동시출현단어 분석 기반 오픈 액세스 분야 지적구조에 관한 연구

Domain Analysis on the Field of Open Access by Co-Word Analysis

서 선 경 (SunKyung Seo)*
정 은 경 (EunKyung Chung)**

초 록

학술 커뮤니케이션의 변화로 인해 오픈 액세스 분야는 상대적으로 최근에 많은 연구가 이루어지고 있다. 본 연구는 동시출현단어 분석을 사용하여 오픈 액세스 분야의 지적구조를 규명하여 연구동향을 제시하고자 하였다. 이를 위해서 데이터 수집은 Web of Science 기반으로 수행하였다. 검색 대상 기간은 1998년 1월 1일부터 2012년 7월 31일까지이며, Topic검색을 통하여 총 479건의 저널 논문을 수집하였다. 총 479건의 저널 논문 제목과 초록에서 명사구 형태의 키워드는 총 8,643개(문헌 당 18.04개)를 추출하였다. 오픈 액세스 분야의 지적구조 규명을 위해 첫째, 네트워크 분석을 통하여 18개의 세부 주제 영역을 밝혔으며, 오픈 액세스 분야 키워드들의 지적 관계를 시각화하여, 키워드 관계, 중심성 분석을 통한 전역 중심 키워드와 지역 중심이 높은 키워드를 제시하였다. 둘째, 군집분석을 실시하여 형성된 4개의 군집을 MDS지도에 표시하였으며, 각 키워드들 간의 상관관계에 따른 지적구조를 제시하였다. 이러한 연구의 결과는 오픈 액세스 분야의 지적구조를 밝히며, 향후 연구 방향성 모색에 유용하게 사용될 수 있을 것으로 기대한다.

ABSTRACT

Due to the advance of scholarly communication, the field of open access has been studied over the last decade. The purpose of this study is to analyze and demonstrate the field of open access via co-word analysis. The data set was collected from Web of Science citation database during the period from January 1998 to July 2012 using the Topic category. A total of 479 journal articles were retrieved and 8,643 noun keywords were extracted from the titles and abstracts. In order to achieve the purpose of this study, network analysis, clustering analysis and multidimensional scaling mapping were used to examine the domain and the sub-domains of open access field. 18 clusters in the network analysis are recognized and 4 clusters are shown in the map of multidimensional scaling. In addition, the centrality analysis in the weighted networks was used to explore the significant keywords in this field. The results of this study are expected to demonstrate and guide the intellectual structure and new approaches of open access field.

키워드: 동시출현단어분석, 지적구조, 오픈액세스
Co-Word Analysis, Intellectual Structure, Open Access

---

* 이화여자대학교 일반대학원 문헌정보학과(welloffssk@ewhain.net)
** 이화여자대학교 사회과학대학 문헌정보학전공 조교수(echung@ewha.ac.kr) (교신저자)
논문접수일자 : 2013년 2월 20일 논문심사일자 : 2013년 2월 26일 게재확정일자 : 2013년 3월 17일

1. 서 론

최근 정보기술의 발전과 학술자료 구입・구독의 문제점, 공공기금에 의한 연구 성과물에 대한 비상업적인 무료 공유 등에서 비롯된 학술 커뮤니케이션의 위기를 해소할 수 있는 대안으로 오픈 액세스에 대한 관심이 고조되고 있다. 이와 같은 학술 정보 유통의 새로운 패러다임인 오픈 액세스의 움직임은 연구자뿐만 아니라 학회, 도서관, 출판사로 확산되고 있으며, 이에 따른 발전과 그 변화의 폭에 주목할 필요가 있다. 이를 위해서 본 연구는 오픈 액세스 분야의 지적구조 분석을 수행하고자 한다. 일반적으로 지적구조 분석은 문헌이나 저자의 동시인용 분석, 서지결합법, 동시출현단어 분석 등의 기법이 활용된다. 오픈 액세스 분야는 비교적 최근에 활발하게 연구가 이루어졌기 때문에 인용 중심의 분석보다는 동시출현단어 분석이 보다 적절하다. 동시출현단어 분석은 텍스트전문이나 제목, 초록, 키워드를 활용하여 분석함으로써 주제 영역이나 하위 분야 또는 분야 속의 패턴 등을 규명할 수 있다.

따라서 본 연구는 동시출현단어 분석을 사용하여 오픈 액세스 분야의 연구 경향을 반영하는 지적구조를 제시하고 하위 주제 영역의 구성을 규명하는데 목적이 있다. 이를 위해서 1998년부터 2012년 사이에 발간된 오픈 엑세스 주제의 저널 논문을 Web of Science 데이터베이스에서 수집하였다. 총 479건의 논문을 수집하였으며, 제목과 초록에서 추출된 키워드는 총 8,643개 이다. 추출된 키워드를 기반으로 오픈 액세스 영역의 지적구조를 다각적으로 분석하기 위해 네트워크 분석을 실시하여, 키워드관계 네트워크의 시각화 통해 중심 주제와 세부 주제 영역을 파악하고, 중심성 분석으로 전역 중심 키워드와 지역 중심 키워드를 확인하고자 한다. 그리고 네트워크 분석을 보완하기 위하여 군집분석을 수행하고, 이 결과를 다차원축척지도로 나타내어 오픈 액세스 영역의 전체적인 주제 영역의 흐름 및 구성을 제시하였다. 본 연구의 분석결과는 오픈 액세스 분야의 두 명의 연구자와의 면담을 통해 결과해석에 대한 전문성을 높였다.

이러한 연구결과는 학술 커뮤니케이션의 패러다임 변화에 있어서 중요한 축인 오픈 액세스 분야에 대한 학문적 구조와 하위 분야에 대한 정보를 제공할 수 있으며, 이를 바탕으로 학문의 발전방향 제시 등에 유용하게 사용될 수 있을 것으로 기대한다.

2. 관련 연구

오픈 액세스 출판물은 일반적으로 온라인에서 누구나, 어디에서나 무료로 이용 가능하도록 만들어진 학술 출판물의 배포 유형이라 정의할 수 있다. 2002년 부다페스트 선언에서는 전통적인 학술 커뮤니케이션의 대안적 전략으로 오픈 액세스 저널과 셀프아카이빙을 제시하였다(Koehler 2006). 본 연구의 분석 대상이 되는 오픈 액세스는 출현배경을 시작으로 개념을 둘러싼 이슈들이 끊임없이 논의되며 발전하고 있다. 이러한 분석 영역의 지적구조를 기술하기 위한 방법론 중 하나인 동시출현단어 분석을 이용하여 주제 영역의 경향과 시기적 변화 등을 파악한 선행연구들을 살펴보면 다음과 같다.

관련된 선행연구는 크게 연구기법에 따라서

두 가지로 구분할 수 있다. 첫째는 군집분석 중심의 지적구조 분석에 관한 연구이다. Milojević 등(2011)은 문헌정보학 분야의 지적구조를 규명하기 위하여, 논문제목 단어 기반의 동시출현단어 분석법을 적용하였다. 이들의 연구에서는 16개의 문헌정보학 학술지에서 1998-2007년 기간 동안의 연구와 리뷰 논문 10,344건을 대상으로 하였다. 이 키워드들은 덴드로그램을 이용하여 계층적인 클러스터링을 통해 3개의 주요 하위분야로써 도서관학, 정보학, 계량과학/계량정보학으로 제시하였다. 추가적으로 독립된 2개의 하위영역인 정보추구행위와 도서관 서지교육을 확인할 수 있었다고 보고하였다. 또한 다차원축적지도를 사용하여 2차원 상에서 시기에 따른 문헌정보학 분야의 지적구조 변화를 나타내었다. 박재신과 정영미(2010)는 다차원척도법과 네트워크 분석을 통해 환경 관련 분야를 학술적 영역과 실천적 영역으로 구분하여 인용분석과 웹링크 분석을 통해 지적구조를 제시하였다.

두 번째는 군집분석, 다차원척도법 이외에 네트워크 분석을 적용하여 동시출현단어 분석을 한 연구들을 찾아볼 수 있다. Liu 등(2012)은 2002-2011년 기간을 범위로 중국의 디지털 도서관분야의 지적구조를 분석하였다. 총 2,647개의 관련 문헌들 중에 9,538개의 키워드들(문헌 당 3.6개)이 수집되었으며, 이 키워드들의 동시출현행렬에 클러스터링과 다차원축적지도 그리고 네트워크 분석을 적용하여 매핑결과를 제시하였다. 분석의 결과는 디지털도서관 분야의 7개의 클러스터로 제시하였다. 각 클러스터들은 디지털 도서관 분야의 연구 방향을 나타내며, 연구 주제 간의 상관관계가 대체로 낮은 점은 다른 나라의 연구들과 비교하여 볼 때, 중국의 디지털 도서관이 상대적으로 분권화 되어져 있음을 보여준다고 하였다. 또한 이 연구가 진행되는 2011년의 중국의 디지털도서관 분야에서 연결중심성이 높은 연구 주제들과 연구 주제 간에 이어주는 역할을 하는 매개중심성이 높은 키워드들을 파악하여 제시하였다. 김희정(2011)은 동시출현단어 분석을 통하여 네트워크 분석만을 적용하여 웹 아카이빙 영역에서 다양한 연구 주제 간의 관련성과 세부 주제 영역을 확인하였다. 이 연구는 검색된 288건의 논문들을 계량분석 소프트웨어인 Network Workbench를 활용하여 최종적 분석 대상으로 93개의 핵심 용어 군을 선정하였다. 93개의 핵심 용어 군을 대상으로 동시출현단어 네트워크를 나타내기 위해 행렬 데이터를 재산출 한 후, 패스파인더 네트워크 방식을 선택하여 NodeXL을 이용하여 네트워크 지도를 작성하였다. 분석 결과 웹 아카이빙 주제 영역의 논문은 1995년도부터 출현하기 시작하였고, 2003년부터 급속히 증가해 왔으며, 의학영역 정보기술 및 시스템과 관련된 이미지 아카이빙 관련 연구들이 가장 중점적으로 수행된 것을 확인할 수 있었다고 하였다. 문헌정보학 및 기록 관리학 영역에서의 웹 아카이빙 연구는 2004년부터 출현하고 있으며, 2009년에 가장 활발하게 이루어졌고, 주제 범주를 크게 웹 아카이빙 및 디지털 보존 프로젝트 영역과 웹 아카이빙 툴과 방법론 영역으로 구분할 수 있다고 하였다. 향후에는 웹 아카이빙 소프트웨어 및 툴 관련 연구가 활성화 될 것으로 예측한다고 제시하였다. 장임숙, 장덕현, 이수상(2011)은 2005년부터 2010년 사이에 발행된 다문화 분야의 논문을 대상으로 동시출현단어 네트워크와 k-core를 제시하였다. 분석된 결과를

