

# 동시출현단어분석을 통한 데이터과학 분야의 지적구조에 관한 연구\*

## A Study on the Intellectual Structure of Data Science Using Co-Word Analysis

김현정 (Hyunjung Kim)\*\*

### 초 록

최근 문헌정보학의 관련 분야로 주목받고 있는 데이터과학은 오랫동안 문헌정보학에서 해오던 정보의 수집, 저장, 조직, 분석, 활용 등의 활동을 데이터에 적용하여 그 가치를 이해하려는 학문이며, 통계학과 컴퓨터공학 등 다른 학문분야와의 연계가 필요한 분야이다. 이러한 데이터과학 분야의 연구 영역을 파악하기 위하여 동시출현단어 분석을 사용하여 Web of Science 핵심컬렉션에 수록된 문헌들 중 데이터과학 관련 자료들을 수집하고, 그 주제범주를 활용하여 네트워크분석을 실시하였다. 총 667건의 자료에 대한 159개의 주제범주를 기술분석하여 데이터과학 관련 연구가 많이 이루어지고 있는 학문분야를 조사하였고, 네트워크분석을 통해 데이터과학 분야 연구영역의 지적구조를 시각적으로 파악하였다. 분석결과, 데이터과학 분야의 연구들은 2개 영역 9개 군집으로 구분되었으며, 주제범주의 용어들 중 중심성이 높은 용어들을 통해 각 군집의 대표적인 주제들을 선정하였다. 연구의 결과는 데이터과학 분야의 연구들에 대한 지적구조를 파악하는데 도움이 될 수 있고, 문헌정보학과의 연계융합전공으로서의 데이터과학 교과과정 개발에 방향성을 제시할 수도 있을 것이다.

### ABSTRACT

Data Science is emerging as a closely related field of study to Library and Information Science (LIS), and as an interdisciplinary subject combining LIS, statistics and computer science in an attempt to understand the value of data by applying what LIS has been doing for collecting, storing, organizing, analyzing, and utilizing information. To investigate which subject fields other than LIS, statistics, and computer science are related to Data Science, this study retrieved 667 materials from Web of Science Core Collection, extracted terms representing Web of Science Categories, examined subject fields that are studying Data Science using descriptive analysis, analyzed the intellectual structure of the field by co-word analysis and network analysis, and visualized the results as a Pathfinder network with clustering created with the PNNC clustering algorithm. The result of this study might help to understand the intellectual structure of the Data Science field, and may be helpful to give an idea for developing relatively new curriculum.

키워드: 동시출현단어분석, 데이터과학, 지적구조, 네트워크분석  
co-word analysis, data science, intellectual structure, network analysis

\* 이 논문은 2017학년도 서울여자대학교 교내학술비의 지원을 받았음.

\*\* 서울여자대학교 사회과학대학 문헌정보학과 조교수(hk48@swu.ac.kr)

■ 논문접수일자: 2017년 11월 18일 ■ 최초심사일자: 2017년 12월 12일 ■ 게재확정일자: 2017년 12월 13일

■ 정보관리학회지, 34(4), 101-126, 2017. [http://dx.doi.org/10.3743/KOSIM.2017.34.4.101]

## 1. 서론

### 1.1 연구의 필요성

모든 학문분야는 지속적으로 크고 작은 변화를 겪는다. 어떤 주제는 많은 사람의 관심을 끌었다가 사라지기도 하고, 어떤 주제는 아주 오랫동안 많은 사람들의 연구 대상이 되기도 한다. 처음에는 한 오래된 학문분야의 세부 분야에 지나지 않았던 작은 주제가 수많은 연구자들의 학술활동에 의해 다양한 시각의 지식이 축적되어 독립적인 학문분야로 자리 잡기도 한다. 그것이 토마스 쿤이 말하는 과학혁명(Kuhn, 1996)의 결과든 아니든 간에, 하나의 학문분야가 여러 주제로 세분화되고, 그 중 하나의 주제가 또 다른 학문분야가 되는 변화는 매우 당연한 현상으로 받아들여진다. 더불어 기존의 여러 주제가 모여 새로운 학문분야를 이루게 되는 변화 역시 마찬가지이다. 최근 대학교육과정에서 문헌정보학과 경영학 등의 사회과학과 IT계열 간의 연계융합전공으로 데이터과학 또는 데이터 사이언스가 등장한 것 역시 이러한 변화를 보여주는 예라고 할 수 있다.

데이터과학은 이미 오랫동안 문헌정보학 분야에서 해오던 정보의 수집, 저장, 조직, 분석, 활용 등의 활동을 다양한 분야에서 방대하게 생성되고 있는 데이터에 적용하여 그 이해를 넓히려는 학문이다. Cervone(2015)은 데이터과학에 대한 소개에서 막대한 양의 데이터를 이해하기 위해 정보학뿐만 아니라 통계학, 컴퓨터공학 등의 학문분야가 연계하여 확률모형, 데이터 마이닝, 그리고 기계학습(machine learning)에 대한 지식을 이용하는 것이라고 설명하였고, 학

문분야로서의 데이터과학은 데이터를 수집, 처리, 저장하기 위한 시스템과 소프트웨어에 관련된 기술적 측면뿐만 아니라, 어떻게 하면 수집된 정보를 실제로 활용하여 현실적 문제를 해결하는데 적용할 수 있을지에 대한 학문적 접근을 포함하는 것이라고 하였다. 엄청난 양의 데이터를 빠른 속도로 수집할 수 있는 정보기술의 발달은 빅데이터라는 용어를 탄생시켰고, 이제는 데이터의 수집과 저장보다 데이터가 갖는 가치를 발견할 수 있는 능력이 더 중요시되고 있다. 따라서 데이터과학에 대한 관심은 하나의 특정한 분야가 아닌 여러 분야에서 나타나고 있으며, 이미 언급한 문헌정보학과 통계학, 컴퓨터공학 외에도 많은 학문분야들과 데이터과학의 관련성은 어느 정도 짐작할 수 있을 것이다. 이와 관련하여 국내 여러 대학에서는 데이터과학을 새로운 교과과정으로 수용하기 위해 다양한 시도를 하고 있다. 일부 대학에서는 독립된 과정으로 데이터과학 학과를 신설하거나, 기존의 학과에 연계융합전공으로 제공하거나, 또는 데이터과학 관련 연구소를 설치하기도 하였다. 그러나 아직까지는 데이터과학과 관련된 학문분야들에 대한 연구가 충분히 이루어지지 않아 실제로 어떠한 학문분야들이 데이터과학과 관련된 연구를 진행하고 있는지 파악되지 않는다. 따라서 본 연구에서는 데이터과학을 주제로 하는 연구들이 주로 어떤 학문분야에서 이루어지고 있는지 알아보기 위해 계량정보학적 연구방법을 적용하여 각 연구의 주제 분야를 나타내는 용어들을 분석하고, 이를 통해 데이터과학 분야의 지적구조를 살펴보고자 한다.

## 1.2 연구문제

이와 같은 연구목적에 위한 본 연구의 연구 문제는 다음과 같다. 첫째, 데이터과학을 주제로 하는 연구들은 주로 어떠한 학문분야에서 이루어지는지 조사하고자 한다. 이를 위해 본 연구는 Web of Science 데이터베이스에 수록된 자료들 중 “Data Science”를 주제로 포함하고 있는 자료들을 수집하여 각 연구의 주제 분야를 나타내는 용어들을 추출한 후, 기술분석을 통해 가장 많은 연구가 이루어진 분야, 가장 많은 자료를 수록하고 있는 학술지 등을 파악할 것이다. 둘째, 데이터과학을 주제로 하는 연구들이 수행되는 주제 분야들 간의 관계를 통해 데이터과학 분야의 지적구조를 파악하고자 한다. 이를 위해 추출된 용어들의 동시출현 빈도분석을 실시하여 연관된 주제 분야들을 나타내는 용어들의 네트워크를 생성하고, 클러스터링을 통해 세부 영역과 주제 분야들로 나누고, 그 결과를 네트워크지도로 시각화하여 데이터과학 분야의 지적구조를 살펴보고자 한다. 네트워크분석 과정에서 다양한 중심성 분석을 통해 전반적인 네트워크 안에서 가장 중심이 되는 주제 분야, 다른 주제 분야와의 매개 역할을 하는 주제 분야 등을 파악할 수 있다.

## 2. 관련연구

### 2.1 문헌정보학 교과과정과 데이터과학

데이터과학 또는 데이터 사이언스는 여러 분

야에 걸쳐 연구되어야 하는 융·복합적 주제 분야이며, 문헌정보학은 그 중 가장 큰 영향을 받은 분야 중 하나일 것이다. 이재운(2015)이 지적한 바와 같이 컴퓨터와 인터넷 등 정보기술의 발달은 문헌정보학계의 변화에 지속적으로 영향을 주었고, 데이터과학의 등장 역시 어떤 방식으로 수용하고 대처해야 할지 고민해야 할 주제이다. 특히 최근 몇몇 대학의 문헌정보학과에서 데이터과학을 교과과정을 일부로, 또는 연계융합전공으로 제공하기 시작하였는데, 이와 관련하여 강지혜(2016)의 연구에서는 69개 iSchool의 커리큘럼을 분석하고 국내 35개 문헌정보학과 교과과정과의 비교를 통해 한국형 데이터과학 교과과정 개발에 대한 방향성을 제시하였다. 해당 연구에 따르면 문헌정보학의 연구대상이 지식에서 정보로, 그리고 정보에서 데이터로 확대되고 있으며, 인접한 학문과의 협업 또한 확장되고 있다는 점을 알 수 있다. 더불어 데이터과학은 단순히 통계적 지식만을 강조하는 것이 아니라 통계적 사고를 활용하여 주제 분야의 데이터를 해석할 수 있는 능력을 키워야 하는 주제 분야이므로 수리적 분석 능력을 키우는 교과가 더욱 필요하며, 국내 사정을 고려하여 각 대학별로 특화된 분야에 집중하는 전략이 필요하다는 것을 제안하였다. 데이터 사이언스 교과과정에 대한 또 다른 연구(이명호, 2016)에서는 데이터 사이언스와 빅데이터 관련 학위를 제공하는 국내 7개 대학과 미국의 10개 대학을 비교하였는데, 그 중 열네 개 학교가 데이터과학을 대학원 과정으로 운영하고 있으며, 각 과정에서 제공되는 국내 181개 과목과 미국의 272개 과목을 분석한 결과 수학 및 통계 지식 영역 과목이 각 10%, 26%로 가장 많은 부분을 차지하

고 있는 것으로 나타났다.

