디지털산업정보학회 논문지 제15권 제1호-2019년 3월
https://doi.org/10.17662/ksdim.2019.15.1.001

빅데이터 연구동향 분석: 토픽 모델링을 중심으로*

박 종 순**·김 창 식***

Research Trends Analysis of Big Data: Focused on the Topic Modeling

Park Jongsoon · Kim Changsik

〈Abstract〉

The objective of this study is to examine the trends in big data. Research abstracts were extracted from 4,019 articles, published between 1995 and 2018, on Web of Science and were analyzed using topic modeling and time series analysis. The 20 single-term topics that appeared most frequently were as follows: model, technology, algorithm, problem, performance, network, framework, analytics, management, process, value, user, knowledge, dataset, resource, service, cloud, storage, business, and health. The 20 multi-term topics were as follows: sense technology architecture (T10), decision system (T18), classification algorithm (T03), data analytics (T17), system performance (T09), data science (T06), distribution method (T20), service dataset (T19), network communication (T05), customer & business (T16), cloud computing (T02), health care (T14), smart city (T11), patient & disease (T04), privacy & security (T08), research design (T01), social media (T12), student & education (T13), energy consumption (T07), supply chain management (T15). The time series data indicated that the 40 single-term topics and multi-term topics were hot topics. This study provides suggestions for future research.

Key Words : Big Data, Topic Modeling, Time Series Analysis, Single-Term Topics, Multi-Term Topics

Ⅰ. 서론

최근 빅데이터 중 텍스트 데이터를 통해서 시사점을 도출하기 위한 연구가 다양한 관점으로 수행되어지고 있다. 특히 국내외 논문의 초록을 위주로, 다양한 분야에서 연구동향을 파악하기 위한 시도가 다수 이루어져 왔다[1-5]. 이러한 현상은 텍스트 데이터를 분석하는 정보통신기술의 발달에 기인하였다. 정보통신기술의 발전은 텍스트 데이터의 빈도분석, 군집화, 분류, 시계열분석, 네트워크분석을 가능하도록 만들었다[6]. 최근에는 빅데이터를 기반으로 인공지능을 활용하는 서비스들에도 관심이 집중되고 있다.

국내외 논문의 초록 데이터를 분석 하면, 연구의 주요 토픽 추출이 가능하고, 연도별 토픽의 변화를

* 본 논문은 2019학년도 서일대학교 학술연구비에 의해 연구되었음.
** 서일대학교 소프트웨어공학과 교수
*** 세종대 박사과정 · 배화여자대학교 글로벌관광과 조교수 (교신저자)

파악할 수 있다. 또한 보다 많은 양질의 데이터가 축적된다면 추가적인 시사점 도출도 가능할 것이다.

일반적으로 국내외 논문을 활용한 텍스트 분석은 초록 또는 키워드를 대상으로 한다. 초기 텍스트 분석은 키워드를 대상으로 빈도분석과 동시출현빈도를 분석하는 형태로 시작하였다. 그러나 최근에는 토픽 모델링 기법을 적용하여, 단일 토픽 또는 다중 토픽을 추출하며, 추출된 토픽들을 대상으로 하여 시계열 분석을 수행하기도 한다. 또한 네트워크분석기법을 적용하여 토픽과 토픽간의 관계를 파악하기 위한 분석도 시도되고 있다. 초기 단일 텍스트 분석기법을 적용한 연구는 다양한 한계점을 가지고 있었다. 그러나 토픽모델링 기법을 활용한 접근은 이러한 한계를 극복하고 있다. 또한 토픽모델링 분석 결과는 시계열 분석의 입력으로 투입되어 각 토픽들의 추세 파악도 가능하게 된다.

본 연구에서는 빅데이터 분야의 연구동향 분석을 위해 텍스트마이닝 기법 중 최근 활발하게 적용되는 토픽 모델링과 시계열분석 기법을 활용하였다. 분석 대상의 선정은 웹오브사이언스 (Web of Science) 데이터베이스에서 'Big Data' 을 검색하여 1995년부터 2018년까지 게재된 논문들을 대상으로 하였다. 즉 본 연구와 기존 선행연구들과의 차이점은 토픽모델링과 시계열분석 기법을 활용하여 'Big Data' 연구 동향을 분석하였다는 점이다.

본 연구의 2장에서는 토픽모델링 관련연구에 대해서 문헌연구 결과를 제시하였으며, 3장에서는 텍스트마이닝에서 중요한 기법 중의 하나인 토픽 모델링 및 시계열 분석을 이용해서 빅데이터 연구 동향을 분석하는 절차와 결과를 제시하였다. 끝으로 4장에서는 결론으로 구성하였다.

Ⅱ. 관련연구

토픽 모델링은 비정형 데이터인 텍스트에서 중요한 토픽을 추출하고, 단일토픽과 단일토픽사이 또는 다중토픽과 다중토픽사이에 어떤 관련성이 있는지를 확인할 수 있다[7].