통해서 다문화 관련 연구 분야의 주요 핵심 주제와 학제성의 정도, 하위 주제 분야의 응집력 등을 제시하였다.

이상의 선행연구들에서 살펴본 것과 같이 동시출현단어 분석법을 사용하여 학문의 연구 분야에 대한 지적구조를 분석하는 연구들이 국내·외에서 다양하게 진행되고 있으며, 동시출현단어 분석을 통해 해당 분야의 개관과 영역들에 관한 정보를 제공하고 있다.

3. 오픈 액세스 분야의 지적구조 분석과정

3.1 자료 수집과 키워드 선정

본 연구의 주제 영역인 오픈 액세스는 부다페스트 오픈 액세스 회의, IFLA 선언, 베를린 선언, 베데스타 선언 등의 국제적인 관심과 지지와 함께 국외에서 이와 관련한 연구가 활발하게 이루어지고 있다. 현재 국외의 오픈 액세스 연구의 결과가 분석할 만큼 축적이 되어있기 때문에 본 연구에서는 Web of Science에 등재된 저널에 게재된 논문을 수집하였다.

WoS 데이터베이스에서 제공하는 저널 범주를 문헌정보학범주로 제한하여, 기간은 1998년 1월 1일부터 본 연구의 수행 시점인 2012년 7월 31일까지로 설정하였다. 분석 대상이 되는 오픈 액세스 관련 문헌들을 추출하기 위해 사용된 질의키워드 "open access", "open access journal", "institutional repositor*"를 이용하여 주제(Topic) 검색을 실시하였다. 검색 결과, 이 기간 동안 발표된 오픈 액세스에 관한 순수 연구 논문 464건과 리뷰 15건, 총 479건이 수집되었다. 단어 추출을 위해 본 연구에서는 수집된 479건의 문헌들을 계량분석을 위한 공개 소프트웨어인 CiteSpace[1]를 활용하여 제목과 초록에서 키워드와 키워드의 빈도수를 추출하였다. 오픈 액세스와 관련한 선행연구들을 검토한 결과, open access라는 단어가 명사구이고 초록에서 키워드를 추출한다는 점을 고려하여 명사와 명사구를 모두 사용하는 것이 적절하며, 명사구의 구성 명사의 숫자는 보다 상세한 주제 표현을 위해서 2개에서 4개 사이로 지정하였다. 초록이 없는 문헌은 제목에서만 명사구를 추출하였다. 479건의 문헌들에서 추출된 명사구는 총 8,643개(문헌 당 18.04)이다. 이 중 CiteSpace에서 자동으로 산출해 주는 가장 자주 출현한 명사구 6,452개가 추출된 리스트를 이용하여, 분석과 해석의 용이성을 위해 불용어들을 제거하고 방대한 규모를 축소하였다.

전처리 과정은 색인자 효과를 최대한 배제하기 위하여, 동시출현단어 분석에서 보편적으로 적용되는 정규화 작업인 단·복수 교정을 중심으로 하였으며, 오픈 액세스 영역에 관한 용어의 표준적인 기준이 없기 때문에 동의어, 약어들은 상위 빈도수의 용어를 기준으로 정리하였다. 따라서 단·복수의 단어표현이나 이형들은 자주 출현한 명사구 리스트의 상위 빈도수의 형태를 그대로 사용하였다. 예를 들면 institutional repositories는 117번, institutional repository는 66번으로 집계되었으므로 institutional repositories

1) Visualizing Patterns and Trends in Scientific Literature. 〈http://cluster.cis.drexel.edu/~cchen/citespace/〉.

의 용어를 최종 키워드 리스트에서 사용하였으며, 두 빈도수를 더한 183번과 다른 모든 명사구에서 institutional repositories와 institutional repository가 모두 포함된 빈도수를 합산하여 집계하였다.

이와 같은 키워드 정규화 과정 후에 빈도수 7회 이상의 키워드 90개의 리스트를 완성하였다. 이 리스트는 가장 자주 출현한 명사구 리스트를 기준으로 만들어졌기 때문에 하나의 논문에서 중복 출현된 단어의 빈도수가 모두 포함하여 합산되어 있다. 그러나 본 연구의 분석에서 필요한 키워드는 각 논문에서 한번 씩 출현한 횟수이다. 그러므로 90개의 리스트를 대상으로 키워드들의 문헌 빈도를 조사하였다. 6개 이상의 논문에서 출현한 키워드(문헌 빈도 6회 이상의 키워드)를 선정하여, 최종 키워드 리스트 총 84개를 재선정하였다. 문헌 빈도수 7회 이상의 키워드는 73개였고, 문헌 빈도수 5회 이상의 키워드는 87개였다. 문헌 빈도수 6회를 기준으로 7회와 차이가 나는 11개의 키워드들은 본 연구의 결과 도출에 필요하다고 판단하였다. 이렇게 완성된 분석 대상이 되는 최종 키워드 리스트는 다음 〈표 1〉과 같다.

3.2 동시출현단어 행렬 작성

동시출현빈도 행렬의 수치를 가공하는 방법에 따라서 생성되는 네트워크의 형태가 달라진다. 빈도 값을 그대로 이용한 경우에는 네트워크 분석에서 핵심노드 위주의 분석이 가능하지만 비 핵심 노드 간의 관계는 드러나지 않는 단점이 있다(이재윤 2006a). 본 연구에서는 분석 대상 키워드 간 연관도를 산출하는 과정에서 벡터 유사도 공식인 코사인 계수와 피어슨 상관계수를 적용하였다.

분석 대상 키워드 84개가 선정된 이후에 각 문헌 479건에 키워드의 출현정보를 액셀에 입력하여 이재윤이 개발한 COOC ver 0.3.1 프로그램을 이용하여 동시출현단어 행렬을 작성하였다. 먼저 액셀에 첫 번째 열에 각 문헌 번호를 입력하고, 두 번째 열에 각 문헌 번호에 해당하는 CiteSpace를 통해 제목과 초록에서 추출된 키워드를 입력하여 (문헌번호, 키워드) 쌍을 만든다. 이렇게 만들어진 리스트는 총 7,879행이었으며, 이 리스트에서 〈표 1〉을 기준으로 키워드 전처리 작업을 끝낸 후, 필요한 최종 키워드를 포함한 행만을 남겨서 총 1,589행의 최종 분석 대상 출현단어 정보 리스트를 〈그림 1〉과 같이 완성하였다.

프로그램 COOC ver 0.3.1을 실행하면 〈그림 1〉의 각 문헌에 포함된 최종 키워드 출현정보 리스트인 b와 분석 대상이 되는 최종 키워드 리스트 〈표 1〉을 가지고 입력된 행렬 데이터를 분석해서 3가지 결과 행렬을 파일로 생성시켜준다. 먼저 정방대칭행렬인 각 키워드끼리의 동시출현빈도 행렬(84×84)이 출력되며, 이 행렬의 대각선 칸에는 해당 키워드의 출현빈도가 표시된다.

다음은 키워드 간의 동시출현빈도를 코사인 유사도계수로 정규화한 행렬이 출력된다. 코사인 유사도 행렬에서 유사도가 1에 가까울수록 두 단어는 유사도가 높고, 0에 가까울수록 두 단어의 유사도는 낮음을 알 수 있다. 유사도가 높은 키워드 쌍은 서로 유사한 주제 분야에서 다루어지고 있으며, 유사도가 낮은 키워드 쌍은 주제적으로 연관성이 적다는 것을 알 수 있다.

