서지결합분석을 통한 빅데이터 활용 분야 연구

An Identification on Big Data Application Fields by Utilizing Journal Bibliographic Coupling Analysis

이보람, 이화여자대학교 문헌정보학과, rajeunir@gmail.com

Boram Lee, Dept. of LIS Graduate School of Ewha Womans University

본 연구는 빅데이터의 처리·저장 등과 같은 기술적 측면이 아닌 분석·활용적 측면에 초점을 맞춰 관련 학문분야를 파악하고 분야 간 지적구조를 규명하고자 하였다. 연구 결과 빅데이터 관련 연구들이 주제분야에 따라 명백한 차이를 보이고 있음을 확인할 수 있었다. 주제범주 분석을 통해 공학·기술(34.60%), 사회과학(25.24%), 자연과학(23.14%), 의학·보건학(14.85%) 등은 관련 연구가 비교적 고르게 분포되어 있지만, 인문학(1.69%)과 농업과학(0.21%)은 연구가 미비함을 알 수 있었다. 네트워크 분석 결과 사회과학 분야(31.58%)에 비해 공학 및 자연과학 분야(68.42%)의 빅데이터 연구가 더 활발함을 확인할 수 있었다. 또한 공학 및 자연과학 분야 연구들은 다양한 주제분야를 다루는 반면 사회과학 분야에서는 아직 한정된 주제분야에서 연구가 진행되고 있음을 알 수 있었다.

1. 서 론

오늘날 인터넷의 대중화와 모바일 기기의 보급 등 정보통신기술의 발전으로 인해 데이터양이 폭발적으로 증가하고 그 유형이 다양화되었다. 지금까지의 방식으로는 이러한 대규모의 데이터 즉 빅데이터를 수집, 저장, 분석, 관리하는 것이 불가능해졌고, 이를 해결하기 위한 도구 및 기술의 개발과 연구가 활발히 일어났다. 또한 현상파악과 미래예측을 위한 데이터 기반 근거를 얻을 수 있어 학계뿐 아니라 산업, 언론, 정부 등 다양한 영역에서 빅데이터에 대한 관심을 나타내고 있다. 또한 빅데이터는 정치, 사회, 경제, 문화, 과학기술 등 전 영역에 걸친 학제적 성격을 지니고 있다(김현영 외 2014).

그러나 현재 빅데이터에 관한 연구 대부분은 빅데이터 처리·분석에 필요한 하드웨어, 소프트웨어, 분석방법 등 기술적 측면에 집중되어있다(김완종 2014). 빅데이터 연구에서 주목해야 할 점은 빅데이터 분석이 가져올 수 있는 데이터의 가치로서(이정미 2013), 이제는 빅데이터 활용이 각 분야에 어떤 도움을 줄 수 있는지에 대한 연구가 필요하다. 따라서 본 연구에서는 빅데이터의 기술적 측면을 배제한 활용적 측면에 초점을 맞춰 빅데이터 관련 연구가 이루어지고 있는 학문분야를 파악하고 서지결합 네트워크를 통해 분야 간의 지적구조를 규명하고자 하였다.

2. 연구방법

빅데이터 관련 분야를 파악하기 위해 본 연구에서는 Thomson Reuters의 인용색인 데이터베이스 Web of Science(WoS)를 검색도구로 선정하였다. 2016년 5월 23일 기준으로 WoS Core Collection에서 주제 필드에 “big data”를 질의어로 입력하였다. 검색기간은 전체로 설정하고, 문서유형은 학술논문으로 언어는 영어로

제한한 결과 총 2,555편의 논문이 검색되었다.

WoS로부터 검색된 논문 2,555편의 서지데이터와 59,912편의 관련 참고문헌 데이터를 반출하였다. 서지데이터 분석프로그램인 Bibexcel을 사용하여 해당 논문의 학술지명, 출판연도, Subject Categories (SC) 그리고 참고문헌의 학술지명을 추출하였다.