Tang과 Sae-Lim(2016)의 연구는 미국에서 제공되고 있는 데이터 사이언스 프로그램에 대하여 내용분석을 통해 각 과정의 프로그램 설명과 교과과정의 구조, 그리고 각 과정에서 주목하고 있는 주제를 분석하였다. 해당 연구의 대상은 iSchool을 포함하여 예술과 과학 분야(Arts & Sciences), 경영학 분야(Business), 컴퓨터공학 분야(Computer Science), 공학 분야(Engineering), 수학/통계 분야(Mathematics/Statistics), 전문기과정 분야(Professional Studies) 등 여덟 개 학문분야에서 제공하고 있는 데이터와 데이터분석 관련 과목들이었으며, 각 분야마다 서너 개 대학원 과정 프로그램들을 표본으로 선택하였다. 연구에 의하면 데이터 사이언스 프로그램에 대한 설명은 데이터과학 학과가 별도로 설치된 경우 다음으로 iSchool이 가장 길었으며, 제공되는 과목들의 학점 수(credit hours)도 평균 71학점으로 연구 대상이었던 여덟 개 분야 중 가장 높았다. 특히 iSchool의 경우 핵심과목과 선택과목 모두에서 다른 분야들에 비해 주제적 지식(domain knowledge)에 대한 내용이 차지하는 부분이 가장 컸고, 정보기술(information skills)을 다루는 선택과목의 비중은 단연코 가장 높았다. 해당 연구의 경우 정보기술 외에도 의사소통기술(communication skills), 수학과 통계(mathematics and statistics), 그리고 시각화기술(visualization skills)을 데이터과학 과목에서 주목해야 할 내용으로 제안하였는데, 모든 분석의 결과 iSchool은 여덟 개 분야 중 가장 광범위한 핵심과목들을 제공하고 있으나 의사소통기술과 시각화기술을 주제로 다루는 선택과목이 매우 부족한 것으로

나타났다.

## 2.2 동시출현단어분석

동시출현단어분석은 텍스트의 덩어리(corpus) 안에 있는 단어 또는 명사구가 동시에 출현하는 양식을 이용하여 해당 텍스트가 나타내는 주제에서의 개념들 간의 관계를 식별해내기 위한 방법(He, 1999)이라고 할 수 있다. 개념들 간의 관계는 네트워크지도를 통해 모형화할 수 있으며(Danowski, 1993), 분석 결과로 얻어진 네트워크모형은 Pajek이나 NodeXL 등의 프로그램으로 시각화할 수 있다. 동시출현빈도는 언어 텍스트에서 키워드들간의 관계를 파악하는데 가장 많이 이용되는 방법이며, 그 관계의 행렬을 기반으로 네트워크 형태를 구성하여 분석할 수 있다(이수상, 2014). 최근 들어 동시출현단어 분석을 대체할 만한 네트워크분석방법으로 토픽 모델링방법이 주목받고 있으나, 동시출현단어 분석과 토픽모델링방법을 비교한 Leydesdorff와 Nerghe(2017)의 연구에서 용어의 수가 1,000개를 넘지 않는 소규모 또는 중규모 텍스트의 경우 동시출현단어분석방법이 더 적합하다는 것을 밝혀낸 바 있다.

용어들 간의 관계를 나타내는 네트워크지도는 해당 주제 분야의 지적구조를 파악하는 중요한 단서가 될 수 있다. 따라서 동시출현단어 분석을 이용하여 새롭게 나타난 주제 또는 학문분야의 지적구조 또는 연구동향을 파악하려는 연구가 많았는데, 서선경과 정은경(2013)은 오픈액세스 분야의 지적구조를 규명하기 위하여 Web of Science의 주제검색을 통해 479건의 학술지 논문을 수집하고 명사구 형태의 키

워드 8,643개를 추출하였으며, 동시출현단어분석의 결과를 네트워크분석과 군집분석에 적용하고 두 가지 분석 결과를 비교하는 방식으로 오픈액세스 분야의 세부 연구 영역에 대해 고찰하였다. 최예진과 정연경(2016)은 메타데이터 분야의 지적구조에 관한 연구에서 동시출현단어 분석을 이용하였는데, 역시 Web of Science의 주제검색을 통해 수집된 410건의 논문에서 1,137개의 저자키워드를 추출하고 그 중 37개의 키워드를 대상으로 네트워크분석과 군집분석을 실시하였다. 네트워크분석을 통해 2개 영역 9개 군집을, 군집분석을 통해 6개 군집을 도출하였고 그 결과를 다차원축적지도로 표현하였다. Cho(2014)의 연구에서는 기관 리포지터리 분야 연구에 대한 지적구조의 파악을 위해 Scopus 데이터베이스에서 수집한 자료들을 이용하여 해당 주제에 대한 여덟 개 군집을 도출하였는데, 각 군집을 대표하는 용어로는 메타데이터, 오픈액세스, 기관 리포지터리, 디지털 도서관, dSpace, 저작권, 보존(preservation), 그리고 시맨틱 웹 등이 선택되었다. 장령령과 홍현진(2014)은 독서분야의 지적구조 분석을 위해 학술지 등급과 키워드 위치를 고려한 혼합 가중치를 기반으로 하는 단어동시출현 분석기법을 이용하여, 독서분야를 4개의 큰 주제와 11개의 하위주제로 구분하였다. 김관준(2015a, 2015b)의 경우 디지털 큐레이션에 대한 연구동향 분석과 국외 독서 및 독서교육에 대한 연구동향 분석을 위해 문헌정보학 분야 학술 데이터베이스인 LISTA에 수록된 논문들의 디스크립터 프로파일링 기법을 사용하였으며, 이재윤과 김수정(2016)은 국내 재난 관련 연구 동향을 파악하기 위해 KCI 데이터베이스에 수록된 재난

관련 학술지 논문 772편에 대해 학문분야의 프로파일링 분석, 저널 프로파일링 분석과 함께 키워드 동시출현분석을 적용하였다. 이 밖에도 프로파일링 분석과 동시출현단어 분석을 함께 적용한 연구는 강범일과 박지홍(2013)의 한국어 교육학의 정체성 분석 연구, 그리고 강범일과 이재윤(2014)의 트위터 관련 연구에 대한 분석 연구 등이 있다.

동시출현단어분석은 이 외에도 다양한 주제에 적용되었는데, 도서관경영분야에 대한 연구(이정규, 이용구, 2013)와 연관영화정보에 대한 분석연구(최상희, 2014), 사회학 분야에 대한 연구(최형욱, 정은경, 2017) 등이 있으며 학술지 단위로 연구동향을 파악하고자 한 연구들(김선희, 윤순진, 2015; 김하진, 송민, 2014; 허고은, 송민, 2013; Ravikumar, Agrahari, & Singh, 2015)도 있다. 특히 최근 몇 년 동안 중국의 학자들이 동시출현단어분석을 이용하여 문헌정보학 분야의 지적구조를 파악하려는 연구가 다수 이루어진 것을 볼 수 있는데, 디지털 도서관 분야에 대한 분석(Liu, Hu, & Wang, 2012)과 문헌정보학 분야 전반에 대한 분석(Hu, Hu, Deng, & Liu, 2013; Zong, Shen, Yuan, Hu, Hou, & Deng, 2013)을 비롯하여 소셜 미디어에 대한 연구(Gan & Wang, 2015)와 마케팅 분야에 대한 연구(Wang, Zhao, & Wang, 2015), 그리고 동시출현단어 분석 자체에 대한 연구들(Wang, Li, Li, & Li, 2012; Yang, Wu, & Cui, 2012) 등이 Scientometrics 저널에 수록되었다.

### 3. 연구방법

#### 3.1 데이터 수집

데이터과학 분야의 지적구조를 파악하기 위하여 Web of Science 핵심컬렉션에 수록된 문헌을 대상으로 주제(Topic)에 대한 검색을 실시하였다. Web of Science의 핵심컬렉션에는 Science Citation Index Expanded(SCIE), Social Science Citation Index(SSCI), Arts & Humanities Citation Index(AHCI), 그리고 Emerging Sources Citation Index(ESCI) 등의 인용색인이 포함되어 있으며, 2006년 1월 1일 자료부터 2017년 11월 24일 업데이트된 자료까지를 그 대상으로 하였다. 단, ESCI의 경우 해당 서비스가 시작된 시점인 2015년부터 포함되었다.

Web of Science의 기본검색 중 주제를 대상 검색필드로 설정할 경우 문헌의 제목, 초록, 저자 키워드, 그리고 인용 논문의 제목에서 생성된 색인 용어인 Keywords Plus까지 포함하여

검색하게 된다. 주제검색의 질의어로는 “Data Science”를 이용하였으며, 따옴표를 사용하여 정확히 일치하는 구문을 검색하였다. 데이터 과학은 비교적 새롭게 등장한 주제이고, 본 연구는 해당 주제가 어떤 분야에서 주로 연구되고 있는지 알아보기 위한 탐색적 연구에 해당하므로 문서유형이나 언어에 별도의 제한을 설정하지 않았다. 최종적으로 총 667건의 자료가 검색되었으며, 문서유형별로 살펴보면 504건이 학술지 논문(article), 77건이 논평(editorial material), 61편이 검토(review), 10건이 논문집 자료(proceedings paper) 등으로 이루어져 있다. 문서유형에 따른 검색결과의 구성은 <표 1>과 같다.