〈표 1〉 문헌 빈도 6회 이상의 최종 84개 키워드 리스트

번호	키워드	빈도수	번호	키워드	빈도수
1	open access	288	43	green open access	12
2	institutional repositories	142	44	golden road	12
3	open access journals	75	45	free access	11
4	scholarly communication	51	46	citation impact	11
5	open access publishing	36	47	citation analysis	11
6	open access movement	30	48	subject repositories	10
7	social sciences	29	49	open source software	10
8	scholarly publishing	25	50	open archives	10
9	metadata	22	51	biomedical	10
10	medical	22	52	document supply	9
11	self-archiving	21	53	web citations	9
12	journal article	21	54	scientific community	9
13	search engine	20	55	research data	9
14	open access repositories	20	56	information resources	9
15	open access model	20	57	citation counts	9
16	information science	20	58	citation advantage	9
17	digital library	20	59	research fund	9
18	developing countries	20	60	scientific literature	8
19	google scholar	19	61	research papers	8
20	oa article	18	62	pubmed central	8
21	long-term	18	63	information professionals	8
22	digital repositories	18	64	digital resources	8
23	business model	18	65	digital preservation	8
24	copyright	17	66	bibliometric	8
25	scientific information	17	67	author-pays model	7
26	research output	16	68	scientific production	7
27	internet	16	69	peer-reviewed journals	7
28	electronic journals	16	70	electronic theses	7
29	university library	15	71	dublin core	7
30	scientific publications	15	72	current trends	7
31	research institutions	15	73	current status	7
32	research article	15	74	usage statistics	6
33	electronic publishing	15	75	university presses	6
34	scientific journals	14	76	scholarly literature	6
35	journal publishing	14	77	quality control	6
36	web page	13	78	protocol	6
37	scientific research	13	79	journal impact factor	6
38	scientific communication	13	80	digital archive	6
39	scholarly journals	12	81	conference proceedings	6
40	open source	12	82	computer science	6
41	open access publication	12	83	commercial publishers	6
42	impact factor	12	84	bibliographic database	6

문헌번호	제목과 초록에서 추출된 키워드
1	stem field
1	scholarly communication
1	producing books
1	open-access scholarly publishing
1	insurmountable obstacle
1	focus lead
1	financial realities
2	status update
2	scientific literature
2	scientific article
2	retraction statement
2	pubmed central
2	personal mendeley library
…	
479	touch screen computer
479	tabloid readers
479	suitable locations
479	psychological state
479	predictor variable
479	outcome variable
479	open access
479	literate group
479	information preference
479	different needs
479	broadsheet readers
479	beats on oncology centre
479	affordable information

▷

문헌번호	최종 키워드
1	open access publishing
1	scholarly communication
2	institutional repositories
2	internet
2	pubmed central
2	scientific literature
3	author-pays model
3	open access
3	research fund
3	scholarly publishing
4	digital repositories
4	institutional repositories
4	pubmed central
…	
471	copyright
471	open access
471	research papers
472	digital archive
472	open access
472	university library
473	open access
474	open access
475	internet
476	open access
477	scholarly communication
478	open access
479	open access

(a) (b)

a : 각 문헌에 포함된 제목과 초록에서 추출된 키워드 리스트
b : 각 문헌에 포함된 최종 키워드 출현정보 리스트

〈그림 1〉 키워드 출현 정보 데이터

마지막으로 산출되는 행렬은 피어슨 상관계수에 의한 2차 연관성 행렬이다. 2차 연관성 행렬은 1차 연관성 행렬을 다시 피어슨 상관계수 등으로 한 차례 더 가공하는 방식으로 산출한다. 이 방법은 White와 Griffith(1981)에 의해 제안되었으며, 이와 같은 2차 연관성 행렬을 이용하여 두 키워드와 제 3의 키워드 간의 동시출현 패턴의 유사함을 측정할 수 있다. 피어슨 상관계수에 의해 산출된 값의 범위는 -1에서 1사이를 가지며, 관계의 크기와 방향을 동시에 나타낸다. 상관계수의 절대치는 관계의 크기를 나타내며, 절대 값이 크면 두 키워드 사이가 밀접하게 관련되어 있음을, 절대 값이 작으면 두 키워드 간의 관련성이 낮음을 의미한다. 그리고 1에

가까울수록 강한 긍정적 관계를 -1에 가까울수록 강한 부정적 관계를 뜻하며, 0은 두 키워드 간에 선형적인 관련성이 없음을 나타낸다.

분석결과, 1차 연관성 행렬인 코사인 유사도 행렬에서 유사도가 가장 높은 키워드 쌍은 golden road-green open access(0.5)로 나타났으며, 최댓값을 제외하고 유사도 값이 0.4 이상인 키워드 쌍은 총 3쌍으로 metadata-dublincore (0.48349), oa article-citation counts(0.4714), research article-webcitations(0.43033) 순으로 나타났다. 유사도가 0인 키워드 쌍 2,268개를 제외하고, 유사도가 가장 낮은 것으로 나타난 키워드 쌍은 information science-institutional repositories(0.01876)으로 나타났으며, 유사도 값이 0.019 이하인 키워드 쌍은 institutional repositories-business model(0.01978), institutional repositories-oa article(0.01978)이었다.

산출된 2차 연관성 행렬의 피어슨 상관관계의 정도에 대하여 Guilford(1950)는 상관계수의 절댓값이 0.9~1.0은 아주 높은 상관관계, 0.7~0.9는 높은 상관관계, 0.4~0.7은 비교적 높은 상관관계, 0.2~0.4는 낮은 상관관계, 0.2 이하는 거의 무시할 정도의 경미한 상관관계를 가진다고 하였다. 피어슨 상관계수의 값을 분석한 결과, 상관관계가 가장 높은 키워드 쌍은 metadata-dublincore(0.8426)이며, 최댓값을 제외하고 상관계수 값이 0.7 이상인 키워드 쌍은 총 5쌍으로 golden road-green open access (0.83907), oa article-citation counts(0.80733), research article-web citations(0.78096), oa article-citation advantage(0.77018), oa article-citation impact(0.74645)순으로 나타났으며, Guilford의 해석에 따라 이들은 정적인 높은 상관관계를 갖는다. 음의 상관관계가 가장 큰 키워드 쌍은 protocol-scholarly literature(-0.06297)이며, digital preservation-scholarly literature (-0.05762), digital preservation-bibliographic database(-0.05178) 순으로 나타났다. 이들은 Guilford의 해석에 따라 음의 무시할 만한 경미한 상관관계를 가지고 있다. 이 값들을 제외하고 음의 상관관계를 지닌 키워드 쌍은 총 31개였다.

4. 오픈 액세스 분야의 지적구조 분석결과

4.1 네트워크 분석에 의한 지적구조

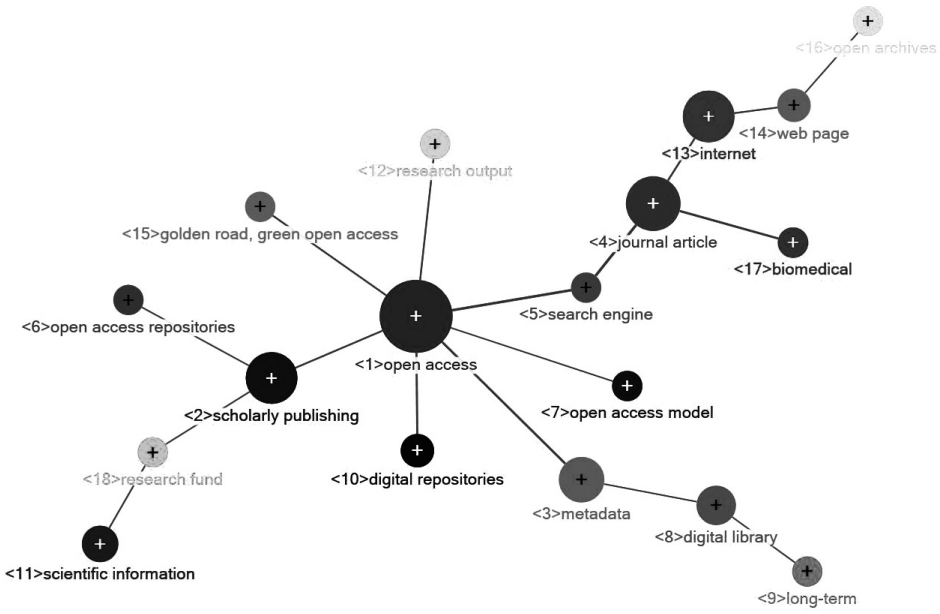
동시출현단어 분석은 네트워크 지도의 시각화를 통하여 선정된 분석 분야의 내부 개념 사이의 관계를 조사할 수 있으며, 네트워크를 여러 개의 군집으로 분할하여 세부 주제 영역을 살펴볼 수 있다. 본 연구에서는 오픈 액세스 분야의 지적구조를 규명하기 위하여 핵심 키워드들의 코사인 유사도 행렬을 산출하였고, 이를 입력데이터로 하여 $r=\infty$, $q=n-1$ 조건의 패스파인더 네트워크 알고리즘을 적용하여 네트워크를 생성하였다. 그리고 패스파인더 네트워크상에서 군집들을 식별하여 세부 주제를 분명하게 파악하고 주제 분석의 식별력을 높이기 위하여 이재윤(2006b)이 제안한 기법인 병렬 최근접 이웃 클러스터링 알고리즘(PNNC)을 적용하였다. 네트워크를 생성하고 클러스터링을 하기 위하여 이재윤의 WNET ver 0.4 프로그램을 사용하였고, 이를 시각화하기 위하여 NodeXL 프로그램(Hansen, Shneiderman, & Smith 2011)을 사

용하였다.