토픽 모델링은 많은 수의 문서들을 그 토픽에 따라 군집한다는 관점에서 문서의 군집화 기법과 유사하다고 볼 수 있지만, 다중 토픽에 하나의 문서가 동시에 대응될 수 있기 때문에 현 세계의 모델링에 부합하다고 할 수 있다. 토픽 모델링은 많은 양의 문서들을 분석하여 관련분야에 통찰을 제공한다는 관점 이외에도, 분석 결과를 활용하여 추가적인 분석을 할 수 있다는 관점에서 그 가치가 높다고 할 수 있다[6].

토픽 모델링을 활용해 텍스트 데이터의 추세분석에 활용되고 있는데, 특정 분야에서 중요하게 다루어지는 토픽의 도출 및 그 토픽을 추적하기 위한 연구에 이용되고 있다. 그 사례로는 국제 관광분야의 연구추세를 확인하기 위해 토픽 모델링을 활용한 연구가 있다. 이 연구에서는 토픽 모델링과 시계열분석을 주로 하였으며, 관광 분야에서 국제논문의 연구동향을 파악하였다[1]. 또한 토픽 모델링과 시계열회귀분석을 활용하여, 국내의 정보시스템분야 대표저널을 대상으로 중요한 토픽 도출과 연구동향 파악을 하였고[2], 더불어 정보보호분야 연구동향을 분석하기 위하여 텍스트마이닝을 활용한 연구도 수행하였다. 이 연구에서는 정보보호 분야의 주요 학술지들의 1991년도부터 2016년도까지 발표된 논문의 초록 1,096편을 대상으로 분석을 수행하였다[3]. 특히 텍스트마이닝과 소셜네트워크 분석기법을 활용하여 호텔분야의 연구 동향을 분석한 사례도 있다. 이 연구는 관광분야 국제선도 저널 4곳에서 1994년부터 2016년까지 발행된 706편을 대상으로 하였다[4]. 이외에도 토픽 모델링 기법은 국내문헌정보학 연구동향 분석[8], 트위터 이슈 트래킹[9], 행복과 불행 이슈[10] 등에 다

양하게 적용되었다. 최근에는 빅데이터의 생성에 큰 역할을 담당하고 있는 소셜네트워크 서비스에 대한 연구동향에도 적용되었다. 이 조사는 1994년도부터 2016년까지 출판된 논문 초록 308편을 대상으로 하여, 소셜네트워크 서비스에 대한 중요한 토픽과 연구동향을 파악되었다[5].

Ⅲ. 토픽 모델링 및 시계열 분석

3.1 분석 대상

빅데이터 분야 연구동향에 대한 분석을 위하여, 웹오브사이언스(Web of Science) 데이터베이스에서 2018년 9월 20일 기준 'Big Data' 를 검색하여 1995년부터 2018년까지 4,245편 논문을 분석 대상으로 선정하였다. 본 연구에서는 데이터 전 처리 후 4,019편의 초록을 대상으로 기간별 연구동향을 파악하였다.

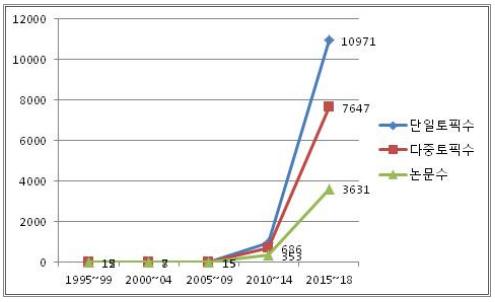
다음의 <표 1>은 5년 단위로 분석한 기간별 단일 토픽 수, 다중토픽 수, 논문 수를 나타낸 것이다.

<표 1> 기간별 논문 수

기간	단일 토픽 수	다중 토픽 수	논문 수
1995~1999	24	15	12
2000~2004	5	7	8
2005~2009	16	15	15
2010~2014	964	686	353
2015~2018	10,971	7,647	3,631
합계	11,980	8,370	4,019

빅데이터 관련 논문 수와, 단일 토픽 수, 다중 토픽 수는 1995년 이후 시간의 흐름에 따라 그 수가 감소하다가, 2005년 이후 급격히 증가하는 추세를 보이고 있다. 특히 2015년 이후는 급격하게 증가하는 추세에 있다. 이러한 추세는 2018년 이후에도 지속될 것으로 예상된다.

다음의 <그림 1>은 분석 대상 논문들의 기간별 단일 토픽 수, 다중 토픽수, 논문 수를 그래프로 표현한 것이다.



<그림 1> 기간별 단일 토픽 수, 다중 토픽 수, 논문 수

3.2 분석 방법 및 절차

분석은 SAS Enterprise Miner(SAS EM), SAS Enterprise Guide 7.2, Excel, SPSS를 활용하였다. 분석의 절차는 첫째, Excel 및 Enterprise Guide 7.2를 활용하여 데이터 전처리, 둘째, SAS를 활용하여 토픽모델링, 셋째, SPSS를 활용하여 시계열분석으로 수행하였다[1-5].