문서유형을 학술지 논문으로 제한하여 검색을 실시한 후 다시 그 결과에 대한 문서유형의 구성을 살펴보면 학술지 논문이면서 동시에 논문집 자료 또는 도서 장에 속하는 경우가 있어 중복 분류가 되어있음을 알 수 있다. 따라서 문서유형별 레코드수의 총합은 664건보다 많은

<표 1> 문서유형별 검색결과의 분포

문서유형	빈도수	비율(%)	설명
Article	504	75.562	논문
Editorial Material	77	11.544	논평
Review	61	9.145	검토
Proceedings Paper	10	1.499	논문집 자료
Book Chapter	9	1.349	도서 장
Meetings Abstract	7	1.049	회의 초록
Letter	6	0.9	서한
Book Review	5	0.75	서평
New Item	5	0.75	뉴스 항목
Correction	1	0.15	수정
Software Review	1	0.15	소프트웨어 검토

686건으로 계산된다. 가장 많은 유형인 학술지 논문 외에도 사설이나 인터뷰, 토론 등을 포함하는 논평자료도 비교적 많은 편이며 “이전에 연구된 자료의 갱신된 연구(a renewed study of material previously studied)”(Web of Science, 2017)라고 소개되는 검토(Review)자료도 적지 않다. 또한 기존의 인용색인에 아직 포함되지 않은, 상대적으로 새로운 주제를 다루는 자료들을 포함하고 있는 ESCI에서만 91건의 자료가 검색되어 아직은 많은 토의가 이루어지고 있는 새로운 주제 분야임을 알 수 있다.

검색된 667건의 자료의 주제범주(Web of Science Categories)를 추출하여 동시출현단어 분석에 이용하였다. Web of Science에서는 총 252개의 주제범주로 자료들을 구분하고 있으며, 자료를 색인할 때 각 자료마다 하나 또는 그 이상을 할당하고, 많게는 다섯 개 범주까지 할당한 경우를 볼 수 있다. 따라서 하나의 연구에 동시에 할당된 주제범주들은 주제적으로 연관성이 있다는 의미가 된다. “Engineering(공학)”이나 “Psychology(심리학)” 등 일부 주제 분야들은 여러 개 하위분야로 세분화되어 있는데, 예를 들어 “Computer Science(컴퓨터공학)”는 “Artificial Intelligence(인공지능)”, “Cybernetics(사이버네틱스)”, “Hardware & Architecture(하드웨어 및 아키텍처)”, “Information Systems(정보시스템)”, “Interdisciplinary Applications(학제간 응용)”, “Software Engineering(소프트웨어 공학)”, “Theory & Methods(이론 및 방법)” 등의 일곱 분야로 분류되어 있다. Web of Science에서는 썬표를 이용하여 세분화된 주제를 표시하고 있으며, 상위주제를 먼저 표기한 후 썬표 뒤에 하위주제를 나열한다. 예를 들

어, 컴퓨터 공학의 인공지능 분야는 “Computer Science, Artificial Intelligence”로 표기하는 방식이다. 본 연구에서는 데이터과학과 관련된 주제 분야를 상세히 살펴보기 위해 세분화된 주제를 각각 다른 주제로 취급하여 분석하였다. <표 2>는 출현빈도가 12회 이상인 25개의 Web of Science 주제범주 리스트이다.

빈도수 상위 25위까지의 주제범주들을 살펴보면, 컴퓨터공학 분야의 세부주제들이 다수 포함되어 있으며, “Statistics & Probability(통계 및 확률)”, “Multidisciplinary Science(다학문 분야 과학)”, “Information Science & Library Science(정보과학 및 도서관학)”, “Materials Science, Multidisciplinary(재료과학, 다학문 분야)”, “Engineering, Electrical & Electronic(공학, 전기 전자)”, “Management(경영학)”, “Biotechnology & Applied Microbiology(생명공학 및 응용 미생물학)”, “Neurosciences(신경과학)”, “Food Science & Technology(식품과학 및 기술)” 등 다양한 학문분야가 포함되어 있다. 특히 상위 25위 안에 드는 컴퓨터공학 관련 분야들의 빈도수는 전체의 42.6%(284건)에 해당한다. 검색결과인 667건의 자료에서 총 1,125개의 주제범주가 추출되었고, 중복데이터를 제거한 후 남은 159개의 고유한 데이터를 동시출현단어 빈도분석의 대상으로 선정하였다.

### 3.2 동시출현단어 행렬 작성과 네트워크 분석

선정된 용어들의 리스트와 해당 용어들이 문헌에 나타난 정보를 기록한 리스트를 이제윤이

〈표 2〉 출현빈도 12회 이상의 데이터과학 관련 Web of Science 주제범주 리스트

순위	Web of Science 주제범주	빈도수	비율(%)
1	Computer Science, Interdisciplinary Applications	78	11.694
2	Computer Science, Artificial Intelligence	59	8.846
3	Computer Science, Information Systems	57	8.546
4	Computer Science, Theory Methods	53	7.946
5	Statistics & Probability	53	7.946
6	Multidisciplinary Sciences	51	7.646
7	Information Science & Library Science	32	4.798
8	Materials Science, Multidisciplinary	31	4.648
9	Computer Science, Software Engineering	24	3.598
10	Engineering, Electrical & Electronic	21	3.148
11	Management	18	2.699
12	Biotechnology & Applied Microbiology	17	2.549
13	Mathematics, Interdisciplinary Applications	16	2.399
14	Neurosciences	16	2.399
15	Food Science & Technology	14	2.099
16	Nursing	14	2.099
17	Operations Research & Management Science	14	2.099
18	Public Environmental & Occupational Health	14	2.099
19	Biochemistry, Molecular Biology	13	1.949
20	Computer Science, Hardware Architecture	13	1.949
21	Environmental Sciences	13	1.949
22	Medical Informatics	13	1.949
23	Engineering, Industrial	12	1.799
24	Pharmacology & Pharmacy	12	1.799
25	Telecommunications	12	1.799

개발한 COOC Version 0.4에 입력하여 동시출현단어 행렬과 이를 정규화한 코사인 연관성 행렬, 그리고 피어슨 상관계수 행렬을 생성하였다. 코사인 계수의 경우 0에서 1 사이 값을 갖는데, 연관성이 있는 것으로 나타난 96쌍의 주제범주들 중 여섯 쌍은 유사도가 1이었으며 0.03501로 가장 적은 유사도를 보인 경우는 “Multidisciplinary Sciences(다학문분야 과학)”와 “Mathematics, Interdisciplinary Applications(수학, 학제간 응용)”였다. “Information Science & Library

Science(정보과학과 도서관학)” 범주의 경우 “Computer Science, Information Systems(컴퓨터공학, 정보시스템)”와의 유사도가 0.23415이고 “Education & Educational Research(교육 및 교육연구)” 범주와의 유사도가 0.0559인 것으로 나타났다.

피어슨 상관계수는 두 주제범주가 제 3의 주제범주와 동시에 출현한 값을 계산한 것이므로 간접적 영향력을 나타내는 2차 연관성 값이며, -1에서 1의 값을 갖는다. 1에 가까울수록 강한

긍정적 관계를 의미하며, -1에 가까울수록 강한 부정적 관계를 의미한다. 피어슨 상관계수 역시 코사인 연관계수가 1이었던 여섯 쌍의 주제범주가 1의 유사도를 보여주었으며, 0.98034의 가장 강한 긍정적 관계를 갖는 경우는 “Obstetrics & Gynecology(산과 및 부인과 의학)”와 “Reproductive Biology(생식 생물학)”이었다.

코사인 연관성 행렬과 피어슨 상관계수 행렬은 이재윤이 개발한 WNET Version 0.4.x를 통해 패스파인더 네트워크(PFNet: Pathfinder Network)와 병렬 최근접 이웃 클러스터링(PNNC: Parallel Nearest Neighbor Clustering), 그리고 가중 네트워크 중심성(Weighted Network Centralities)을 구하는데 이용되었다. PFNet 알고리즘은 노드들 간에 최소한의 가중치를 갖는 경로를 찾는 것으로,  $r$ 과  $q$ 의 두 변수가 필요한데,  $r$ 은 경로의 거리를 계산하는 민코프스키 측정법을 정의하는 변수이며,  $q$ 는 최종적으로 그려질 최소한의 가중치 경로의 범위를 제한하는데 쓰인다(White, 2003). 따라서  $r$ 이 커질수록 경로의 거리는 짧아지고,  $q$ 가 커질수록 조사 대상인 경로의 수가 늘어나므로 패스파인더 네트워크로 학문분야의 지적구조를 표현하기 위해서는  $r = \infty$ ,  $q = n-1$ 로 설정하여야 한다(이재윤, 2006). 패스파인더 네트워크 알고리즘은 중요한 링크가 존재하는 노드들만을 남기는 방식으로 네트워크를 표현하는 것이므로 전체 구조를 이해하는데 매우 유용한데, 세부 주제 분야를 파악하기에는 부족한 면이 있다. 따라서 데이터 과학 분야의 자료들에 할당된 Web of Science 주제범주들을 적당한 군집으로 나누어 세부 주제 분야를 파악할 수 있는 방법으로 PNNC 기법을 활용하였다.

## 4. 분석결과

### 4.1 연도별, 국가별, 학술지별 분석

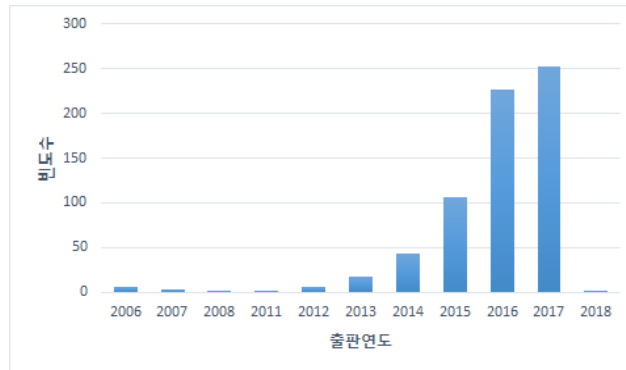
검색결과를 출판연도별로 살펴보면, 2006년의 6건에 비해 2017년에 252건의 자료가 검색되는 등, 지난 10여 년간 꾸준히 증가하는 추세에 있음을 알 수 있다. 특히 2015년에 폭발적으로 증가하기 시작하여 2016년에는 이전 해에 비해 두 배로 늘어났으며, 2017년에도 그 추세를 유지하였다. 또한 2018년에도 한 건이 있는 것으로 검색되었는데, 이는 Computers & Structures라는 학술지에 실린 논문(Nguyen & Keip, 2018)으로 해당 학술지의 경우 한 달에 두 번씩 발행되며, 2017년 11월 24일 현재 2018년 1월 15일 자 이슈까지 발행되어 있는 것을 볼 수 있다. <표 3>과 <그림 1>은 검색결과와 출판연도별 분포를 보여주는 표와 그래프이다.

