동시출현 단어분석을 활용한 빅데이터 관련 연구동향 분석[1)]

The Research Trends about the Big Data Using Co-word Analysis

김완종, 한국과학기술정보연구원, wjkim@kisti.re.kr
Wanjong Kimg, Korea Institute of Science and Technology Information

본 연구는 동시출현 단어분석 기법을 이용하여 최근 전세계적으로 많은 주목을 받고 있는 빅데이터(Big Data) 관련 연구 동향과 연구 영역을 분석하는 것을 목적으로 한다. 이를 위하여 인용색인데이터베이스인 Web of Science SCIE(Science Citation Index Expanded)에서 분석 대상 논문을 수집하였다. 논문 수집을 위한 검색식은 은 Title(논문 제목), Abstract(초록), Author Keywords(저자 키워드), Keywords Plus®의 네 가지 필드를 동시에 검색하는 주제어(topic)가 “big data”를 포함하고 있는 논문 563편을 대상으로 동시출현단어 분석을 수행하였다.

1. 서 론

최근 들어 IT 분야는 물론이거니와 사회 곳곳에서 ‘빅데이터’에 논의가 활발하게 진행되고 있다. 빅데이터는 일반적으로 기존의 컴퓨터 환경에서는 다룰 수 없는 큰 규모의 데이터를 의미하고 있다. 현재 전 세계적으로 빅데이터와 관련된 웹문서와 보고서들이 계속적으로 출판되고 있다. 이러한 자료들은 빅데이터 분석에 필요한 하드웨어, 소프트웨어, 분석 방법 등 기술적 측면이나 미래에 대한 시장 전망 등에 대한 내용이 주를 이루고 있다. 하지만 실제 빅데이터를 분석한 결과에 대한 사례는 국내를 비롯하여 해외에서도 아직까지 많은 논의가 이루어지지 않고 있다. 국내의 경우 이정미(2013)의 연구가 문헌정보학 분야에서 발표된 유일한 논문이라 할 수 있다. 따라서 본 연구에서는 현재까지 출판된 빅데이터와 관련된 연구 동향을 비롯하여 다양한 연구 주제 간의 관련성과 세부 연구 영역 등을 확인하고자 한다.

본 연구의 분석을 위하여 인용색인 데이터베이스인 Web of Science SCIE를 활용하였으며, 군집분석 및 네트워크 표현을 위한 시각화 도구로는 공개 소프트웨어인 NodeXL을 활용하였다.

본 연구에서의 수행하고자 하는 내용은 다음과 같다.

첫째, 빅데이터 관련 논문에 대한 연도별 발표 추이를 살펴본다. 둘째, 빅데이터 관련 논문이 수록된 학술지 현황을 알아본다. 셋째, 빅 데이터 관련 연구 동향을 분석하기 위하여 저자 키워드와 키워드 플러스[2)]에 출현한 주제어를 중심으로 동시출현 단어분석을 기초로 한 네트워크 분석을 통해 빅데이터 관련 연구 분야를 알아본다.

2. 데이터 수집

2.1 데이터 수집

빅데이터 관련 연구 동향 분석을 위한 데이터

1) 본 연구는 문화체육관광부의 ‘2014년 도서관 빅데이터 분석활용 체계 구축’ 사업으로 수행되었음.
2) Web of Science에서 직접 색인한 키워드

<표 1> 검색조건 및 검색 결과

구분	검색조건
검색조건	Topic = (“big data”) Timespan=All years. Databases=SCI-EXPANDED
검색기간	SCIE : 1986년 ~ 현재
검색결과	563편

를 수집하기 위하여 인용색인데이터베이스인 Web of Science SCIE를 선정하였다. 데이터 수집은 2014년 5월 23일 실시하였으며, 논문명(Title), 초록(Abstract), 저자 키워드, 키워드 플러스 네 개의 필드에 “big data”를 포함하고 있는 논문을 검색하기 위하여 검색 필드를 주제어(Topic)로 선택한 후 검색을 실시하였다. 검색결과 563편의 분석용 데이터를 수집하였다.

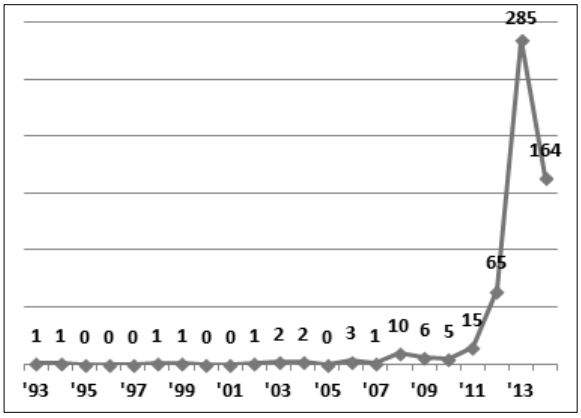
2.2 데이터 전처리

Web of Science 검색을 통해 얻어진 563편의 논문에서 동시단어분석을 위해 Author Keywords와 Keywords Plus® 필드에서 주제어를 추출하였다. 분석 대상 논문들을 대상으로 데이터 분석을 위해 다음과 같이 전처리 작업을 수행하였다. 첫째, 동일한 용어가 대소문자 구분에 따라 다른 용어로 식별되는 것을 방지하기 위하여 각 필드에 포함된 주제어를 대문자로 변환하였다. 둘째, 단복수 명사, 하이픈(-) 등으로 구분된 용어를 동일한 용어로 변환하는 작업을 실시하였다. 셋째, 동일한 논문의 저자 키워드와 키워드 플러스 필드에 동시에 출현한 중복 주제어는 하나의 주제어로 처리하였다.