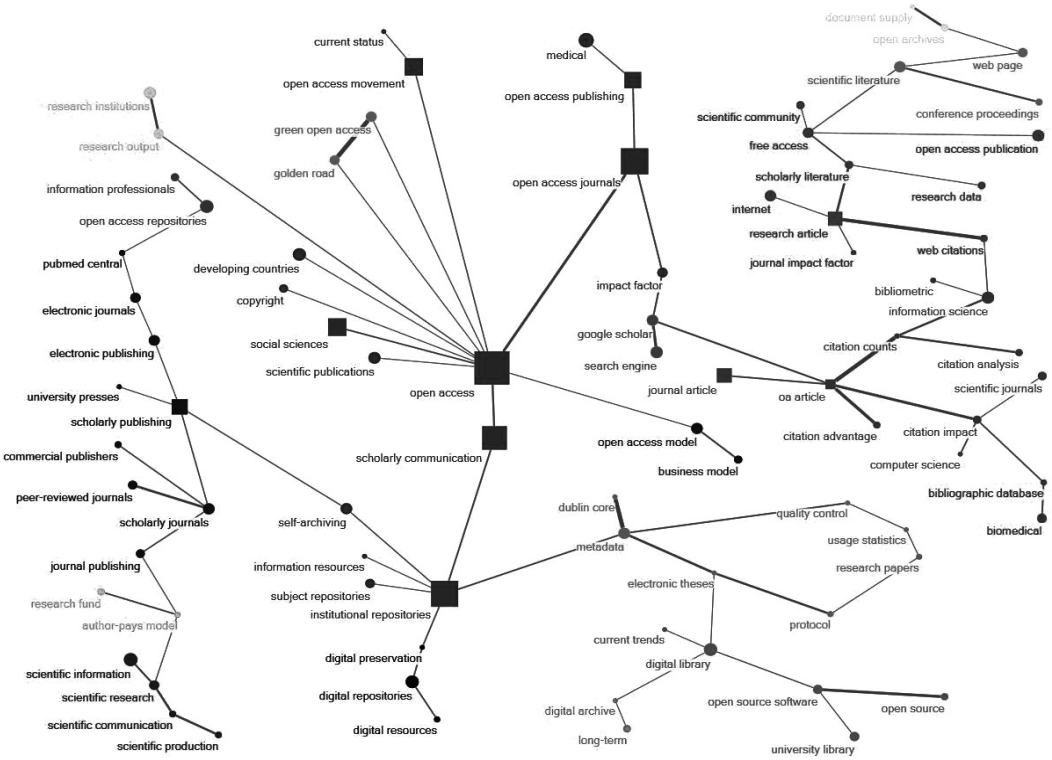
키워드 84개에 관한 동시출현단어의 네트워크 분석 결과, PNNC알고리즘으로 18개의 최적의 군집이 생성되었다. 〈그림 2〉는 18개의 군집의 각 세부 영역을 병합하여 나타내었으며 각 군집의 번호를 부여하였다. 각 군집에서 빈도수가 가장 높은 키워드를 각 군집을 대표하는 주제명으로 부여하였고, golden road와 green open access의 빈도수는 12로 동일하였으므로 golden road, green open access로 표현하였다. 〈그림 2〉의 병합을 해제하여, 〈그림 3〉과 〈그림 4〉와 같이 각 18개의 군집에 속한 하위 주제영역을 나타내고 네트워크 중심성 분석을 통하여 오픈 액세스 분야의 전역중심성이 높은 주제어와 지역중심성이 높은 주제어, 매개중심성이 높은 주제어를 확인하였다. 전역중심성이 높은 키워드는 오픈 액세스 연구 영역 전반에 걸쳐 다른 키워드들과 폭넓게 연계되어 있는 주제어이고, 지역중심성이 높은 키워드는 자기가 속한 군집내의 세부 연구 영역에서 영향력이 있는 키워드이다. 또한 매개중심성은 네트워크를 구성할 때 각 노드들을 연결해 주는 중개자 역할을 하는 주제어이다.

동시출현단어 분석은 동시에 출현한 빈도수에 의한 연결 강도를 나타내는 가중 네트워크이므로 중심성 분석에는 가중 네트워크에 적용되는 이재윤(2006c)이 제안한 중심성 분석 척도를 사용하였다. 전역중심성이 높은 키워드들을 확인하기 위하여 삼각매개중심성과 평균연관성을 측정하였다. 그리고 링크의 굵기에 키워드 간의 빈도에 의한 연관도 가중치를 반영하여 연결 강도를 나타냈다.

키워드들의 각 전역중심성 지수의 측정된 결과 값을 반영하여 네트워크 지도를 〈그림 3〉과



〈그림 2〉 PFNET 지도에 나타낸 18군집



〈그림 3〉 전역중심성에 의한 키워드 관계 네트워크

같이 작성하였다. 상대적 삼각매개중심성 값은 노드의 크기에 반영하였고, 평균연관성은 중심성 값이 0.05이상인 노드들의 형태를 사각형으로 바꾸어 표현하였다. 상대적 삼각매개중심성 지수와 평균연관성 지수를 비교하기 위해 상위 10위 이상까지의 결과 값들을 살펴본 후, 측정값의 상위 11위까지가 0.05 이상이므로 이것을 기준으로 잡았다. 이로써 두 전역중심성 지수를 비교해 볼 수 있으며, 오픈 액세스 전역에 중심이 되는 키워드들을 구별하여 확인할 수 있다.

다음 〈표 2〉와 같이 평균연관성 지수가 0.05 이상인 노드는 11개였으며, 이를 기준으로 상대적 삼각매개중심성 값의 상위 11위까지의 노드와 비교하였을 때, 각 전역중심성 지수의 상위 4위까지는 순서가 동일하였으나 5위부터 11위까지의 순위는 서로 일치하지 않았다. 상대적 삼각매개중심성 지수의 상위 11위까지에는 키워드 medical(10)이 포함되어 있었고, 평균연관성 지수의 상위 11위까지에는 키워드 medical(10)이 아닌 oa article(20)이 포함되어 있었다. 키워드 medical(10)은 평균연관성 지수에서는 18위를 oa article(20)은 상대적 매개중심성 지수에서는 34위를 차지하였다. 일반적으로 상대적 매개중심성와 평균연관성으로 전역중심성을 측정할 수 있기 때문에 키워드 medical(10)과 oa article(20)를 모두 상위 11위 안에 포함되는 전역중심성이 높은 주제어로 판단하여 해석하였다.

〈표 2〉 가중 네트워크에 대한 두 가지 전역중심성 지수의 상위 11위

순위	키워드(번호)	상대적 삼각매개 중심성(rtbc,0~1)	키워드(번호)	평균연관성 (AVGSIM)
1	open access(1)	0.88775	open access(1)	0.13834
2	open access journals(3)	0.55422	open access journals(3)	0.08828
3	institutional repositories(2)	0.54246	institutional repositories(2)	0.07992
4	scholarly communication(4)	0.44578	scholarly communication(4)	0.06742
5	social sciences(7)	0.24743	scholarly publishing(8)	0.05334
6	open access movement(6)	0.24567	journal article(12)	0.0527
7	open access publishing(5)	0.20893	research article(32)	0.05263
8	scholarly publishing(8)	0.18542	open access publishing(5)	0.05259
9	medical(10)	0.17308	social sciences(7)	0.05115
10	journal article(12)	0.16603	open access movement(6)	0.05103
11	research article(32)	0.1481	oa article(20)	0.05021

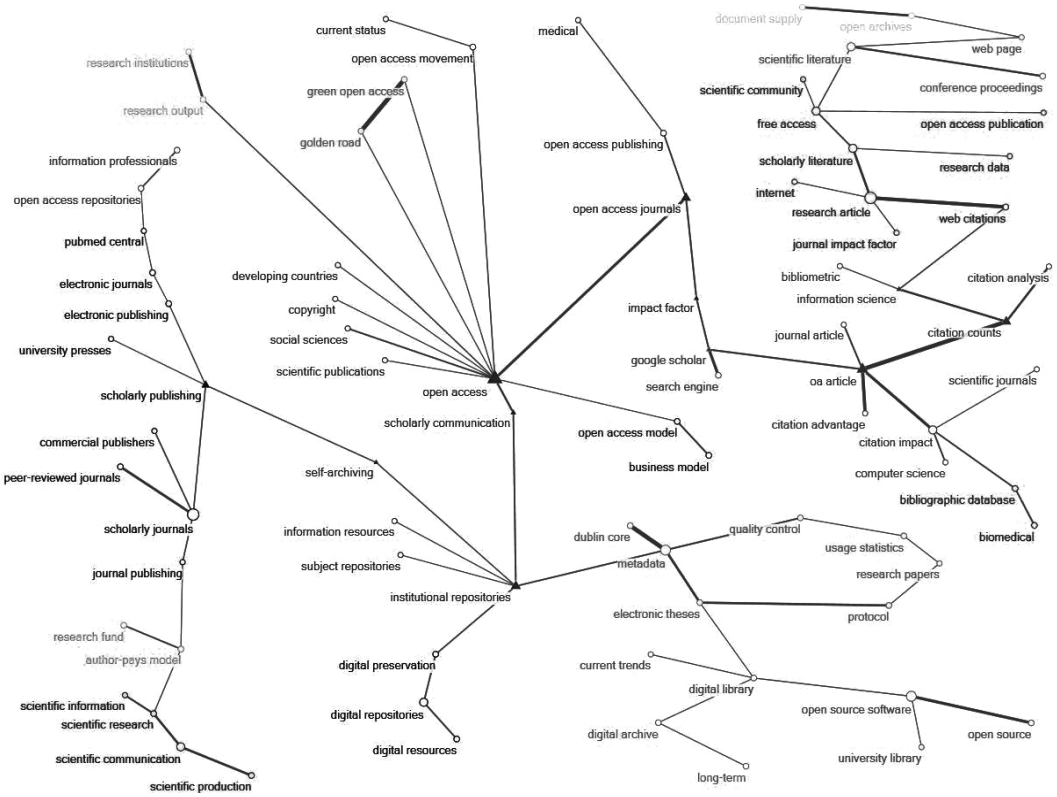
이 전역중심성 지수 상위 11위까지의 순위를 바탕으로 오픈 액세스 분야에서 open access(1)를 중심으로 키워드 open access journals(3), institutional repositories(2), scholarly communication(4), social sciences(7), open access movement(6), open access publishing(5), scholarly publishing(8), medical(10), journal article(12), research article(32), oa article(20)는 다른 주제어들과 광범위하게 연관되어 다루어지는 주제어들이다.