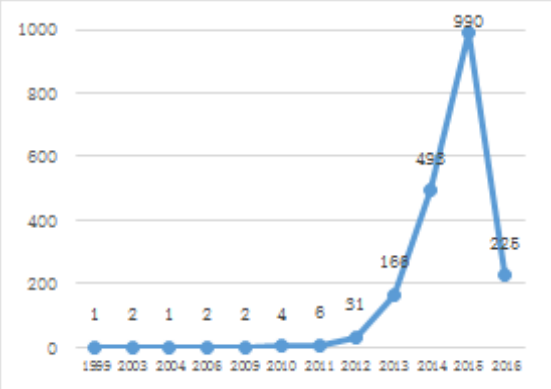
수집된 논문 중 빅데이터의 처리, 저장 등 기술적인 측면을 다루는 논문을 제외하기 위해서 SC가 컴퓨터공학인 학술지에 실린 논문을 파악한 결과 총 1,029편이 발견되었다. 이 중 주제분야가 단독으로 컴퓨터공학에만 해당하는 논문 632편은 분석대상에서 제외하였다. 그리고 컴퓨터공학 외에 다른 주제분야가 복수로 지정된 논문 397편은 제외하지 않고 컴퓨터공학을 제외한 나머지 주제분야만을 분석에 사용하였다. 이상의 과정에 따라 본 연구를 위해서 총 1,923편의 논문이 수집되었다.

빅데이터 관련 분야를 분석하기 위하여 본 연구에서는 두 가지 방법을 시도하였다. 첫째, 빅데이터를 연구주제로 다루는 학술지에 부여된 컴퓨터공학을 제외한 주제범주를 분석하여 빅데이터 분야가 다루어지는 분야를 파악하고자 하였다. 둘째, 학술지 단위 서지결합분석을 실시하여 네트워크상에서 나타나는 군집과 최근접중심성, 매개중심성, 삼각매개중심성을 통해 그 의미를 해석하고자 하였다. 본 연구에서는 네트워크 분석·시각화 소프트웨어인 NodeXL을 이용하여 코사인 연관성 행렬의 네트워크를 생성하고 각 노드의 매개중심성을 측정하였으며, 이재윤의 WNET 프로그램(v. 0.4.1)을 이용하여 최근접중심성, 삼각매개중심성을 측정하였다.

3. 연구결과

3.1 출판연도별 추이

본 연구에서 수집한 총 1.923편의 논문에 대한 출판연도별 게재 빈도 분포는 <그림 1>과 같다. 처음으로 관련 논문이 출현한 시기는 1999년이며, 그 때부터 현재까지의 빅데이터 분야의 논문을 살펴보면 2012년부터 급격한 증가추세가 나타남을 확인할 수 있다. 2015년에는 990편으로 가장 높게 나타났다. 2016년의 경우 2016년 5월 23일까지 WoS에 포함된 논문만을 집계하였으므로 2016년의 모든 논문을 집계할 2017년에는 더욱 증가한 수치가 나타날 것이라 예상된다. 따라서 빅데이터 분야에 대한 관심이 최근 3-4년간 크게 증가하여 활발한 연구가 수행되고 있음을 알 수 있다.



<그림 1> 출판연도별 논문수

3.2 주제범주 분석

모든 주제범주를 대상으로 산출한 결과 해당 논문들에는 총 127개의 SC가 2,869번 부여되어 논문 당 평균 약 1.5개의 SC범주가 부여되었으며, <표 1>에서 보이듯이 공학이 전체 빈도인 2,869번 중 약 14.57%인 418번 부여되어 가장 많이 부여된 주제범주임을 알 수 있었다. 그 뒤로 정보통신, 경영·경제, 수학, 과학·기술, 환경과학·생태학, 생화학·분자생물학, 문헌정보학, 수리·전산 생물학이 빅데이터 분야의 주요 주제범주로서 나타났다.