3. 데이터 분석

3.1 출판연도별 분석

먼저 본 연구에서 수집한 563편에 대한 출판 연도별 분석을 수행하였다. 빅데이터 관련



<그림 1> 빅데이터 관련 연도별 논문수

논문이 최초로 출판된 것은 1993년이다. 이후 2007년까지 15년간 빅데이터 관련 논문이 출판되지 않은 해와 1편의 논문이 출판된 해가 6회, 2편의 논문이 출판된 해가 2회, 3편의 논문이 출판된 해가 1회로 적은 수의 논문이 출판되었다. 하지만 2008년 10편, 2009년 6편, 2010년 5편, 2011년 15편, 2012년 65편으로 논문 발표량이 점차 증가하다 2013년 285편으로 빅데이터 관련 연구가 급격히 증가하고 있다. 또한 2014년 5월 23일 기준으로 2014년에 출판된 논문수가 164편으로 2014년도 논문이 최종 집계되는 2105년 상반기에는 관련 논문이 300편 이상일 것으로 예측할 수 있다(<그림 1> 참조).

3.2 학술지 분석

빅데이터 관련 논문이 수록된 학술지는 총 321종이었으며 학술지 한 종당 평균 1.8편의 빅데이터 관련 논문이 출판되었다. 빅데이터 관련 논문이 10편 이상 수록된 학술지는 Nature를 비롯한 총 4종에 불과했다. *Nature*가 가장 많은 19편을 수록하였으며, *Computer* 17편, *PLOS One* 12편, *Behavioral and Brain Sciences* 11편 등의 순서로 나타났다. 이 4종의 학술지를 보면 *Computer* 한 종만 IT 분야이고 다른 3종의 학술지는 다학제 분야 및 뇌 과학 분야인 것을 알 수 있다(<표 2> 참조).

<표 2> 학술지별 논문수,

학술지명	논문수
Nature	19
Computer	17
PLOS One	12
Behavioral and Brain Sciences	11
기타	504
총합계	563

3.2 동시출현단어 네트워크 분석

빅 데이터 관련 연구 동향을 분석하기 위하여 “2.2 데이터 전처리”에서 언급한 바와 같이 Web of Science SCIE 검색을 통해 얻어진 563편의 논문의 저자 키워드와 키워드 플러스 2개의 필드에 출현한 주제어를 중심으로 동시출현단어분석을 기초로 한 네트워크 분석을 통해 빅데이터와 관련된 연구 분야를 알아보았다.

데이터 전처리 과정을 통해 저자 키워드와 키워드 플러스 필드에서 분석을 위해 추출된 주제어는 총 2,055개였으며, 해당 주제어들의 총 누적 출현빈도는 3,063회로 주제어 하나당 평균 1.5회의 낮은 출현빈도를 보이는 것으로 나타났다.

주제어별 출현 빈도를 보면 Big Data 133회 Systems 29회, Mapreducs 28회, Model 23회의 순으로 나타났다. 주제어별 출현빈도가 10회 이상 되는 주제어는 20개에 불과했으며, 5회 이상 되는 주제어는 62개였다.