다음으로 네트워크 기반으로 형성된 18개 각 군집의 세부 주제 영역 분석을 위해 지역중심성 지수를 측정하였다. 군집 내의 중심 키워드를 파악하여 이를 중심으로 군집의 세부 키워드들을 해석하고자, 상대적 최근접이웃중심성 값을 구하였다. 상대적 최근접이웃중심성은 최근접이웃중심성의 값을 정규화 한 것으로 군집에서 영향력이 높은 키워드를 확인할 수 있는 지역중심성의 지표이다. 각 군집에서 지역중심성이 가장 높은 중심 주제어는 제 1군집 open access(1), 제 2군집 scholarly journals(39), 제 3군집 metadata(9), 제 4군집 oa article(20), 제 8군집 open source software(49), 제 10군집 digital repositories(22), 제 11군집 scientific communication(38), 제 13군집 research article(32), 제 14군집 scientific literature(60)이었으며, 제 5, 6, 7, 9, 12, 15, 16, 17, 18군집은 군집에 속한 키워드 수가 2개이며, 두 키워드의 상대적 최근접이웃중심성 지수가 동일하게 측정되었다.

또한 NodeXL을 이용하여 매개중심성 지수를 측정하였다. 이를 통해 오픈 액세스 연구 영역에서 각 군집들을 연결해 주는 역할을 하는 키워드들을 확인하였다. 오픈 액세스 분야에서 각 주제어들을 연결해 주는 역할을 하는 키워드들은 매개중심성 값 상위 11위까지를 기준으로 open access(1), institutional repositories(2), scholarly communication(4), open access journals(3), oa article(20), impact factor(42), google scholar(19), citation counts(57), scholarly publishing(8), self-archiving(11), information science(16) 순이었다.

〈그림 4〉와 같이 지역중심성 값에 비례하도록 노드의 크기를 설정하고, 매개중심성 값이 1.000 이상인 11개의 노드들을 삼각형 형태로 표시하여 네트워크 지도를 작성하였다.

제 1군집은 open access(1)에 대한 논의와 동향을 다루고 있는 전반적인 주제어들이다. 이 군집에 속해 있는 키워드들을 바탕으로 오픈 액세스에서 포괄적으로 다루어지는 세부 주제 분야를 요약 할 수 있다. 이 군집에는 전역·지역·매개중심성 지수가 1위부터 4위까지인 키워드들을 포함하고 있으며, 그 안에서 약간의 순위 변동이 있지만 이를 바탕으로 이 키워드 open access(1), open access journals(3), institutional repositories(2), scholarly communication(4)들이 오픈 액세스 분야 연구의 핵심이며, 가장 대두되는 주제어들임이 틀림없음을 확인할 수 있다. 또한 다른 노드들보다 오픈 액세스와 오픈 액세스 저널 사이의 링크 굵기에 반영된 가중치가 눈에 띄게 높았으며, 이를 바탕으로 이 두 주제어가 강한 연결 관계를 가지고 있고, 오픈 액세스 저널은 오픈 액세스 영역에서 독립된 연구 영역이라기보다는 자주 함께 다루어지는 연구 키워드라는 것으로 이해할 수 있다. 이 군집의 키워드 사회 과학(social sciences(7))의 경우 문헌 빈도수 29회로 빈도 순위 7위를 차지하고 있으며, 이는 오픈 액세스



〈그림 4〉 지역, 매개중심성에 의한 키워드 관계 네트워크

가 STM(Science, Technology and Medicine) 저널에서 차지하는 그 의미의 비중에 비해 지나치게 높은 빈도수와 전역중심성 지수를 가지고 있었다. 키워드 social sciences(7)가 포함된 29건의 문헌을 검토한 결과, '사회 과학 분야에서 셀프아카이빙 실현과 출판사 정책의 영향' 등과 같은 엄격한 기준에서 오픈 액세스 기반으로 사회 과학 분야를 접근한 연구는 2개였다. 10건의 문헌은 과학과 사회 과학 분야로 분석 대상을 구분하여, 비교 및 분석이 이루어지는 연구들이었고, 계량정보학 측면에서 다양한 분야를 분석 대상으로 할 때 그 중의 한 분야로 출현한 경우가 2건, 기타로 초록에서 연구의 의의나 결과를 제시할 때 언급되는 등의 경우가 6건이었다. 남은 9건은 인문 사회 과학을 의미하는 'social sciences and humanities' 또는 'humanities and social sciences'의 형태로 각 제목 또는 초록에 포함되어 있었다. 이는 본 연구에서 키워드를 명사구 형태로 추출하였고, 단어와 명사구를 이어주는 접속사 'and'로 인해 'social sciences'만 추출된 것임을 알 수 있다. 이를 근거로 키워드 social sciences(7)가 오픈 액세스 연구 영역에서 다양하게 쓰이고 있음을 확인할 수 있다.

제 2군집의 주제어들은 학술 출판과 그를 둘러싸고 있는 세부 주제어들이다. 이들의 연관성을 바탕으로 제 2군집은 오픈 액세스의 등장과 발달에 관한 학술 출판의 역할 변화와 관련한 주제어들로 구성되어진 군집으로 분류할 수 있다. 이 군집의 생성을 바탕으로 새로운 기술과 오픈 액세스의 도입에 따라 기존의 학술 출판 시스템에서 새롭게 변모된 학술 출판의 흐름에 관한 연구들이 부분적으로 진행되고 있음을 확인할 수 있다.

제 3군집은 제 1군집의 키워드 기관 리포지터리와 연결되어 있으며, 이 군집의 형성을 통해 리포지터리 개발 및 보급에 관하여 기술적인 측면에서 접근하여 메타데이터의 공유 및 연계에 관한 연구들이 부분적으로 실시되고 있음을 유추할 수 있다.

제 4군집에는 전역중심성이 높은 주제어 journal article(12)이 포함되어 있지만, 이 군집 내에서 가장 영향력이 있는 주제어는 지역중심성이 높은 oa article(20)이다. 이 군집을 구성하고 있는 주제어들을 기반으로 오픈 액세스 영역에서 오픈 액세스 논문을 대상으로 계량서지학(bibliometric(66)) 분야의 연구들이 활발하게 진행되고 있음 확인할 수 있다. 키워드 oa article(20)과 인용 빈도(citation counts(57)), 인용 영향력(citation impact(46)), 인용의 장점(citation advantage(58)) 간의 링크의 굵기를 바탕으로 연결 강도가 높은 관계이며, 정보학 분야에서 연구자들이 이 키워드들을 자주 함께 사용하고 있음을 확인할 수 있다.

제 5군집은 search engine(13)과 google scholar(19)로 구성되어 있으며, 제 1군집의 키워드 impact factor(42)와 링크가 형성되어 있다. 이 군집에 속하는 두 키워드가 빈도수 상위 20위 안에 들어 있는 것을 바탕으로 새로운 과학 검색서비스를 제공해주는 구글 학술검색엔진을 이슈로 다수의 연구들이 수행되고 있으며, 영향력 지수와의 연결을 바탕으로 계량정보학 측면에서 접근하는 연구들에서 학술 데이터베이스로서 구글 스칼라가 자주 선정되는 데이터베이스임을 유추할 수 있다.

제 6군집은 제 2군집인 학술 출판과의 상호

관계가 형성되었다. 분석을 위해 오픈 액세스 리포지터리와 함께 군집을 이루고 있는 키워드 information professionals(63)가 포함된 문헌들의 내용을 추적한 결과, 오픈 액세스를 둘러싸고 있는 주제어들과 향후 정보전문가들을 위한 기술 및 역량들을 제시하는 연구들이 대부분이었다.

제 7군집은 오픈 액세스의 전반적인주제어들을 포함하고 있는 제 1군집에서 독립적으로 나누어진 군집이다. 이 같은 제 7군집의 생성으로 오픈 액세스 분야에서 비즈니스 측면에서 접근하여 오픈 액세스의 모델과 그에 따른 경제 모형을 다루는 연구가 별도로 꾸준히 진행되고 있음을 유추할 수 있다.

제 8군집이 제 3군집인 metadata(9)와의 링크가 형성되어 있는 것을 바탕으로 같은 맥락에서 오픈 액세스 분야에서 디지털도서관에 관한 연구들이 기술적인 측면에 치중되는 편이라고 해석할 수 있다.

제 9군집은 제 8군집의 키워드 digital library(17)를 매개로 연결되어 있다. 관련연구들에서 디지털 객체의 접근 및 메타데이터의 관리, 제공에 관한 주제를 다루고 있으며, 키워드 digital archive(80)와 장기 저장 및 보존(long-term(21))을 중심으로 흐르고 있음을 유추 해석할 수 있다.