언어에 대한 제한을 두지 않고 검색을 실시하였으나 667건의 자료 중 영어로 된 자료가 660건이었고, 러시아어로 된 자료 두 건을 제외하면 나머지 다섯 건은 중국어, 독일어, 헝가리어, 일본어, 포르투갈어로 각 한 건씩 검색되었다. 그에 비해 저자들의 출신국가 또는 지역별 분포는 매우 다양하게 나타났는데, 총 65개 국가 또는 지역 중 미국과 영국이 각각 53.5%와 13.8%를 차지하고 그 외에 중국, 독일, 호주, 프랑스, 스페인, 이태리, 인도, 일본 순으로 나타났다. 한국의 경우 7편의 자료가 검색되어 18위를 차지하였다. <표 4>는 10건 이상 수록된 상위 14개 국가별 분포를 나타내는 표이다.

데이터과학 관련 자료 667건이 발행된 학술지들을 살펴보면, 통계학 관련 학술지인 Statistical

〈표 3〉 출판연도별 분포

출판연도	빈도수
2006	6
2007	4
2008	1
2011	2
2012	6
2013	18
2014	44
2015	106
2016	227
2017	252
2018	1



〈그림 1〉 출판연도별 분포

〈표 4〉 국가/지역별 분포

국가 또는 지역	빈도수	비율(%)
미국	357	53.523
영국	92	13.793
중국	53	7.946
독일	42	6.297
호주	24	3.598
프랑스	24	3.598
스페인	23	3.448
이탈리아	22	3.298
인도	19	2.849
일본	19	2.849
스코틀랜드	19	2.849
캐나다	16	2.399
네덜란드	16	2.399
이스라엘	10	1.499

Analysis and Data Mining에 가장 많은 논문이 수록되었고, 단 한 편씩 실려 있는 학술지가 무려 348종으로, 아직까지는 많은 주제에서 다양하게 다뤄지고 있다는 것을 알 수 있다. 다섯 편 이상 수록하고 있는 학술지 16종을 살펴보면, 통계학 관련 학술지 2종(Statistical Analysis and Data Mining, American Statistician), 빅데이터 관련 학술지 2종(Big Data, Big Data Society), 정보학 관련 학술지 2종(Information

Sciences, International Journal of Information Management), 재료 관련 2종(Integrating Materials and Manufacturing Innovation, Acta Materialia), 식품 관련 1종(Agro FOOD Industry Hi Tech), 컴퓨터 시스템 관련 1종(Future Generation of Computer Systems), 커뮤니케이션 관련 1종(Communications of the ACM), 그리고 PLOS One이나 Royal Society Open Science, IEEE Access 등과 같이 과학기

〈표 5〉 학술지별 데이터과학 관련 논문 출판 건수

순위	학술지명	빈도수
1	Statistical Analysis and Data Mining	28
2	Big Data	18
3	Philosophical Transactions Of The Royal Society A: Mathematical, Physical and Engineering Sciences	14
4	Agro FOOD Industry Hi Tech	11
5	PLOS One	9
6	Integrating Materials And Manufacturing Innovation	8
7	American Statistician	7
8	Royal Society Open Science	7
9	Communications Of The ACM	6
10	Information Sciences	6
11	Proceedings of the National Academy of Sciences of The United States of America	6
12	Acta Materialia	5
13	Big Data Society	5
14	Future Generation Computer Systems: The International Journal Of eScience	5
15	IEEE Access	5
16	International Journal of Information Management	5

술 전 분야를 다루는 학술지들로 이루어져 있다.

총 430종의 학술지 중 데이터과학 관련 논문이 10건 이상 수록된 학술지는 총 4종밖에 되지 않고, 문헌정보학 분야의 주요 학술지 중 하나인 Journal of the Association for Information Science and Technology의 경우 세 편의 논문이 수록되어 있어 상위 16종 안에는 포함되지 않았다. 〈표 5〉는 데이터과학 관련 논문들이 다섯 편 이상 수록되어 있는 학술지들의 리스트이다.

## 4.2 네트워크 분석 결과

### 4.2.1 중심성 비교

데이터과학을 주제로 하는 667건의 자료에 할당된 159개의 Web of Science 주제범주에 대한 동시출현빈도 행렬을 코사인 계수로 정규화한 후 각 용어의 중심성을 산출하였다. 네트워크분석에 있어서 단순한 이진네트워크가 아닌 가중네트워크의 경우 기존의 중심성지수와

는 다른 방법으로 산출하는 중심성지수가 필요한데, 대표적인 가중네트워크의 중심성지수는 최근접이웃중심성, 평균연관성, 평균프로파일연관성, 삼각매개중심성 등이 있다(이재운, 2013). 이 중 평균연관성과 삼각매개중심성은 각각 이진네트워크의 근접중심성과 매개중심성의 대안 개념으로 제안된 것으로, 네트워크 전반에 영향을 미치는 전역중심성 지수이다. 평균연관성은 한 노드가 다른 노드들과 가진 가중치의 평균값으로, 네트워크의 밀도가 높은 경우 중심성을 측정하기에 적합한 지수이며, 삼각매개중심성은 다른 노드들 사이를 매개시켜주는 정도를 측정하는 값으로, 본 연구에서는 이를 0과 1 사이의 값으로 정규화한 상대적 삼각매개중심성을 비교하였다. 다음은 코사인 유사도에 기반한 상대적 삼각매개중심성과 평균연관성이 높은 상위 15위 주제범주의 리스트이다.

상대적 삼각매개중심성과 평균연관성 모두가 가장 높은 주제범주는 “Computer Science,

<표 6> 데이터과학 분야 주제범주 159개에 대한 가중 네트워크의 전역중심성 지수 상위 15위

Web of Science 주제범주	상대적 삼각매개 중심성 (rTBC:0~1)	Web of Science 주제범주	평균연관성 (AVGSIM)
Computer Science, Interdisciplinary Applications	0.02693	Computer Science, Interdisciplinary Applications	0.01819
Statistics & Probability	0.01000	Engineering, Environmental	0.01641
Computer Science, Information Systems	0.00492	Materials Science, Multidisciplinary	0.01625
Neurosciences	0.00468	Geography	0.01441
Computer Science, Artificial Intelligence	0.00411	Health Care Sciences & Services	0.01420
Operations Research & Management Science	0.00379	Social Sciences, Biomedical	0.01381
Engineering, Environmental	0.00347	Environmental Studies	0.01363
Mathematics, Interdisciplinary Applications	0.00323	Neurosciences	0.01353
Medical Informatics	0.00323	Environmental Sciences	0.01317
Materials Science, Multidisciplinary	0.00306	Mathematics, Interdisciplinary Applications	0.01298
Information Science & Library Science	0.00258	Statistics & Probability	0.01271
Health Care Sciences & Services	0.00242	Physics, Applied	0.01262
Computer Science, Theory & Methods	0.00226	Operations Research & Management Science	0.01237
Engineering, Industrial	0.00218	Ethics	0.01237
Geography	0.00194	Engineering, Manufacturing	0.01213

Interdisciplinary Applications(컴퓨터공학, 학제간 응용)”으로 나타났다. 이는 해당 분야가 데이터과학 분야에서 다른 주제범주들과 광범위하게 연관되어 있음을 의미한다. 두 가지 중심성 척도 모두에서 상위 15위 안에 포함된 주제범주는 아홉 개로, “Computer Science, Interdisciplinary Applications(컴퓨터공학, 학제간 응용)”, “Engineering, Environmental(공학, 환경)”, “Geography(지리학)”, “Health Care Sciences & Services(보건과학 및 서비스)”, “Materials Science, Multidisciplinary(재료과학, 다학문분야)”, “Mathematics, Interdisciplinary Applications(수학, 학제간 응용)”, “Neurosciences(신경과학)”, “Operations Research & Management Science(오퍼레이션 리서치 및 경영과학)”, “Statistics & Probability(통계 및 확률)” (이상 알파벳순) 등으로 나타났다. 컴퓨터공학이나 통계학, 경영학 관련 분야들의 경우

데이터과학을 소개하는 많은 논문들이 주요 관련 분야로 언급한 바 있으며, 이 외에도 환경과 지리분야, 수학분야, 신경과학과 보건과학 등 자연과학 계열의 다양한 분야에서 데이터과학 연구가 진행되고 있음을 알 수 있다. 또한 상위권에 있는 많은 범주들이 다학문적인 연구 또는 학제간 응용연구인 것에도 주목할 필요가 있다.

주제범주의 영향력이 전체 네트워크에 미치는 범위를 파악할 수 있는 상대적 삼각매개중심성만 살펴보면 가장 상위에 위치한 “Interdisciplinary Applications(학제간 응용)” 외에도 “Information Systems(정보시스템)”, “Artificial Intelligence(인공지능)”, “Theory & Methods(이론 및 방법)” 등 컴퓨터공학 관련 세부 주제 분야들이 다수 포함되어 있다. 이는 데이터과학에 관련된 여러 분야들 중 컴퓨터공학 분야가 가장 많은 영향을 미치는 중심적인 분야임을 의미한다. 한 노드가 다른 노드들과의 가까운 정도를 평균적

으로 계산한 평균연관성의 경우 “Engineering, Environmental(공학, 환경)”, “Environmental Studies(환경 연구)”, “Environmental Sciences(환경 과학)” 등 환경관련 분야 연구가 다수 상위권에 포함되어 있어, 해당 주제 분야가 네트워크 전체에서 다른 분야들과 상당히 많이 연관되는 분야임을 알 수 있다.

