증가하고 있었다. 셋째, 연관분석에서는 총 80개의 연관 규칙이 생성되었다. 환경 분야에서 가장 많은 연관 규칙을 나타낸 것은 "환경일반"으로 17개의 규칙을 보였으며, 그 외에 "기후" 16개, "생물다양성" 11개, "토양" 11개, "수질" 9개, "대기" 5개, "해양" 5개, "유해물질 및 보건" 3개, "폐기물" 2개, "소음" 1개의 연관 규칙이 생성되었다.

본 연구에서는 텍스트 마이닝 기법을 활용하여 과거부터 현재까지 각 환경 분야 별 환경공간정보 활용의 변화를 정량적으로 분석하였다. 그러나 본 연구는 향후 2가지 개선사항이 있다. 첫째, 환경 전 분야를 연구의 범위로 설정하였기 때문에 환경 분야의 분류체계가 대분류의 단계에서 진행되었다. 비록, 환경공간정보의 활용 관점에서 전체의 환경 분야의 연구 경향의 파악에는 의미가 있으나 실질적 활용을 위해서는 보다 세분화된 연구가 요구된다. 예를 들어, "환경일반"에 포함되어 있는 농업, 환경교육 등 세분화된 분야를 대상으로 한 추가연구가 필요할 것이다. 둘째, 환경공간정보 또한 대분류의 단계에서 진행된 문제점이 있다. 예를 들어, 본 연구에서는 "위성영상"의 활용 비율의 가장 높았으나 해당 자료는 광학 영상, 레이더 영상, 적외선 영상, 초분광 영상 등 센서의 특징에 따라 또는 시간 및 공간해상도에 따른 세분화가 가능하다. 향후 추가적인 연구를 통하여 제시된 개선사항

을 보완한다면, 환경 분야별 환경공간정보의 융합 활용의 범위를 확장시킬 수 있는 기초 연구로써 그 의미가 있을 것으로 사료된다. KAGIS

주

1) 마인즈인사이트(http://www.mindsinsight.co.kr)는 ETRI(Electronics and Telecommunications Research Institute)에서 개발된 빅데이터 분석 도구로 신문기사, 블로그 등의 웹 자료를 기반으로 검색어에 대한 연관 키워드 등 선정하여 제공한다.

REFERENCES

Bae, K.Y., J.H. Park, J.S, Kim, and Y.S. Lee. 2013. Analysis of the abstracts of research articles in food related to climate change using a text-mining algorithm, Journal of the Korean Data & Information Science Society 24(6):1429-1437 (배규용, 박주현, 김정선, 이영섭. 2013. 텍스트 마이닝 기법을 활용한 기후변화관련 식품분야 논문초록 분석, 한국데이터정보과학회지 24(6):1429-1437).

Cho, H.G., Y.S. Kim, and S.E. Kim. 1998. A study on circulation and management of spatial data. Journal of the Korean Association of Geographic Information Studies 1(1):28-38 (조혜경, 김영섭, 김상은. 1998. 공간정보 유통 및 관리에 관한 연구. 한국지리정보학회지 1(1):28-38).

Feldman, R. and H. Hirsh. 1996. Mining associations in text in the presence of background knowledge. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. pp.343-346.

Feldman, R. and I. Dagan. 1995. Knowledge discovery in Textual Databases (KDT). Proceedings of the First International Conference on Knowledge Discovery and Data Mining. pp.112-117.

Hotho, A., A. Nürnberger, and G. Paaß. 2005. A brief survey of text mining. A brief survey of text mining 20(1): 19-62.

Kim, H.J. and J,Y. Chang. 2011. Discovering news keyword associations using association rule. The journal of the Institute of Internet Broadcasting and Communication 11(6):63-71 (김한준, 장재영. 2011. 연관 규칙 마이닝을 활용한 뉴스기사 키워드의 연관성 탐사. 한국인터넷방송통신학회 논문지 11(6):63-71).

Korea National Spatial Data Infrastructure Portal. http://www.nsdi.go.kr (Accessed June 1, 2016)

Lee, M.S., C.H. Lee, and J.Y. Kim. 2014. Big data analysis on demands for environmental policies. Korea Environment Institute. Research report. pp.47 (이미숙, 이창훈, 김지연. 빅데이터를 활용한 환경분야 정책수요 분석. 한국환경정책평가연구원. 연구보고서. 47쪽).

Lim, S.Y., M.S, Yi, G.H. Jin, and D.B, Shin. 2014a. A study on the research trends in the area of geospatial-information using text-mining technique focused on national R&D reports and theses. Journal of Korea Spatial Information Society 22(4):11-20 (임시영, 이미숙, 진기호, 신동빈. 2014. 텍스트 마이닝 기술을 이용한 공간정보 분야의 연구 동향에 관한 고찰 -국가연구개발사업 보고서 및 논문을 중심으로-. 한국공간정보학회지 22 (4):11-20).

Lim, S.Y., Y.M. Lim, and J.Y. Lee. 2014b. Study on the trends of U-City and smart city researches using text mining

technology. Journal of the Korean Society for Geospatial Information Science 22(3): 87-97 (임시영, 임용민, 이재용. 2014. 텍스트 마이닝 기법을 이용한 U-City와 Smart City의 연구 동향에 대한 분석. 한국지형공간정보학회지 22(3):87-97).

Mindsinsight. http://www.mindsinsight.co.kr. (Accessed November 1, 2016).

NDSL(National Digital Science Library). http://www.ndsl.kr(Accessed June 1, 2016).

Sakong, H.S. and K.H. Seo. 2007. A review on GIS research trends using content analysis method -focus on the GIS journals published from 1993 to 2006-. Journal of the Korean Association of Geographic Information Studies 10(3): 104-112 (사공호상, 서기환. 2007. 내용분석 기법을 활용한 GIS관련 연구 동향 분석 - 최근 14년(1993~2006)간 학회지 투고논문을 중심으로-. 한국지리정보학회지 10(3): 104-112).

Shin, H.S., O.J. Kwon, Y.D. Koo, Y.W. Shon, and Y.C. Bae. 2013. Scientometric analysis through linkage relation of keyword. The Journal of the Korea Institute of Electronic Communication Sciences 8(10):1467-1475 (신현식, 권오진, 구영덕, 손영우, 배영철. 2013. 키워드 연결 관계를 통한 계량정보 분석. 전자통신학회 8(10):1467-1475).

Oh. K.Y., M.J. Lee, and W.Y. No. 2016. A study on the improvement of sub-divided land cover map classification system: based on the land cover map by Ministry of Environment. Korean Journal of Remote Sensing 32(2):105-118 (오관영, 이명진, 노우영. 2016. 세분류 토지피복지도 분류체계 개선방안 연구: 환경부 토지피복지도를 중심으로. 대한원격탐사학회지 32(2):105-118).

Yun, H.C., K.S. Min, and M.G. Kim. 2010. Construction of multi-purpose hazard information map based on digital image using geospatial information. Journal of the Korean Association of Geographic Information Studies 13(3):91-101(윤희천, 민관식, 김민규. 2010. 지형공간정보를 활용한 수치영상기반의 다목적 재해정보지도 구축. 한국지리정보학회지 13(3): 91-101).