제 10군집은 제 1군집의 키워드 institutional repositories(2)를 매개로 하여 "디지털"을 중심으로 분리된 영역이다. 이 군집에 "digital"이라는 용어가 집중적으로 배치되어 있는 것은 기관 리포지터리의 연구 영역에서 전자 자원과 그에 관한 연구 또는 프로젝트들이 독립적으로 이루어지고 있음을 파악할 수 있다. 대표적인 예로 MIT 대학 도서관과 HP가 공동으로 개발한 Dspace를 근거로 제시할 수 있다. Dspace의 주요 특징으로는 디지털 보존(digital preservation(65))을 들 수 있으며, Dspace는 수록되는 정보원인 디지털 자원(digital resources(64))의 장기적 저장 및 관리 기능을 제공한다.

제 11군집에서 가장 영향력 있는 주제어는 지역중심성이 가장 높은 scientific communication(38)이다. 이 군집의 생성은 2000년 10월에 출범한 PLoS(Public Library of Science)가 배경이 되는 것을 뒷받침 할 수 있다. PLoS는 과학의 발전, 교육, 공익을 위하여 과학 문헌이 전 세계 과학자와 공공에게 자유롭게 접근될 수 있기를 바라는 과학자들을 주축으로 결성된 비영리조직이다. 이들은 온라인상에서 과학 분야 공공도서관을 구축한다는 목적을 지니고 있다. 생성된 scientific information(25) 군집의 세부 영역이 scientific으로 시작되는 용어들의 출현으로 이루어져 있는 것을 기반으로 오픈 액세스의 배경에 과학 분야와 과학정보의 자유로운 유통이 차지하는 비중이 크다는 것을 알 수 있다.

제 12군집은 제 1군집 오픈 액세스와의 연결을 토대로 생성된 군집으로 이 군집의 생성은 오픈 액세스 분야에서 연구 성과 정보 오픈 액세스와 관련한 연구 및 프로젝트들이 부분적으로 다수 수행되고 있다는 실증적인 근거가 된다.

제 13군집 내에서 가장 영향력 있는 주제어는 research article(32)이며, 키워드 web citations(53)를 매개로 제 4군집의 키워드 information science(16)와 연결되어 있다. 이 군집은 세부 주제어들의 구성만으로 생성의 의의를 직관적으로 파악하는 것이 쉽지 않다. 키워드 research

article(32)과 키워드 web citations(53)의 관계는 1차 연관성 행렬인 코사인유사도의 결과값에 따라 주제적 연관성이 높으며, 2차 연관성인 피어슨 상관계수를 이용한 행렬의 결과 값 역시 정적인 높은 상관관계로 나타났다. 이 군집은 제 4군집과 제 14군집과의 네트워크 구성이 형성되어 있으며, 이를 바탕으로 크게 계량정보학 분야에서 오픈 액세스 영역을 분석 대상으로 하여 사용되는 주제어들로 유추할 수 있다.

제 14군집에서 지역중심성 지수가 가장 높은 키워드는 scientific literature(60)이다. 이 군집의 생성은 오픈 액세스 영역에서 과학 문헌을 배포하기 위한 새로운 경로로 회의록과 웹페이지가 있다는 것을 확인할 수 있으며, 특히 두 키워드 과학문헌과 회의록 사이의 굵은 링크는 서로 연관도가 높은 관계임을 파악할 수 있다.

제 15군집은 제 1군집과의 네트워크 관계를 바탕으로 독립되어진 군집이다. 두 키워드 golden road(44)와 green open access(43)의 코사인 유사도와 피어슨 상관관계에 따른 가중치를 반영할 때, 오픈 액세스 영역에서 두 주제어는 동시에 자주 출현되는 용어로서 주제적 연관성이 매우 높은 것을 알 수 있다.

제 16군집 생성의 의의를 찾기 위해 두 키워드 open archives(50)와 document supply(52)를 포함한 문헌을 추적한 결과, 2010년 프랑스의 학술 상호대차와 문헌제공에 관한 보고서(A review of interlending and document supply in France: 2010)에서 오픈 아카이브가 변화를 겪으면서 상당히 발전하고 있지만 아직까지 현실적인 대안을 제공하지 못하고 있으며, 전통적인 상호대차와 문헌제공에 통합되지는 않는다고 제언하였다.

제 17군집은 제 4군집의 키워드 citation impact(46)와 연결되어 있으며, 2개의 키워드로 구성된 독립된 군집이다. 오픈 액세스 활동의 배경에는 과학 분야뿐만 아니라 생의학(biomedical(51)) 분야 연구자들 역시 주도적인 역할을 맡고 있다. 생의학 과학자들은 학제적인 학문 분야에서 커뮤니티의 기량을 공유하는 것의 필요성을 느꼈고, 이에 오픈 액세스에 대한 관심이 고조되었다. 그러나 키워드 pubmed central(62)는 제 2군집에 속해있는 반면에, 키워드 biomedical(51)은 제 16군집에 포함되어 있다. 그리고 키워드 biomedical(51)의 매개중심성 값이 0인데 반해, 같은 군집내의 키워드 bibliographic database(84)의 매개중심성 지수가 82이며, 키워드 citation impact(46)와의 링크가 남아있다. 이를 바탕으로 수집된 논문들을 검토한 결과, 각 주제 분야별로 데이터베이스를 선정하여 오픈 액세스 저널과 논문에 관한 계량서지적 연구가 실시되고 있으며, 특정한 주제 분야인 생의학 분야를 대상으로 한 연구들이 수행되고 있음을 확인할 수 있었다.

제 18군집의 research fund(59)와 author-pays model(67)의 관계는 오픈 액세스를 둘러싸고 있는 논쟁 중에서 경제 모델과 관련 있는 주제어이다. 그러나 오픈 액세스 모델과 경제 모델이라는 키워드를 포함하고 있는 제 7군집이 생성되어 있으며, 제 7군집의 생성으로 오픈 액세스 분야에서 오픈 액세스의 모델과 그에 따른 경제 모형을 다루는 연구가 별도로 진행되고 있음을 확인할 수 있었다. 제 18군집이 제 7군집이 아닌 제 2군집의 키워드 journal publishing

(35)과 연결되어 더 강하게 네트워크를 구성하고 있는 것을 바탕으로 학술 출판 영역에서 연구자금제공자와 저자 지불형 모델을 둘러싼 논쟁과 취해야 할 전략 및 발전 방향에 관한 연구가 진행되고 있음을 유추 해석할 수 있다.

지금까지 키워드 84개의 네트워크 지도를 기반으로 병렬 최근접 이웃 클러스터링 알고리즘에 의해 형성된 18개 군집의 세부 주제 영역을 확인하였다. 또한 네트워크 중심성 분석을 통해 오픈 액세스 분야의 핵심 주제어들과 각 군집내의 영향력이 높은 주제어들 그리고 각 군집을 연결 시켜주는 매개 주제어들을 확인하여 네트워크 기반 오픈 액세스 분야의 지적구조를 규명하였다.

4.2 군집분석에 의한 지적구조

네트워크 분석 알고리즘에 의해 형성된 군집의 결과를 보완하기 위하여, 산출된 2차 연관성 행렬을 가지고 통계프로그램 SPSS ver 20.0을 이용하여 군집분석을 실시하였다. 클러스터링 알고리즘으로는 계층적 클러스터링 기법인 Ward 기법을 사용하였으며, z점수로 표준화하고 제곱 유클리디안 거리를 사용하여 덴드로그램으로 나타내었다.

네트워크 분석 및 선행연구들의 검토를 기반으로 오픈 액세스의 연구 경향을 잘 나타내주는 적절한 군집의 수를 최종 4개로 결정하였다. 군집명은 네트워크 분석과 선행연구를 기반으로 전문가 검토 후, 각 군집에 속한 전체 키워드를 대표할 수 있는 단어를 사용하여 부여하였다. 이를 〈표 3〉과 같이 정리하였으며, 〈표 3〉은 군집분석에 의한 오픈 액세스 분야의 4개의 군집별 세부 주제 분야이다.

군집분석에 의해 형성된 계층적 분류는 수행 알고리즘과 통계프로그램 SPSS ver 20.0에 의해 나온 결과이기 때문에 제 1군집, 제 2군집, 제 3군집, 제 4군집과 같은 군집의 순서와 관계없이 주제 영역인 오픈 액세스에 관한 연구의 흐름을 바탕으로 각 군집의 살펴보면 다음과 같다.

제 1군집 Institutional Repositories에서는 기관 리포지터리와 관련한 주제어들을 확인할 수 있다. 연구자가 self-archiving(11)을 하기 위해서는 institutional repositories(2) 즉, 기관의 커뮤니티 구성원이 연구 성과물을 디지털 형태로 생성, 수집, 보존, 관리, 유통시킬 수 있는 저장소가 필요하다. 오픈 액세스기반의 지식정보 유통체계를 구축하기 위해서 리포지터리는 주제 기반, 대학, 연구기관 또는 상업적 기관, 국가 별로 접근할 수 있다. 그러므로 리포지터리의 개발 및 보급을 위해서 각 운영 주체에 따라 목적성을 가지고 추진 체계과 전략을 제시하기 위한 연구들이 다수 진행되고 있다. 초기에 구축된 기관 리포지터리에 대한 사례 조사에 관한 연구들, 각 기관 리포지터리들과 중앙 리포지터리 및 국가별 상호 연계를 위한 표준 메타데이터 형식에 관한 연구들, 기술적 측면에서 공유 인프라 구축을 위한 정보 자원의 수집, 오픈 소스 소프트웨어 활용에 관한 연구들과 기관 리포지터리와 연구자를 둘러싼 저작권에 관한 연구 등이 그 예이다.