#### 4.2.2 패스파인더 네트워크와 병렬 최근접 이웃 클러스터링을 이용한 군집분석

코사인 연관성 행렬을 이용하여 생성된 패스파인더 네트워크에서 연관성이 있는 것으로 나타난 96쌍의 주제범주들을 통해 최종적으로 링크가 96개이고 노드가 134개 주제범주인 네트워크를 생성하였다. 결과적으로 동시출현빈도 분석을 실행한 159개 주제범주 중 25개 범주가 다른 주제범주와 연관성이 없는 것으로 밝혀져 네트워크분석의 대상에서는 제외되었다. 병렬 최근접 이웃 클러스터링을 적용한 결과 2개 영역과 9개 군집, 45개의 세부 군집이 생성되었는데, 세부 군집의 경우 “Biochemistry & Molecular Biology(생화학 및 분자 생물학)”, “Cell Biology(세포 생물학)”, 그리고 “Biophysics(분자 생물물리학)”와 같이 주제 연관성이 매우 높은 분야들이 군집화된 경우가 많아, 본 연구에서는 두 번째 단계인 아홉 개 군집까지를 집중적으로 살펴보기로 하였다.

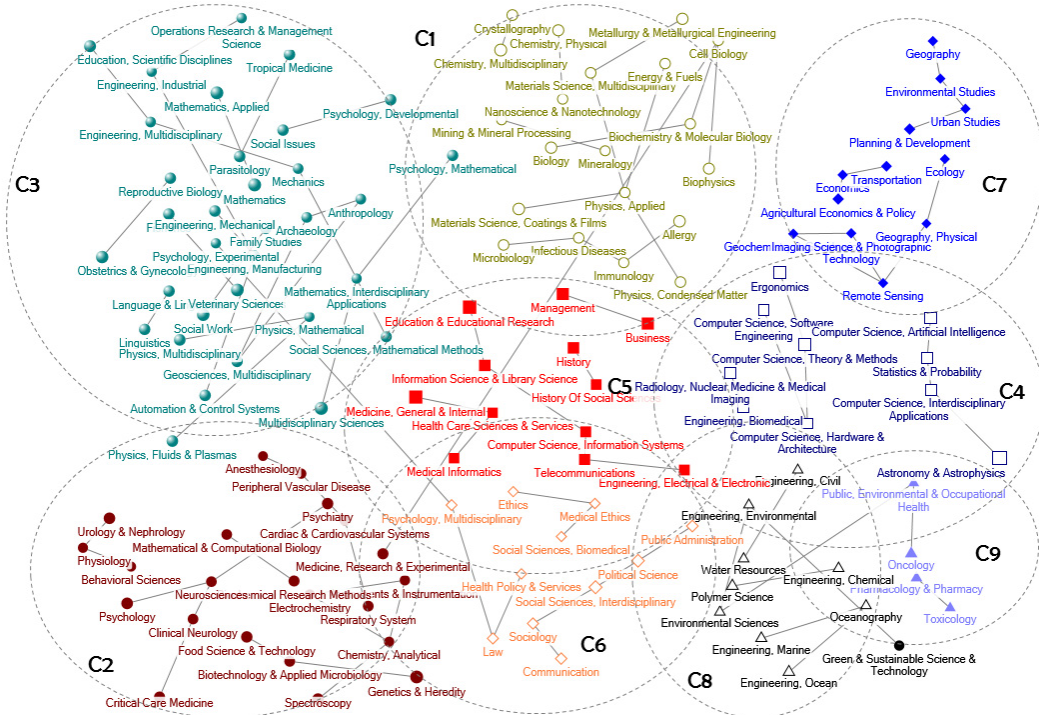
군집화한 결과를 살펴보면 중 네트워크분석의 대상에서 제외된 25개 주제범주의 경우 총 45개 세부 군집 중 하나의 군집을 이루는 용어들임을 알게 되었다. 대부분의 세부 군집은 세계에서 여섯 개 정도의 주제범주로 이루어진 것에 비해 제외된 군집의 경우 상대적으로 많

은 주제범주가 포함되어 있었으나, 다른 주제범주들과의 유사도가 0으로 나타나 네트워크분석의 대상에서 제외할 수밖에 없었다. 나머지 134개 주제범주의 영역별/군집별 목록은 [부록 1]에 첨부하였다. 굵은 글씨로 표기된 주제범주는 각 군집 내에서 상대적 삼각매개중심성이 가장 높은 주제범주를 표시한 것이다.

〈그림 2〉는 동시출현빈도의 코사인 유사도 행렬을 이용하여 생성된 패스파인더 네트워크( $q=n-1, r=\infty$ )를 NodeXL로 시각화한 것이다. 병렬 최근접 이웃 클러스터링(PNNC) 알고리즘을 이용하여 생성된 아홉 개 군집을 네트워크지도에 적용하였고, 각 군집에 포함된 노드의 색깔과 모양으로 구분하였다. 각 군집별로 적게는 4개, 많게는 33개의 주제범주가 포함되어 있으며 평균 개수는 14.8개이고, 각 군집의 노드 수와 그래프 밀도를 살펴보면 〈표 7〉과 같다.

가장 많은 노드수를 갖고 있는 세 번째 군집(C3)의 경우 그래프 밀도는 가장 낮았는데, 수학과 관련된 주제범주들과 공학 분야의 다양한 세부 주제들, 그리고 인류학, 고고학, 언어학 등 수학과나 공학과는 관련이 적은 분야들로 이루어져 있어 주제범주들 간 연관성이 높지 않은 것으로 보인다. 반면 네 번째 군집(C4)의 경우 노드 수는 열 개로 적은 편이나 그래프 밀도가 가장 높으며, 컴퓨터공학 분야의 다수 세부 주제들과 통계 및 확률 등 데이터과학의 주요 관련 주제범주들이 포함되어 있다.

총 134개의 주제범주를 PNNC 알고리즘으로 클러스터링한 결과 생성된 2개의 영역과 9개의 군집을 자세히 살펴보면 다음과 같다. [부록 1]에 첨부한 영역별/군집별 주제범주 목록



〈그림 2〉 데이터과학 분야 Web of Science 주제범주의 동시출현빈도 네트워크

〈표 7〉 군집별 노드수와 그래프 밀도

영역	군집	노드수	그래프 밀도
A	C1	20	0.074
	C2	22	0.069
	C3	33	0.042
	C4	10	0.156
	C5	12	0.121
	C6	11	0.145
B	C7	12	0.152
	C8	10	0.139
	C9	4	0.333

은 각 군집별로 상대적 삼각매개중심성이 높은 주제범주에서 낮은 주제범주로 나열되어 있으므로, 상위에 기록된 주제범주들 위주로 살펴보면 각 군집의 특성을 파악할 수 있다. 첫 번째 영역에는 C1군집에서 C6군집까지 여섯 군집이

해당되며, 두 번째 영역은 C7군집부터 C9군집까지 세 군집이 포함되어 있다.

첫 번째 영역 안에는 통계학이나 컴퓨터공학, 그리고 경영학 등 중심성이나 평균연관성 등이 높은 주제들이 다수 포함되어 있어, 데이

터과학 분야와 상당히 높은 관련이 있는 분야 들임을 알 수 있다. 첫 번째 군집(C1)의 경우 주로 화학, 생물학, 물리학, 재료과학 등의 주제를 포함하고 있고, 두 번째 군집(C2)에 포함된 주제범주들은 신경과학, 생화학, 생명공학, 미생물학, 생리학 등과 정신의학, 비뇨기과학, 신장학 등 의학 분야의 세부 주제들이다. 세 번째 군집(C3)에는 경영과학과 수학 관련 주제들과 산업공학, 제조공학 등 공학 계열 주제들이 포함되었다. 네 번째 군집(C4)과 다섯 번째 군집(C5)에는 데이터과학의 가장 큰 부분을 차지하는 컴퓨터공학 부분이 포함되어 있는데, 그 중에서도 네 번째 군집에는 인공지능, 소프트웨어공학, 하드웨어 및 아키텍처 등의 컴퓨터공학 분야와 통계 및 확률이 자리잡고 있으며, 다섯 번째 군집은 정보시스템과 의료 정보학, 정보과학 및 도서관학 등 정보학 관련 주제들과 경영학과 비즈니스 등 경영학 관련 분야들

이 포함되어 있다. 여섯 번째 군집(C6)에는 사회학, 윤리와 법률 등 사회과학 관련 분야들이 있다.

두 번째 영역에 해당하는 일곱 번째 군집(C7)에는 지리학과 생태학, 그리고 환경 관련 분야가 포함되어 있으며, 여덟 번째 군집(C8) 역시 환경과학과 토목, 해양 등의 주제를 다루고 있다. 마지막 군집인 아홉 번째 군집(C9)의 경우 네 개의 주제범주만 포함되어 있는데, 주로 공중 위생과 약학, 종양학 등에 관련된 주제들이다. 다음은 아홉 개 군집의 명칭과 주요 주제범주를 간단한 표로 정리해놓은 것이다. 주요 주제범주 예시 중 이탤릭체로 표기한 주제범주들은 각 군집 내에서 상대적 최근접이웃 중심성이 높은 경우이다.

각 군집 내에서 지역중심성인 상대적 최근접이웃중심성이 가장 높고 전역중심성인 상대적 삼각매개중심성도 비교적 높은 범주(〈표 6〉참

〈표 8〉 군집명칭과 군집별 주요 주제범주 목록

영역	군집	군집명칭	주요 주제범주 예시
A	C1	화학, 생물학, 물리학	<b>응용물리학</b> , 재료과학, 생화학 및 분자생물학, 세포 생물학, 분자생물 물리학, 고체물리학, 에너지 및 연료
	C2	신경과학, 생명공학, 의학	<b>신경과학</b> , 신경학, 생화학, 생명공학 및 응용 미생물학, 생리학, 의학, 중환자 의학, 정신의학, 신장학
	C3	수학, 공학, 경영과학	<b>응용수학</b> , 수학, 경영과학, 산업공학, 제조공학, 기계공학
	C4	컴퓨터 공학, 통계 및 확률	<b>하드웨어 및 아키텍처, 통계 및 확률</b> , 컴퓨터 공학, 인공지능, 소프트웨어 공학, 생물의학공학, 영상의학, 핵의학 및 의료영상
	C5	정보과학, 정보시스템, 경영학	<b>보건과학 및 서비스, 통신</b> , 정보시스템, 의료정보학, 정보과학 및 도서관학, 사회과학 역사, 역사학, 경영학, 비즈니스
	C6	사회과학, 사회학, 법률, 윤리	<b>사회학, 심리학, 의학윤리, 법률, 정치학</b> , 윤리학, 행정학, 사회과학, 커뮤니케이션
B	C7	지리학, 환경연구, 생태학	<b>지리학, 환경연구, 경제학</b> , 도시학, 자연지리학, 지구과학, 농업경제학
	C8	환경공학, 토목공학, 해양공학	<b>환경공학, 해양학, 해양공학</b> , 환경과학, 토목공학, 화학공학, 친환경 과학, 수자원, 선박공학
	C9	위생, 약학, 종양학	<b>공중, 환경 및 노동 위생, 약리학 및 약학, 종양학</b> , 독성학

조)들은 C2의 “Neuroscience(신경과학)”과 C8의 “Engineering, Environmental(환경공학)”, 그리고 C9의 “Public, Environmental & Occupational Health(공중, 환경 및 노동 위생)” 등이 있다. 위의 주제분야들은 각 군집 내에서 영향력이 가장 크고, 데이터과학 연구들의 네트워크 전반에 있어서도 그 영향력의 범위가 좁지 않다고 볼 수 있다.