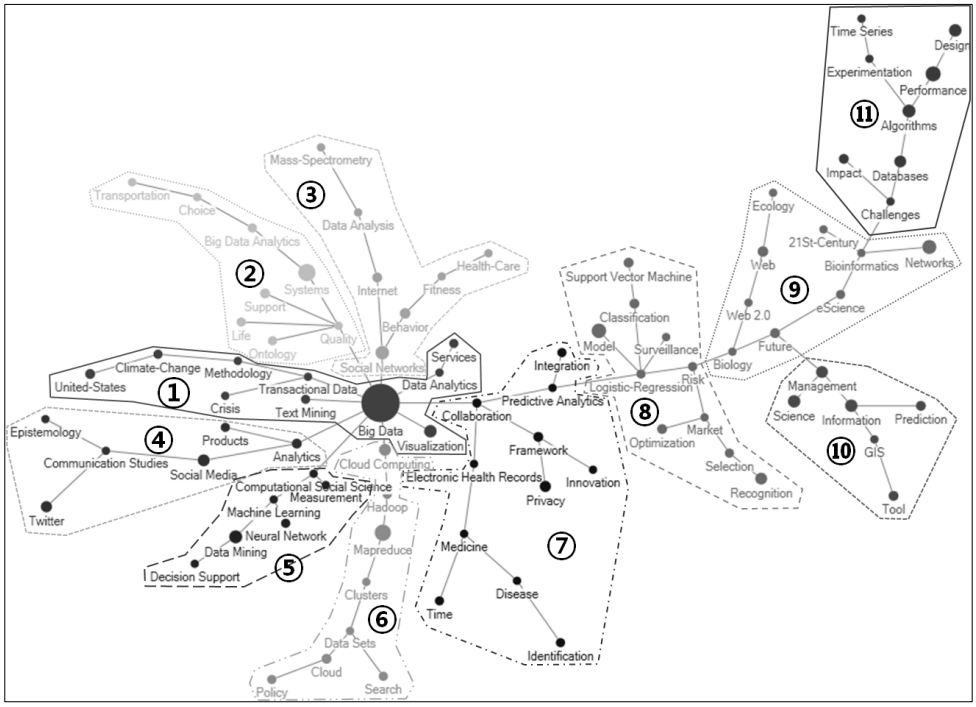
<표 3> 주제어별 출현빈도

주제어	출현빈도
Big Data	133
Systems	29
Mapreduce	28
Model	23
기타	1,842
총합계	2,055

본 연구에서는 빅데이터와 관련된 연구 영역을 파악하기에는 출현빈도가 낮아 정확한 연구영역을 분석하기 어렵다고 판단된 출현빈도 2회 이하인 1,993개 주제어를 제외하고, 3회 이상 출현한 89개 주제어를 대상으로 동시출현단어분석을 기초로 한 네트워크 분석을 실시하였다.

네트워크 분석은 주제어 간의 동시출현빈도를 산출한 후 각 주제어 간의 상관계수를 구하여 이를 패스파인더 네트워크로 표현하였다. 네트워크 표현을 위한 시각화 도구로는 공개 소프트웨어인 NodeXL을 활용하였다.

네트워크 분석 결과 총 11개 군집이 형성되었다. 10개의 주제어가 포함된 제1군집은 Big Data를 포함하고 있는 군집으로서 Text Mining, Visualization 이라는 분석 및 시각화 용어와 함께 미국, 기후변화라는 용어가 함께 나타났다. 8개의 주제어를 가진 제2군집은 시스템, 빅데이터 분석, 품질 등의 용어가 나타났으며, 7개의 용어가 속한 제3군집은 소셜 네트워크와 함께 건강 관련 용어들이 포함되었다. 6개의 주제어를 지닌 제4군집과 5군집은 소셜 미디어, 트위터, 언론정보학 등의 용어와 사회과학, 신경망, 데이터마이닝, 의사결정 지원 등의 용어를 포함하고 있었다. 8개의 용어를 지닌 제6군집은 하둡, 맵리듀스, 클라우드 컴퓨팅과 같은 빅데이터 플랫폼과 관련된 용어들이 속해있다. 네트워크 군집 가운데 가장 많은 11개의 주제어를 가진 제7군집은 프라이버시, 질병, 약물, 식별, 시간 등과 같은 건강 및 의료 분야의 용어가 다수 포함되어 있다. 10개의 주제어를 포함하고 있는 제8군집은 시장, 인지, 감시, 위험 등에 대한 주제어와 최적화, 분류, 지지벡터머신(SVM) 등의 용어가 함께 포함되어 있다. 9개의 주제어가 수록된 제9군집은 21세기, 미래, eScinece, 웹 2.0 등에 대한 시대적 용어와 함께 생태학, 생물학, 생물정보학 등의 용어가 포함되어 있다. 6개의 주제어가 포함된 제10군집은 과학, 경영, 정보, 예측, GIS, 도구 등 다른 군집들에 비해 좀 더 일반적인 용어들이 포함되어 있다. 마지



<그림 2> 빅데이터 분야 11개 주제영역 분석

막으로 8개의 용어가 포함된 제11군집은 데이터베이스, 알고리즘, 실험, 도전, 설계, 영향력 등의 용어가 포함되어 새로운 모형이나 실험 등에 대한 용어가 주를 이루었다.

4. 결론 및 시사점

본 연구는 SCIE에 등재된 321종의 학술지에 수록된 빅데이터 관련 논문 563편의 저자 키워드 및 키워드 플러스에 출현한 주제어를 대상으로 동시출현 단어분석을 실시하였다. 분석 결과 총 11개의 주제 영역으로 구분할 수 있었으며, 하드웨어, 소프트웨어, 데이터 분석 기법 등을 포함한 IT 전반적인 분야뿐만 아니라 사회과학, 경영, 생물학, 의학 등에 관련된 주제어들이 포함되어 있는 것으로 나타났다. 하지만 물리학이나 화학, 재료과학 등과 관련된 주제어는 확인할 수 없었다.

본 연구에서는 각 주제어에 대한 네트워크 지수는 파악하지 않았다. 향후 연구에서는 이에 대한 후속 연구가 수행되어 개별 주제어들이 네트워크 내에서 지니는 영향력이나 위상 등에 대한 분석도 함께 이루어져야 할 것이다.

참고문헌

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