제 2군집 Informetric Analysis on Open Access는 국외의 문헌정보학 분야에서 오픈 액세스를 주제로 하여 다수의 계량정보학적 접근이 시도되고 있음을 확인할 수 있는 뚜렷한

〈표 3〉 군집분석에 의한 4군집의 세부 주제 분야

군집명	세부키워드(키워드번호)	군집명	세부키워드(키워드번호)
제 1군집 (27개) Institutional Repositories	institutional repositories(2) metadata(9) self-archiving(11) open access repositories(14) digital library(17) long-term(21) digital repositories(22) copyright(24) university library(29) web page(36) open source(40) subject repositories(48) open source software(49) research data(55) information resources(56) information professionals(63) digital resources(64) digital preservation(65) scientific production(68) electronic theses(70) dublin core(71) current trends(72) usage statistics(74) quality control(77) protocol(78) digital archive(80) conference proceedings(81)	제 3군집 (20개) Open Access Journals	open access(1) social sciences(7) open access model(15) business model(23) scientific information(25) electronic journals(28) scientific publications(30) scientific journals(34) journal publishing(35) scientific research(37) scientific communication(38) green open access(43) golden road(44) free access(45) open archives(50) document supply(52) research fund(59) scientific literature(60) pubmed central(62) author-pays model(67)
제 2군집 (20개) Informetric Analysis on Open Access	open access journals(3) open access publishing(5) medical(10) journal article(12) search engine(13) information science(16) google scholar(19) oa article(20) research article(32) impact factor(42) citation impact(46) citation analysis(47) biomedical(51) web citations(53) citation counts(57) citation advantage(58) scholarly literature(76) journal impact factor(79) computer science(82) bibliographic database(84)	제 4군집 (17개) Scholarly Publishing	scholarly communication(4) open accessmovement(6) scholarly publishing(8) developing countries(18) research output(26) internet(27) research institutions(31) electronic publishing(33) scholarly journals(39) open access publication(41) scientific community(54) research papers(61) bibliometric(66) peer-reviewed journals(69) current status(73) university presses(75) commercial publishers(83)

예가 된다. 오픈 액세스 도입 이후, 오픈 액세스는 학술 정보 유통에 있어 점차 주요 정보원으로 인식되어가고 있다. 따라서 급격하게 증가한 오픈 액세스 논문들에 대한 질적인 요소를 평가하거나 학술 이용자의 이용행태를 분석할 필요가 있으며, 이는 인용 분석을 연구하는 연구자들에게 다양한 기회를 제공시켜준 셈이다.

제 3군집 Open Access Journals에서는 오픈 액세스 저널과 관련이 있는 이슈들을 확인할 수 있다. 오픈 액세스 저널은 오픈 액세스 방식으로 수록된 논문을 이용시키는 저널이다(golden road). 오픈 액세스 저널은 제한 없는 무료접근이 가능한 학술지로, 특히 저작권으로 인한 접근과 이용 제한이 발생하지 않아야하며, 구독료나 접근 비용이 부과되지 않는 학술지를 말한다(정경희 2011). 과학 분야와 의학 분야에서 오픈 액세스 출판사들이 다수의 OAJ를 출판하였으며, 이러한 OAJ는 새로운 비즈니스 모형을 요구한다. 그에 따라 제안되고 채택된 오픈 액세스 모델들과 모형 개발, 비용부담의 문제, 기존 무료 액세스(free access(45)) 저널의 오픈 액세스화 추진 등 OAJ와 관련한 연구들이 다양한 형태로 전개되고 진행 중임을 확인할 수 있다.

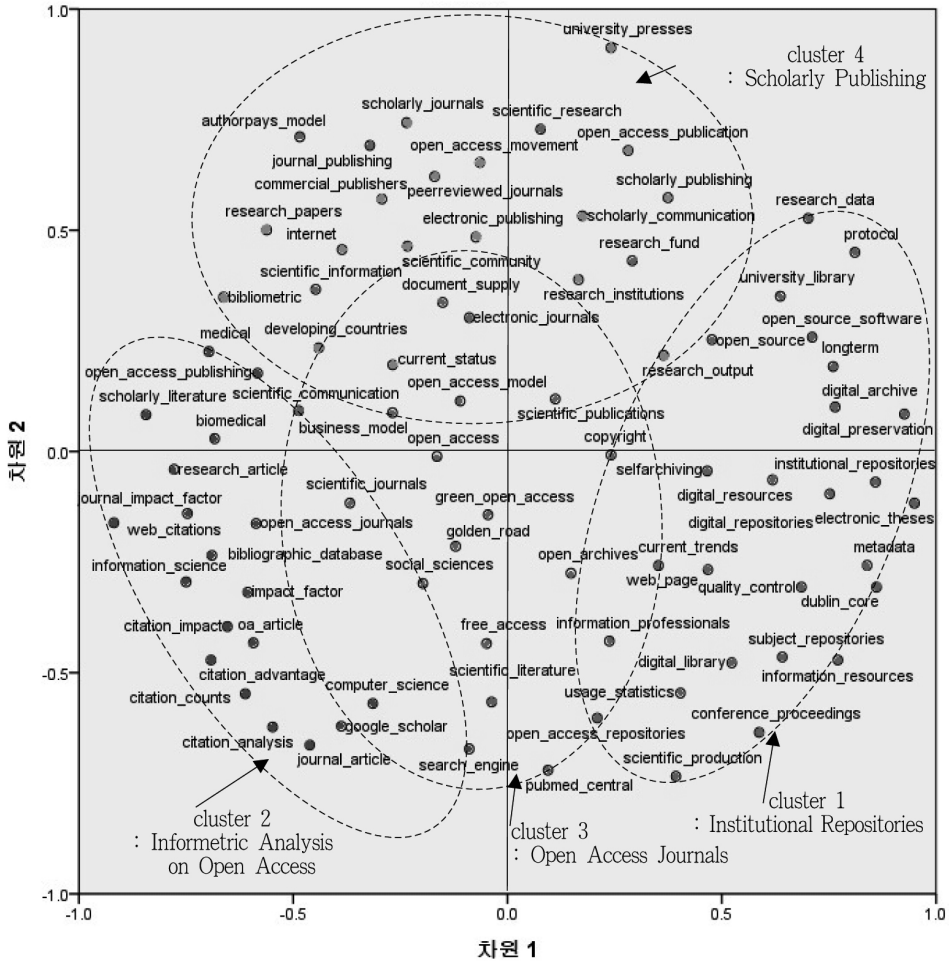
마지막으로 오픈 액세스는 복잡한 그 등장의 배경에 관하여 기술, 경제, 사회, 법적인 측면에서 접근할 수 있다. 이에 대한 연구의 흐름은 학술 커뮤니케이션의 새로운 패러다임인 오픈 액세스의 동향 분석 및 정책마련으로 진행되고 있다. 따라서 오픈 액세스와 관련한 제반요소를 살펴보기 위해서는 전통적인 학술 커뮤니케이션의 기능 및 관련 요소들의 역할을 숙지하고, 새롭게 변화된 학술 커뮤니케이션의 추세에 관한 연구가 필수적이다. 생성된 제 4군집 Scholarly Publishing은 이 같은 견해를 뒷받침해주는 군집으로서 학술 커뮤니케이션의 공식 수단인 학술 출판의 역할과 대안 모색으로서의 오픈 액세스와 그 배경에 관련한 주제어들을 확인할 수 있다.

4.3 다차원축적지도에 의한 지적구조

군집분석을 기반으로 다차원축적지도를 작성하기 위해 통계프로그램 SPSS ver 20.0으로 PROXSCAL 알고리즘을 사용하면서 변수를 z 점수로 표준화하여 처리하였다. 본 연구에서의 스트레스 값은 0.08506이었으며, 이는 Kruskal이 제시한 스트레스 값을 기준으로 할 때, 대상들 간의 적합도는 보통이었다.

각 키워드들 간의 위치를 2차원 공간상에 점으로 나타냈고, 군집분석 결과에 따라 4개의 클러스터영역에 속한 점을 색깔별로 구분하였다. 이를 〈그림 5〉와 같이 키워드지도상에 나타냈으며, 각 클러스터영역의 경계를 점선으로 표시하고 대표되는 군집명을 표기하였다. 유사도에 의해 점으로 나타낸 각 키워드들의 좌표 값을 확인한 결과, 키워드지도상의 키워드의 위치는 키워드 사이의 상관관계에 따라 나타나는 것임을 알 수 있다.

군집분석 결과를 MDS 지도에 표시하여 해석한 결과, 지도의 X축을 기준으로 살펴보면 우측에는 군집분석에 의해 형성된 제 1군집인 Institutional Repositories의 세부 키워드들과 유사하게 키워드들이 분포되어 있는 것을 확인할 수 있다. 그리고 좌측에 위치한 키워드들은 제 2군집인 Informetric Analysis on Open



〈그림 5〉 MDS 지도에 나타낸 오픈 액세스 분야 지적구조

Access의 세부 키워드들과 유사하게 나타났다. 마지막으로 중앙부분에는 제 3군집인 Open Access Journals와 제 4군집인 Scholarly Publishing의 세부 키워드들이 두루 분포되어 있는 것을 확인할 수 있다.