## 5. 결론

본 연구는 데이터과학에 대한 연구들이 수행되고 있는 연구분야를 파악하고, 데이터과학 분야의 지적구조를 파악하기 위하여 Web of Science에 수록된 자료들 중 데이터과학을 주제로 하는 667건의 자료의 주제범주를 나타내는 용어들을 수집하여 기술분석과 동시출현단어 빈도분석을 실시하고, 네트워크 분석을 통해 데이터과학 분야 연구들의 주제범주들을 군집별로 시각화하였다. Web of Science에서 “Data Science”를 주제로 하는 자료들에 할당된 Web of Science 주제범주는 총 159개이고, 동시출현단어 빈도분석에 의해 그 중 연관성이 있다고 나타난 134개의 주제범주에 대해 네트워크 분석을 실시하였다. 동시출현단어 빈도를 정규화한 코사인 계수로 연관성 행렬을 구한 후, 이를 패스파인더 네트워크로 시각화하였고, PNNC 알고리즘을 이용하여 총 2개의 영역과 9개의 군집을 생성하였다. 네트워크분석 과정에서 산출된 중심성 지수를 활용하여 각 군집에서 네트워크 전반에 미치는 영향력이 큰 주제범주들을 선정하였으며, 이를 각 군집의 특성을 가장 잘 설명

하는 대표적 주제로 정의하였다.

수집된 주제범주들에 대한 기술분석을 통해 파악된 내용을 살펴보면, 데이터과학을 주제로 하는 연구들은 지난 10여 년간 지속적으로 증가하는 추세에 있으며, 가장 많은 연구를 수행한 학문분야는 컴퓨터공학 분야인 것으로 나타났다. 구체적으로는 컴퓨터공학 분야의 세부 주제인 다학제적 응용분야, 인공지능분야, 정보시스템분야, 이론과 방법분야 모두 데이터과학 관련 연구들이 많이 이루어지고 있음을 알 수 있었고, 뒤이어 확률과 통계, 정보과학과 도서관학 분야 역시 다수의 연구가 진행되고 있는 분야로 나타났다. 이 외에도 경영학, 수학, 신경과학 등 다양한 주제 분야에서 데이터과학 관련 연구를 진행하고 있고, 문서유형별 구성에서도 학술지 논문뿐만 아니라 논평이나 검토(Review) 등 다양한 유형의 자료에서 데이터과학을 주제로 논의를 하고 있는 것으로 나타났다. 학술지 중에서는 통계학 관련 학술지인 Statistical Analysis and Data Mining에서 가장 많은 데이터과학 관련 연구를 수록하고 있는 것으로 나타났으며, Big Data나 Information Sciences와 같이 특정한 주제를 다루는 학술지들도 있었는데, PLOS One이나 IEEE Access와 같이 다학문적인 성격의 학술지들도 상위권에 위치하고 있었다. 다만 총 667건의 자료들이 430종의 학술지에 수록되어 있고 10건 이상의 관련 논문을 수록한 학술지가 4종밖에 되지 않아, 주제범주별 분포와 마찬가지로 매우 다양한 주제의 학술지에서 데이터과학 관련 연구들을 수록하고 있음을 알 수 있다.

네트워크분석을 통해 생성된 9개 군집을 대표하는 주제들은 다음과 같다. 화학, 생물학, 물리

학으로 이루어진 첫 번째 군집(C1), 신경과학, 생명공학, 의학으로 이루어진 두 번째 군집(C2), 수학, 공학, 경영과학의 세 번째 군집(C3), 컴퓨터공학, 통계 및 확률의 네 번째 군집(C4), 정보과학, 정보시스템, 경영학의 다섯 번째 군집(C5), 사회과학, 사회학, 윤리, 법률 등의 여섯 번째 군집(C6)이 첫 번째 영역에 해당하고, 지리학, 환경연구, 생태학으로 이루어진 일곱 번째 군집(C7)과 환경공학, 토목공학, 해양공학으로 이루어진 여덟 번째 군집(C8), 그리고 위생, 약학, 중앙학으로 구성된 아홉 번째 군집(C9)은 두 번째 영역에 해당한다. 첫 번째 영역 중 네 번째 군집과 다섯 번째 군집은 컴퓨터공학 분야의 세부주제들과 통계 및 확률, 그리고 정보학 주제들이 포함되어 있어 데이터과학과 가장 깊은 관련이 있는 주제들로 이루어져 있다. 그 외에는 수학, 공학 등 자연과학 분야의 다양한 주제들을 다루는 군집들(C1, C2, C3, C7, C8, C9)이 대부분이고 유일하게 여섯 번째 군집(C6)은 사회과학 분야의 여러 세부 주제들로 이루어져 있다. 따라서 데이터과학 분야와 관련된 학문분야는 기존의 데이터과학에 대한 소개에서 다루고 있는 통계학, 컴퓨터공학, 정보학 외에도 화학, 물리학, 생물학, 지리학, 환경학, 공학 등 다양한 자연과학 분야의 연구가 있다는 것을 주목해야 할 것이다. 더불어 커뮤니케이션이나 심리학, 윤리학, 정치학, 행정학 등 사회과학의 많은 세부 주제들 역시 데이터과학과 관련된 연구를 진행하고 있다는 점도 눈여겨볼 필요가 있다.

본 연구는 데이터과학 관련 분야에 대한 자료를 Web of Science 데이터베이스에만 한정하여 수집한 점, 특히 그 중에서도 주제검색을

통해 명백하게 “Data Science”라는 표현이 들어 있는 논문만을 대상으로 한 점과 Web of Science에서 부여한 주제범주들만을 분석대상으로 했다는 점에서 더 많은 자료와 충분한 주제어를 포함하지 못했다는 한계점이 있다. 비록 Web of Science에 수록된 자료들은 질적인 수준이 어느 정도 보장된다는 장점이 있으나, 한 편으로는 매우 엄격한 기준으로 선별된 자료들이기 때문에 비교적 새롭게 등장하거나 연구되는 주제에 대해서는 다양한 자료를 포함하고 있지 않을 수 있다는 단점이 있다. 그리고, “Data Science”라는 표현이 명시되어 있지 않아도 데이터 과학 분야의 논문으로 포함되어야 할 논문들이 다수 있을 수 있다. 예를 들어 EPJ Data Science나 Advances in Data Science and Adaptive Analysis 등과 같이 학술지명 자체에 “Data Science”라는 표현을 포함하고 있는 학술지에 실린 모든 논문들은 데이터 과학 분야 관련 논문으로 보아도 무방할 것이다. 그러나 본 연구에서는 “Data Science”라는 표현을 대상으로 하는 주제검색에서 누락될 경우 분석대상으로 포함되지 않았다. 따라서 본 연구의 분석대상에 포함된 논문들이 데이터과학 분야를 대표하거나 포괄하는 논문의 전체라고 보기에 어려울 수 있다. 또한 Web of Science 주제범주의 경우 데이터베이스에 자료를 색인할 때 시스템이 일방적으로 부여하는 주제어이기 때문에 상당히 객관적일 수는 있으나 간혹 자료의 주제를 충분히 파악하지 못하는 경우도 있을 수 있으며, 덧붙여 각 주제범주에 대한 한글번역이 자연스럽지 못하다는 단점이 있다. 주제범주에 대한 분석에 혼동이 없도록 하기 위해 Web of Science에서 제공하는 한글 번역

을 그대로 가져왔으나, 일부 주제범주는 기계를 통해 자동으로 번역한 것 같은 어색한 표현이거나 더 이상 쓰지 않는 표현으로 번역되어 있다. 예를 들어 “Information Science & Library Science” 범주의 경우 Web of Science에서 제공하는 한글번역은 “정보과학 및 도서관학”이고, “Public, Environmental & Occupational Health”는 “공중, 환경 및 노동 위생”으로 번역되어 있다. 따라서 각 군집을 대표하는 주제어는 Web of Science 주제범주가 아닌 일반적으로 학문분야를 표현하는 용어들을 이용하여 선정하였다. 또한, 본 연구는 데이터과학의 지적구조를 파악하기 위해 네트워크분석 결과만을 이용했다는 점에서 지적구조에 관한 타 연구들에 비해 다양한 측면에서의 시도가 이루어지지 않았다는 한계점이 있다. 지적구조에 관한 다른 연구들을 살펴보면, 네트워크분석 뿐만 아니라 피어슨 상관계수를 이용한 군집분석

을 실시하고, 그 결과를 네트워크분석 결과와 비교하는 등 여러 가지 방법을 활용하여 지적구조를 파악하였다.

이러한 한계점을 보완하기 위해 향후 연구에서는 Web of Science 이외의 데이터베이스까지 포함하여 데이터과학 관련 자료들을 수집하고, 저자 키워드나 초록 등 다양한 방법으로 키워드를 추출하여 분석을 시도할 필요가 있다. 또한 더 많은 자료의 분석을 위해 동시출현단어 빈도분석뿐만 아니라 토픽모델링 등 다양한 분석방법을 활용하면 데이터과학 분야의 지적구조를 좀 더 자세히 살펴볼 수 있을 것이다. 데이터과학 분야에 대한 충분한 분석이 이루어진다면, 데이터과학 연구가 필요한 분야를 파악하는데 도움이 될 수 있고, 일부 대학에서 신설되었거나 연계융합전공으로 제공되는 데이터과학 분야의 교과과정을 개발하는데 있어서도 그 방향성을 제시할 수 있을 것으로 기대된다.