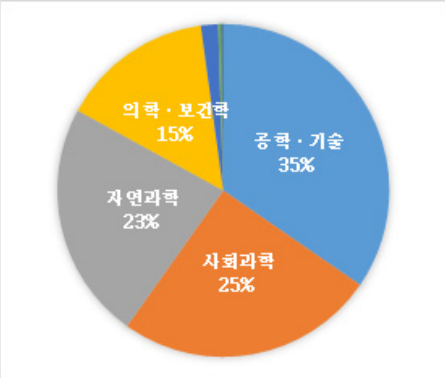
<표 1> 상위 10위 SC 주제범주 및 논문수

순위	SC 주제범주	논문수
1	공학	418
2	정보통신	191
3	경영·경제	154
4	수학	132
5	과학·기술	117
6	환경과학·생태학	85
7	생화학·분자생물학	84
8	문헌정보학	84
9	수리·전산 생물학	81
10	운영연구·경영과학	73

또한 SC를 OECD의 과학기술분야 주제분류(Field of Science and Technology Classification)의 6개 대분류 영역인 농업과학, 공학·기술, 인문학, 의학·보건학, 자연과학, 사회과학으로 분류하였다. 그 결과, <그림 2>와 같이 공학·기술(34.60%)이 가장 높은 비율을 차지하였는데 이는 SC범주 중 가장 높은 비율을 차지하고 있는 공학, 정보통신, 과학·기술 등이 속하기 때문으로 해석할 수 있다. 그 뒤를 이어 사회과학(25.24%), 자연과학(23.14%), 의학·보건학(14.85%) 순으로 많이 나타나고 있다. 인문학(1.69%)과 농업과학(0.21%)을 제외하고 나머지 주제분야가 고르게 분포하고 있는 것을 확인할 수 있었다.

3.3 서지결합 네트워크 분석

본 연구에서는 빅데이터 관련 분야의 구조분석을 위해 논문 게재 빈도 상위 42위(49종)을 대상으로 서지결합분석을 실시하였다. 서지결합이 발생하지 않은 11종을 제외한 38종의 학술지 간 서지결합빈도를 이용하여 코사인 유사도 행렬을 작성하였다. 코사인 유사도 행렬을 패스파인더 네트워크 알고리즘을 적용하여 네트워크를 생성하고, 이재윤의 WNET 프로그램을 사용하여 병렬최근접이웃클러스터링(PNNC) 알고리즘을 적용한 결과 2개의 대



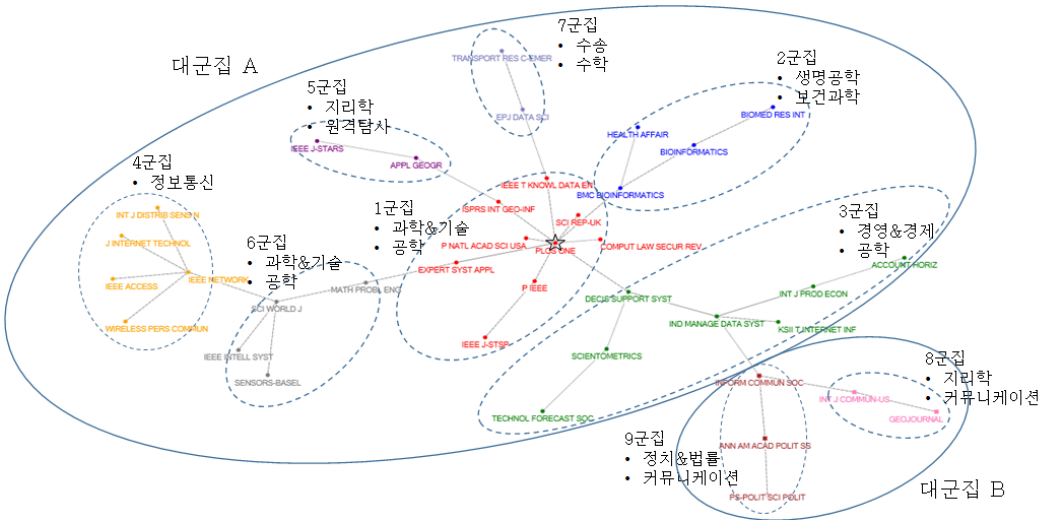
<그림 2> OECD 대분류별 논문 비율

군집과 9개의 소군집이 생성되었다. <그림 3>의 각 노드는 학술지를 나타내고, 대군집은 실선, 소군집은 점선으로 표시하였다. 또한 해당 학술지에 부여된 모든 SC와 최근접중심성을 기반으로 각 군집을 대표하는 주제를 부여하였다.