키워드지도의 Y축(차원2) 기준으로 하단에는 제 1군집 Institutional Repositories의 세부 키워드 약 3분의 2 정도와 제 2군집 Informetric Analysis on Open Access의 세부 키워드들을 주로 확인할 수 있으며, 제 3군집 Open Access Journals의 세부 키워드들을 반 정도를 확인할 수 있다. 그리고 상단에는 제 4군집 Scholarly Publishing의 세부 키워드들이 주로 위치하고 있음을 확인할 수 있다. Y축을 기준으로 중앙부분에서 Y축 상향으로 갈수록 전통적인 학술 커뮤니케이션의 대안 모색으로서 오픈 액세스 기반 학술 출판에 관한 연구를 진행하는 경향이 있다.

다차원축척지도에서의 제 2군집 Informetric

Analysis on Open Access의 세부 키워드들은 좌측 하단에 위치하여 다른 군집의 키워드들과 비교적 확연하게 분리되어 있음을 확인할 수 있다. 그러나 제 3군집인 Open Access Journalsl과 제 4군집인 Scholarly Publishing의 세부 키워드들은 지도 중심부의 상단과 하단에서 서로 조금 겹쳐져서 위치해 있다. 이는 제 3군집과 제 4군집의 주제 분야 영역이 다른 모든 연구 영역과 상관관계가 높으며, 연구의 중심축에 위치하고 있음을 보여주는 결과이다. 제 1군집인 Institutional Repositories의 세부 키워드들은 우측 상단과 하단에 위치하며 비교적 차별성이 드러나 있지만, X축, Y축을 기준으로 중앙부분의 키워드 copyright과 Y축 하단의 키워드 information professionals, open access repositories와 같이 간혹 겹쳐지는 키워드들을 확인할 수 있다. 각 군집을 묶었을 때, 군집에 속해 있지만 그 경계가 불분명하게 교집합을 이루듯이 위치해 있는 키워드들은 오픈 액세스에서 보편적으로 사용되는 주제어들임을 직관적으로 알 수 있다.

5. 결 론

본 연구는 Web of Science에서 오픈 액세스를 주제로 한 연구 문헌들을 수집하여, 동시출현단어기법을 활용하여 분석함으로써 오픈 액세스 연구 경향을 반영하는 지적구조 분석의 결과를 제시하였다. 네트워크 분석을 실시하여, 키워드들 간의 관계를 패스파인더 네트워크로 시각화하고, 병렬 최근접 이웃 클러스터링 군집으로 형성하여 오픈 액세스 분야의 세부 주제 영역을 살펴보았다. 중심성 분석으로 전역중심의 키워드와 지역 중심의 키워드, 매개 키워드를 확인하여, 오픈 액세스 분야의 핵심 주제어와 군집 내에서 영향력이 있는 주제어, 군집들의 매개가 되는 주제어를 파악하였다. 또한, 군집분석을 실시하여 네트워크 분석을 보완하였고, 군집분석의 결과를 다차원축척 지도에 반영하여 키워드지도를 통해 오픈 액세스 분야의 지적구조를 제시하고 세부 주제 영역의 구성을 규명하였다.

패스파인더 네트워크와 병렬 최근접 이웃 클러스터링 기법으로 키워드들 사이의 관계를 시각화 하고 군집을 형성한 결과, 18개의 군집으로 파악되었다. 중심성을 살펴보면, 전역중심성이 가장 높은 키워드는 open access였으며, 그 다음으로는 open access journals, institutional repositories, scholarly communication 등의 순으로 나타났다. 이러한 결과는 오픈 액세스 분야는 크게, 오픈 액세스 저널, 기관리파지토리, 학술커뮤니케이션 등으로 구성되었다고 볼 수 있다. 또한 이 결과는 매개중심성에서도 유사하게 나타났다. 즉 open access, institutional repositories, scholarly communication, open access journals 등의 순으로 나타났다.

군집분석 결과를 살펴보면, 총 4개의 군집으로 분류되었으며, 형성된 군집은 제 1군집 Institutional Repositories, 제 2군집 Informetric Analysis on Open Access, 제 3군집 Open Access Journals, 제 4군집 Scholarly Publishing으로 표현될 수 있다. 군집분석 결과를 MDS 지도에 표시하여 해석한 결과, MDS 지도상에서 군집분석에 의해 형성된 제 1군집과 제 2군집의 세부 키워드들이 확연하게 분리되어 위치

하였다. 지도상의 중심부에는 제 3군집과 제 4군집에 해당하는 키워드들이 위치하고 있으며, 이는 이 두 군집의 키워드들이 다른 모든 연구와 상관관계가 높으며, 오픈 액세스 분야 연구의 중심축에 위치하고 있음을 보여준다.

이와 같이 네트워크 기반의 지적구조와 군집분석과 MDS 지도에 나타낸 지적구조의 결과를 비교해 볼 수 있다. 네트워크 분석에 의해 구분되는 18개의 연구 영역과 군집분석에 의해 구분되는 4개의 연구 영역을 비교할 때, 전체 오픈 액세스 분야의 연구 경향에 따른 세부 주제 키워드는 약 60퍼센트 일치한다. 이는 두 분석 기법의 알고리즘에 의한 차이로 판단되며, 세부 연구 영역에 대한 해석에 관하여 융통성을 가질 필요가 있다.

이상의 결과를 종합해 보면, 1998년부터 2012년까지 문헌정보학 범주에서 수행된 오픈 액세스 분야의 핵심적인 연구 영역은 오픈 액세스 기반의 학술 출판을 둘러싼 연구들을 중심으로 기관 리포지터리에 관한 연구 영역, 오픈 액세스 저널과 논문을 분석 대상으로 실시되는 계량정보학적 연구 영역들이다. 본 연구에서는 오픈 액세스에 관한 주제를 다루는 문헌 수집의 범주를 문헌정보학으로 한정하였고, 이러한 특성이 결과 해석에 있어서 제한적이라고 볼 수 있다. 오픈 액세스의 범주를 의학 분야 등으로 더 확대하여 분석을 실시한다면, 학제적 성격의 오픈 액세스 분야의 지적 구조로서 결과가 다르게 도출될 가능성이 있다. 그러나 본 연구는 오픈 액세스 분야가 먼저 이루어진 국외 문헌정보학기반 오픈 액세스 분야의 지적구조를 나타낼 수 있다는 점에서 의의를 가지며, 오픈 액세스 분야의 연구 방향성 모색에 유용하게 사용될 수 있을 것으로 기대한다.

참 고 문 헌

김희정. 2011. 네트워크 분석을 기반으로 한 웹 아카이빙 주제 영역 연구.『한국비블리아학회지』, 22(2): 235-248.

박재신, 정영미. 2010. 지구적 환경문제 해결을 위한 학술활동과 환경운동 경향 연구.『정보관리학회지』, 27(3): 83-102.

이재윤. 2006a. 지적구조의 규명을 위한 네트워크 형성 방식에 관한 연구.『한국문헌정보학회지』, 40(2): 333-355.

_____. 2006b. 지적구조 분석을 위한 새로운 클러스터링 기법에 관한 연구.『정보관리학회지』, 23(4): 215-231.

_____. 2006c. 계량서지적 네트워크 분석을 위한 중심성 척도에 관한 연구.『한국문헌정보학회지』, 40(3): 191-214.

장임숙, 장덕현, 이수상. 2011. 다문화연구의 지식구조에 관한 네트워크 분석.『한국도서관·정보학회지』, 42(4): 353-374.

정경희. 2011. 국내 오픈 액세스 학술지 특성에 관한 연구.『한국비블리아학회지』, 22(3): 373-391.

정용일, 이준영, 이방래, 유선희, 원동규, 정성

창, 주시형. 2005.『계량정보분석을 통한 지식의 mapping과 활용』. 서울: 한국과학기술정보연구원.

Ding, Y., G. G. Chowdhury, and S. Foo. 2001. "Bibliometric cartography of information retrieval research by using co-word analysis." *Information Processing & Management*, 37(6): 817-842.

Guilford, J. P. 1950. *Fundamental statistics in psychology and education*. New York: McGraw-Hill.

Hansen, D.L., B. Shneiderman, and M. A. Smith. 2011. *Analyzing social media networks with NodeXL: insights from a connected world*. MA: Morgan Kaufmann.

Koehler, A. E. C. 2006. "Some Thoughts on the Meaning of Open Access for University Library Technical Services." *Serials Review*, 32(1): 17-21.

Liu, G. Y., J. M. Hu, and H. L. Wang. 2012. "A co-word analysis of digital library field in China." *Scientometrics*, 91(1): 203-217.

Milojević, S., C. R. Sugimoto, E. J. Yan, and Y. Ding. 2011. "The cognitive structure of library and information science: Analysis of article title words." *Journal of the American Society for Information Science and Technology*, 62(10): 1933-1953.

Noyons, E. C. M. and A. F. J. van Raan. 1998. "Advanced mapping of science and technology." *Scientometrics*, 41(1-2): 61-67.

White, H. D. and B. C. Griffith. 1981. "Author cocitation: A literature measure of intellectual structure." *Journal of the American Society for Information Science & Technology*, 32(3): 163-171.

[프로그램]

이재윤. COOC ver 0.3.1 [cited 2012.9.22].

_____. WNET ver 0.4 [cited 2012.9.22].