## 참 고 문 헌

- 강범일, 박지홍 (2013). 프로파일링 분석과 동시출현단어 분석을 이용한 한국어교육학의 정체성 분석. 정보관리학회지, 30(4), 195-213. <http://dx.doi.org/10.3743/KOSIM.30.4.195>
- 강범일, 이재운 (2014). 트위터 관련 연구에 대한 계량정보학적 분석. 정보관리학회지, 31(3), 293-311. <http://dx.doi.org/10.3743/KOSIM.31.3.293>
- 강지혜 (2016). 문헌정보학과와 데이터 사이언스 커리큘럼 개발 실태와 방향성 고찰. 한국도서관·정보학회지, 47(3), 343-363. <http://dx.doi.org/10.16981/kliss.47.201609.343>
- 김선희, 윤순진 (2015). 한국 환경사회학의 지적 구조: ECO 논문 제목의 동시출현단어분석을 중심으로. 환경사회학연구 ECO, 19(2), 165-211.
- 김판준 (2015a). 국외 독서 및 독서교육 연구동향 분석. 정보관리학회지, 32(3), 69-97. <http://dx.doi.org/10.3743/KOSIM.2015.32.3.069>

- 김판준 (2015b). 디지털 큐레이션 연구동향 분석과 과제: 문헌정보학 분야를 중심으로. 정보관리학회지, 32(1), 265-295. <http://dx.doi.org/10.3743/KOSIM.2015.32.1.265>
- 김하진, 송민 (2014). 동시출현단어 분석을 통한 국내외 정보학 학회지 연구동향 파악. 정보관리학회지, 31(1), 99-118. <http://dx.doi.org/10.3743/KOSIM.2014.31.1.099>
- 서선경, 정은경 (2013). 동시출현단어 분석 기반 오픈 액세스 분야 지적구조에 관한 연구. 한국비블리아학회지, 24(1), 207-228. <https://doi.org/10.14699/kbiblia.2013.24.1.207>
- 이명호 (2016). 데이터 사이언스 교과과정에 대한 연구. 한국비블리아학회지, 27(1), 263-290. <http://dx.doi.org/10.14699/kbiblia.2016.27.1.263>
- 이수상 (2014). 언어 네트워크 분석 방법을 활용한 학술논문의 내용분석. 정보관리학회지, 31(4), 49-68. <http://dx.doi.org/10.3743/KOSIM.31.4.049>
- 이재운 (2006). 지적 구조 분석을 위한 새로운 클러스터링 기법에 관한 연구. 정보관리학회지, 23(4), 215-231. <https://doi.org/10.3743/kosim.2006.23.4.215>
- 이재운 (2013). tnet과 WNET의 가중 네트워크 중심성 지수 비교 연구. 정보관리학회지, 30(4), 241-264. <http://dx.doi.org/10.3743/KOSIM.2013.30.4.241>
- 이재운 (2015). 데이터 사이언스와 데이터 리터러시. 한국정보관리학회 학술대회 논문집, 11-15.
- 이재운, 김수정 (2016). 국내 재난 관련 연구 동향에 대한 계량정보학적 분석. 정보관리학회지, 33(3), 103-124. <http://dx.doi.org/10.3743/KOSIM.2016.33.4.103>
- 이정규, 이용구 (2013). 동시출현단어 분석을 이용한 도서관경영 분야의 지적구조 분석. 한국정보관리학회 학술대회 논문집, 23-26.
- 장령령, 홍현진 (2014). 학술지 중요도와 키워드 순서를 고려한 단어동시출현분석을 이용한 독서분야의 지적구조 분석. 한국비블리아학회지, 25(1), 295-318. <http://dx.doi.org/10.14699/kbiblia.2014.25.1.295>
- 최상희 (2014). 동시출현단어분석을 이용한 연관영화정보 분석 연구. 정보관리학회지, 31(4), 161-178. <http://dx.doi.org/10.3743/KOSIM.2014.31.4.161>
- 최예진, 정연경 (2016). 동시출현단어 분석에 기반한 메타데이터 분야의 지적구조에 관한 연구. 정보관리학회지, 33(3), 63-83. <http://dx.doi.org/10.3743/KOSIM.2016.33.3.063>
- 최형욱, 정은경 (2017). 사회학 분야의 연구데이터 특성과 지적구조 규명에 관한 연구. 정보관리학회지, 34(3), 109-124. <http://dx.doi.org/10.3743/KOSIM.2017.34.3.109>
- 허고은, 송민 (2013). 저자동시인용 분석과 동시출현단어 분석을 이용한 의료정보학 저널의 지적구조 분석. 정보관리학회지, 30(2), 207-225. <https://doi.org/10.3743/kosim.2013.30.2.207>
- Cervone, H. F. (2015). Informatics and data science: An overview for the information professional. Digital Library Perspectives, 32(1), 7-10. <http://dx.doi.org/10.1108/DLP-10-2015-0022>
- Cho, J. (2014). Intellectual structure of the institutional repository field: A co-word analysis. Journal

- of Information Science, 40(3), 386-397. <http://dx.doi.org/10.1177/0165551514524686>
- Danowski, J. (1993). Network analysis of message content. In W. D. Richards Jr. & G. A. Barnett (Eds.), *Progress in Communication Sciences IV* (pp. 197-221). Norwood, NJ: Ablex.
- Gan, C., & Wang, W. (2015). Research characteristics and status on social media in China: A bibliometric and co-word analysis. *Scientometrics*, 105(2), 1167-1182.  
<https://doi.org/10.1007/s11192-015-1723-2>
- He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends*, 48(1), 133-159.
- Hu, C. P., Hu, J. M., Deng, S. L., & Liu, Y. (2013). A co-word analysis of library and information science in China. *Scientometrics*, 97(2), 369-382.  
<https://doi.org/10.1007/s11192-013-1076-7>
- Kuhn, T. (1996). *The Structure of Scientific Revolutions* (3rd ed.). Chicago: The University of Chicago Press.
- Leydesdorff, L., & Nerghe, A. (2017). Co-word maps and topic modeling: A comparison using small and medium-sized corpora (N<1,000). *Journal of the Association for Information Science and Technology*, 68(4), 1024-1035. <http://dx.doi.org/10.1002/asi.23740>
- Liu, G. Y., Hu, J. M., & Wang, H. L. (2012). A co-word analysis of digital library field in China. *Scientometrics*, 91(1), 203-217. <https://doi.org/10.1007/s11192-011-0586-4>
- Nguyen, L. T. K., & Keip, M. (2018). A data-driven approach to nonlinear elasticity. *Computers and Structures*, 194, 97-115. <http://dx.doi.org/10.1016/j.compstruc.2017.07.031>
- Ravikumar, S., Agrahari, A., & Singh, S. N. (2015). Mapping the intellectual structure of scientometrics: A co-word analysis of the journal *Scientometrics* (2005-2010). *Scientometrics*, 102, 929-955. <http://dx.doi.org/10.1007/s11192-014-1402-8>
- Tang, R., & Sae-Lim, W. (2016). Data Science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus. *Education for Information*, 32, 269-290. <http://dx.doi.org/10.3233/EFI-160977>
- Wang, Z. Y., Li, G., Li, C. Y., & Li, A. (2012). Research on the semantic-based co-word analysis. *Scientometrics*, 90(3), 855-875. <https://doi.org/10.1007/s11192-011-0563-y>
- Wang, Z., Zhao, H., & Wang, Y. (2015). Social networks in marketing research 2001-2014: A co-word analysis. *Scientometrics*, 105(1), 65-82.  
<https://doi.org/10.1007/s11192-015-1672-9>
- Web of Science (2017, November 24). Web of Science Core Collection Help. Retrieved from <http://images.webofknowledge.com/WOKRS526R11/help/WOS/contents.html>
- White, H. D. (2003). Pathfinder networks and author cocitation analysis: A remapping of

paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, 54(5), 423-434. <https://doi.org/10.1002/asi.10228>

Yang, Y., Wu, M., & Cui, L. (2012). Integration of three visualization methods based on co-word analysis. *Scientometrics*, 90(2), 659-673. <https://doi.org/10.1007/s11192-011-0541-4>

Zong, Q. J., Shen, H. Z., Yuan, Q. J., Hu, X. W., Hou, Z. P., & Deng, S. G. (2013). Doctoral dissertations of library and information science in China: A co-word analysis. *Scientometrics*, 94(2), 781-799. <https://doi.org/10.1007/s11192-012-0799-1>

• 국문 참고문헌에 대한 영문 표기  
(English translation of references written in Korean)

Choi, Hyung Wook, & Chung, EunKyung (2017). An investigation on characteristics and intellectual structure of sociology by analyzing cited data. *Journal of the Korean Society for Information Management*, 34(3), 109-124. <http://dx.doi.org/10.3743/KOSIM.2017.34.3.109>

Choi, Sanghee (2014). An analysis of related movie information using the co-word method. *Journal of the Korean Society for Information Management*, 31(4), 161-178. <http://dx.doi.org/10.3743/KOSIM.2014.31.4.161>

Choi, Ye-Jin, & Chung, Yeon-Kyoung (2016). A study on the intellectual structure of metadata research by using co-word analysis. *Journal of the Korean Society for Information Management*, 33(3), 63-83. <http://dx.doi.org/10.3743/KOSIM.2016.33.3.063>

Heo, Go Eun, & Song, Min (2013). Examining the intellectual structure of a medical informatics journal with author co-citation analysis and co-word analysis. *Journal of the Korean Society for Information Management*, 30(2), 207-225. <https://doi.org/10.3743/kosim.2013.30.2.207>

Kang, Beomil, & Lee, Jae Yun (2014). A bibliometric analysis on twitter research. *Journal of the Korean Society for Information Management*, 31(3), 293-311. <http://dx.doi.org/10.3743/KOSIM.31.3.293>

Kang, Beomil, & Park, Ji-Hong (2013). Profiling and co-word analysis of teaching Korean as a foreign language domain. *Journal of the Korean Society for Information Management*, 30(4), 195-213. <http://dx.doi.org/10.3743/KOSIM.30.4.195>