대군집 A는 제 1군집부터 제 7군집까지 대부분의 군집을 포함하며 공학, 자연과학 분야로 이루어져 있다. 다만 대군집 B와 가장 가까이 위치한 제 3군집은 사회과학 분야에 해당하는 경영·경제도 포함하고 있어 대군집 간을 연결하는 역할을 하고 있다. 대군집 B는 제 8군집과 제 9군집으로 이루어져 있으며 사회과학 분야로 이루어져 있다.

제 1군집은 과학·기술과 공학을 다루는 총 9종의 학술지로 이루어져 가장 규모가 크며, 근접한 5개의 군집과 직접적으로 연결되어 있어 네트워크 내 군집들 간의 매개 역할을 하고 있다. 또한 전체 네트워크에서 매개중심성 및 삼각매개중심성이 가장 큰 학술지 PLOS ONE이 속한 군집으로서, 해당 학술지가 모든 과학 및 의학 분야를 다루는 다학제적 성격을 지녀 전체 네트워크에서 중심적인 역할을 하고 있음을 보여준다.

제 3군집은 학술지 7종이 사회과학과 공학 분야로 이루어져 있으나, 대다수가 경영·경제



<그림 3> 빅데이터 관련 학술지 38종의 패스파인더 네트워크와 PNNC 군집

에 해당하며 공학 분야의 경우에도 산업공학을 다루고 있어 군집이 일관적인 성격을 띠고 있다. 이러한 성격으로 인해 공학 및 자연과학을 다루는 대군집 A와 사회과학을 다루는 대군집 B를 연결하는 역할을 한다.

4. 결 론

본 연구는 빅데이터 관련 분야와 그 지적구조를 파악하고자 하였다. 연구 결과 빅데이터 관련 연구들이 주제분야에 따라 차이가 두드러짐을 확인할 수 있었다. 우선 주제범주를 분석한 결과 OECD 대분류 기준 공학·기술(34.60%), 사회과학(25.24%), 자연과학(23.14%), 의학·보건학(14.85%) 등은 비교적 고르게 분포하고 있는 것으로 나타났지만, 인문학(1.69%)과 농업과학(0.21%) 분야에는 빅데이터와 관련한 연구가 미비한 것으로 나타났다.

네트워크 분석 결과 공학 및 자연과학과 사회과학 분야 간 차이를 확인할 수 있었다. 제 3군집을 사회과학 분야에 포함시켜도, 전체 9개 군집 중 6개 군집(68.42%)이 공학 및 자연과학 분야에 속하고 3개 군집(31.58%)만이 사회과학 분야에 속한다. 또한 공학 및 자연과학 분야에 해당하는 학술지들은 다양한 주제분야를 다루는 반면 사회과학 분야의 학술지들은 커뮤니케이션, 정치·법률, 지리학에 한정되어 있어 아직 사회과학의 다양한 주제분야에서 빅데이터 관련 연구가 활발히 진행되지 못하고 있음을 알 수 있었다.

본 연구는 학술지에 부여된 SC를 중심으로 빅데이터 관련 학술지 및 군집의 주제적 성격을 규정하여 큰 주제범주 간 관계를 파악하였다. 따라서 후속 연구에서는 대상 학술지에 대한 더욱 상세한 주제적 설명을 바탕으로 세부 주제범주 간 관계를 파악하는 것도 의미 있을 것이다.

참고문헌

김완종 (2014). 동시출현 단어분석을 활용한 빅데이터 관련 연구동향 분석. 한국정보관리학회 학술대회 논문집, 17-20.

김현영, 지현수, 이화순, 지종덕 (2014). 빅 데이터에 따른 지적정보의 효율화 방안 연구. 한국지적정보학회지, 16(1), 29-48.

이정미 (2013). 빅데이터의 이해와 도서관 정보서비스에의 활용. 한국비블리아학회지, 24(4), 53-73.