Kang, Ji Hei (2016). Study on the current status of Data Science curriculum in Library and Information Science and its direction. *Journal of Korean Library and Information Science Society*, 47(3), 343-363. <http://dx.doi.org/10.16981/kliss.47.201609.343>

- Kim, Ha Jin, & Song, Min (2014). A study on the research trends in domestic/international information science articles by co-word analysis. *Journal of the Korean Society for Information Management*, 31(1), 99-118. <http://dx.doi.org/10.3743/KOSIM.2014.31.1.099>
- Kim, Pan Jun (2015a). An analytical study on research trends of reading and reading instruction in overseas: Focused on Library and Information Science. *Journal of the Korean Society for Information Management*, 32(3), 69-97. <http://dx.doi.org/10.3743/KOSIM.2015.32.3.069>
- Kim, Pan Jun (2015b). An analytical study on research trends of digital curation: Focused on Library and Information Science. *Journal of the Korean Society for Information Management*, 32(1), 265-295. <http://dx.doi.org/10.3743/KOSIM.2015.32.1.265>
- Kim, Sun-Hoi, & Yun, Sun-Jin (2015). The knowledge structure of environmental sociology in Korea: Based on the co-word analysis of article titles in ECO. *ECO*, 19(2), 165-211.
- Lee, Jae Yun (2006). A novel clustering method for examining and analyzing the intellectual structure of a scholarly field. *Journal of the Korean Society for Information Management*, 23(4), 215-231. <https://doi.org/10.3743/kosim.2006.23.4.215>
- Lee, Jae Yun (2013). A comparison study on the weighted network centrality measures of tnet and WNET. *Journal of the Korean Society for Information Management*, 30(4), 241-264. <http://dx.doi.org/10.3743/KOSIM.2013.30.4.241>
- Lee, Jae Yun (2015). Data Science and data literacy. *Proceedings of the 22nd Conference of Korean Society for Information Management*, 11-15.
- Lee, Jae Yun, & Kim, Soojung (2016). A bibliometric analysis of research trends on disaster in Korea. *Journal of the Korean Society for Information Management*, 33(3), 103-124. <http://dx.doi.org/10.3743/KOSIM.2016.33.4.103>
- Lee, Jung-Gyu, & Lee, Yong-Gu (2013). A study on the analysis of intellectual structure of library management studies using co-word analysis. *Proceedings of the 20th Conference of Korean Society for Information Management*, 23-26.
- Lee, Soo-Sang (2014). A content analysis of journal articles using the language network analysis methods. *Journal of the Korean Society for Information Management*, 31(4), 49-68. <http://dx.doi.org/10.3743/KOSIM.31.4.049>
- Seo, SunKyung, & Chung, EunKyung (2013). Domain analysis on the field of open access by co-word analysis. *Journal of the Korean Biblia Society for Library and Information Science*, 24(1), 207-228. <https://doi.org/10.14699/kbiblia.2013.24.1.207>
- Yi, Myongho (2016). A study on the curriculums of Data Science. *Journal of the Korean Biblia*

Society for Library and Information Science, 27(1), 263-290.

<http://dx.doi.org/10.14699/kbiblia.2016.27.1.263>

Zhang, Ling Ling, & Hong, Hyun Jin (2014). Examining the intellectual structure of reading studies with co-word analysis based on the importance of journals and sequence of keywords.

Journal of the Korean Biblia Society for Library and Information Science, 25(1), 295-318.

<http://dx.doi.org/10.14699/kbiblia.2014.25.1.295>

[부록 1] 데이터과학 관련 Web of Science 주제범주에 대한 영역별/군집별 목록

영역	군집	Web of Science 주제범주	국문명칭
A	C1	<b>Materials Science, Multidisciplinary</b>	재료과학, 다학문분야
		Biochemistry & Molecular Biology	생화학 및 분자 생물학
		Physics, Applied	물리학, 응용
		Chemistry, Multidisciplinary	화학, 다학문분야
		Cell Biology	세포 생물학
		Energy & Fuels	에너지 및 연료
		Biophysics	분자 생물 물리학
		Immunology	면역학
		Metallurgy & Metallurgical Engineering	금속 공학
		Microbiology	미생물학
		Physics, Condensed Matter	물리학, 고체
		Crystallography	결정학
		Infectious Diseases	전염병
		Nanoscience & Nanotechnology	나노 과학 & 나노 기술
		Allergy	알레르기
		Biology	생물학
		Chemistry, Physical	화학, 물리
		Materials Science, Coatings & Films	재료과학, 코팅 및 필름
		Mineralogy	광물학
	Mining & Mineral Processing	채광 및 선광	
	C2	<b>Neurosciences</b>	신경 과학
		Biochemical Research Methods	생화학 연구방법
		Mathematical & Computational Biology	수학 및 컴퓨터 생명공학
		Biotechnology & Applied Microbiology	생명공학 및 응용 미생물학
		Medicine, Research & Experimental	의학, 연구 및 실험
		Anesthesiology	마취학
		Chemistry, Analytical	화학, 분석
		Clinical Neurology	임상 신경학
		Physiology	생리학
		Behavioral Sciences	행동과학
		Food Science & Technology	식품 과학 및 기술
		Cardiac & Cardiovascular Systems	심장 및 심혈관계
		Peripheral Vascular Disease	말초 관다발병
Respiratory System		호흡기계	
Critical Care Medicine	중환자 의학		
Electrochemistry	전기화학		
Genetics & Heredity	유전학 및 유전		
Instruments & Instrumentation	계기 및 계측		
Psychiatry	정신의학		
Psychology	심리학		
Spectroscopy	분광학		
Urology & Nephrology	비뇨기과학 및 신장학		

영역	군집	Web of Science 주제범주	국문명칭
A	C3	<b>Operations Research &amp; Management Science</b>	오퍼레이션 리서치 및 경영과학
		Mathematics, Interdisciplinary Applications	수학, 학제간 응용
		Engineering, Industrial	공학, 산업
		Physics, Mathematical	물리학, 수학
		Engineering, Multidisciplinary	공학, 다학문분야
		Psychology, Experimental	심리학, 실험
		Social Sciences, Mathematical Methods	사회과학, 수학방법
		Engineering, Manufacturing	공학, 제조
		Education, Scientific Disciplines	교육, 과학 분야
		Mathematics, Applied	수학, 응용
		Mechanics	역학
		Anthropology	인류학
		Archaeology	고고학
		Automation & Control Systems	자동화 및 제어시스템
		Engineering, Mechanical	공학, 기계
		Family Studies	가족연구
		Fisheries	어업
		Geosciences, Multidisciplinary	지구과학, 다학문분야
		Language & Linguistics	언어 및 언어학
		Linguistics	언어학
	Mathematics	수학	
	Multidisciplinary Sciences	다학문분야 과학	
	Obstetrics & Gynecology	산과 및 부인과 의학	
	Parasitology	기생충학	
	Physics, Fluids & Plasmas	물리학, 액체 및 플라즈마	
	Physics, Multidisciplinary	물리학, 다학문분야	
	Psychology, Developmental	심리학, 발달	
	Psychology, Mathematical	심리학, 수리	
	Reproductive Biology	생식 생물학	
	Social Issues	사회적 문제	
	Social Work	사회 사업	
	Tropical Medicine	열대 의학	
	Veterinary Sciences	수의학	
	C4	<b>Computer Science, Interdisciplinary Applications</b>	컴퓨터 공학, 학제간 응용
		Statistics & Probability	통계 및 확률
Computer Science, Artificial Intelligence		컴퓨터 공학, 인공지능	
Computer Science, Theory & Methods		컴퓨터 공학, 이론 및 방법	
Engineering, Biomedical		공학, 생물의학	
Computer Science, Hardware & Architecture		컴퓨터 공학, 하드웨어 및 아키텍처	
Computer Science, Software Engineering		컴퓨터 공학, 소프트웨어 공학	
Radiology, Nuclear Medicine & Medical Imaging		영상 의학, 핵의학 및 의료영상	
Astronomy & Astrophysics		천문학 및 천체물리학	
Ergonomics		인체 공학	

영역	군집	Web of Science 주제범주	국문명칭
A	C5	<b>Computer Science, Information Systems</b>	컴퓨터 공학, 정보시스템
		Medical Informatics	의료 정보학
		Information Science & Library Science	정보과학 및 도서관학
		Health Care Sciences & Services	보건 과학 및 서비스
		Engineering, Electrical & Electronic	공학, 전기 전자
		Management	경영학
		Business	비즈니스
		History Of Social Sciences	사회과학 역사
		Telecommunications	통신
		Education & Educational Research	교육 및 교육 연구
	History	역사학	
	Medicine, General & Internal	의학, 일반 내과	
	C6	<b>Social Sciences, Biomedical</b>	사회과학, 생물의학
		Social Sciences, Interdisciplinary	사회과학, 학제간
		Health Policy & Services	보건정책 및 서비스
		Ethics	윤리학
		Sociology	사회학
		Medical Ethics	의학 윤리
		Psychology, Multidisciplinary	심리학, 다학문분야
		Communication	커뮤니케이션
Law		법률	
Political Science		정치학	
Public Administration	행정학		
B	C7	<b>Geography</b>	지리학
		Environmental Studies	환경 연구
		Urban Studies	도시학
		Geography, Physical	자연 지리학
		Economics	경제학
		Remote Sensing	원격 탐사
		Transportation	교통
		Agricultural Economics & Policy	농업 경제학 및 정책
		Ecology	생태학
		Geochemistry & Geophysics	지구과학 및 지구물리학
	Imaging Science & Photographic Technology	영상 과학 및 사진 기술	
	Planning & Development	기획 및 개발	
	C8	<b>Engineering, Environmental</b>	공학, 환경
		Environmental Sciences	환경과학
		Engineering, Civil	공학, 토목
		Engineering, Chemical	공학, 화학
		Green & Sustainable Science & Technology	지속가능한 친환경 과학 및 기술
		Water Resources	수자원
		Oceanography	해양학
		Engineering, Marine	공학, 선박
Engineering, Ocean		공학, 해양	
Polymer Science		고분자 과학	
C9	<b>Public, Environmental &amp; Occupational Health</b>	공중, 환경 및 노동 위생	
	Pharmacology & Pharmacy	약리학 및 약학	
	Oncology	종양학	
		Toxicology	독성학