- 전처리: 분석 과정에서 가장 많은 시간이 소요되는 전처리 작업은 주로 Excel을 활용하였다.
- 토픽모델링: SAS EM을 활용하여 토픽모델링을 수행하였고, ① 데이터 업로드 ② 데이터 파싱 ③ 데이터 필터 ④ 토픽모델링의 순서로 이루어졌다.
- 시계열분석: SAS EM을 활용한 토픽모델링 결과를 기반으로 SPSS를 활용하여 시계열분석을 수행하였다.

3.3 분석 결과

토픽모델링의 빈도는 논문에서 토픽들이 얼마나

많이 다루어지는지를 나타낸다. 토픽은 단일 토픽, 다중 토픽으로 분석가능하다.

다음 <표 2>는 1995년에서 2018년까지의 빅데이터에 대한 단일 토픽 모델링 결과이다.

<표 2> 단일 토픽 모델링 결과

토픽id	토픽	빈도
T09	model	955
T13	technology	914
T04	algorithm	832
T18	problem	801
T11	performance	779
T01	network	697
T17	framework	660
T05	analytics	624
T08	management	620
T12	process	594
T16	value	533
T02	user	523
T06	knowledge	516
T19	dataset	497
T07	resource	478
T03	service	467
T15	cloud	384
T14	storage	379
T20	business	379
T10	health	348

<표 3> 다중 토픽 모델링 결과

토픽id	토픽	빈도
T10	sensor, storage, volume, technology, architecture	596
T18	system, decision, process, knowledge, model	566
T03	algorithm, classification, method, feature, learning	560
T17	analytics, data analytics, learning, business, technique	510
T09	query, performance, memory, cluster, system	508
T06	science, research, big data, data science, scientist	506
T20	model, sample, variable, method, distribution	502
T19	user, service, dataset, web, search	487
T05	network, traffic, node, communication, transmission	482
T16	customer, business, product, market, company	446

이 기간 논문에서는 다중 토픽 모델링 결과, 기술 아키텍처, 의사결정시스템, 학습 및 분류 알고리즘, 데이터 애널리틱스, 질의 성과, 데이터사이언스, 샘플 모델, 사용자 서비스 데이터 셋, 네트워크 트래픽, 고객 및 제품, 클라우드 컴퓨팅, 건강 관리, 스마트 시티, 개인 정보 보안, 디자인 리서치, 소셜미디어, 학습과 기술, 에너지 소비와 센스, 공급체인관리 순으로 중요하게 도출되었다.

다음 <표 4>, <표 5> 및 <그림 2>, <그림 3>은 단일 및 다중 토픽별 분석결과를 나타낸 것이다.

<표 4> 단일 토픽 시계열 분석

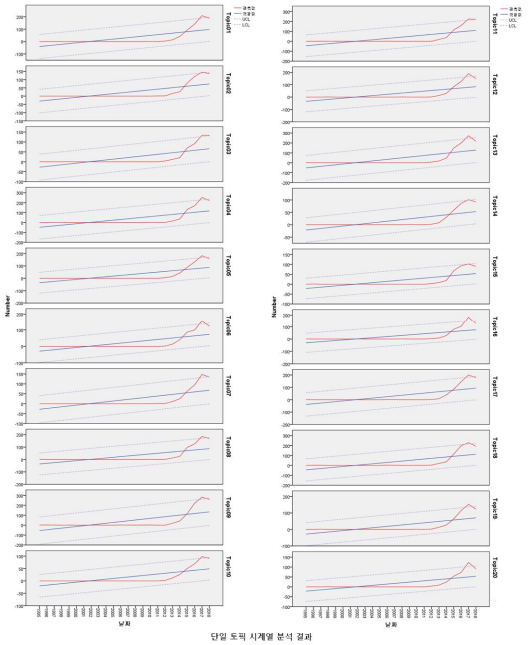
토픽id	추정 값	p-value	Hot/Cold
T01	6.008	.000	Hot
T02	4.520	.000	Hot
T03	4.050	.000	Hot
T04	7.196	.000	Hot
T05	5.379	.000	Hot
T06	4.428	.000	Hot
T07	4.176	.000	Hot
T08	5.383	.000	Hot
T09	8.218	.000	Hot
T10	2.991	.000	Hot
T11	6.761	.000	Hot
T12	5.145	.000	Hot
T13	7.830	.000	Hot
T14	3.241	.000	Hot
T15	3.277	.000	Hot
T16	4.596	.000	Hot
T17	5.731	.000	Hot
T18	6.862	.000	Hot
T19	4.287	.000	Hot
T20	3.254	.000	Hot

<표 5> 다중 토픽 시계열 분석

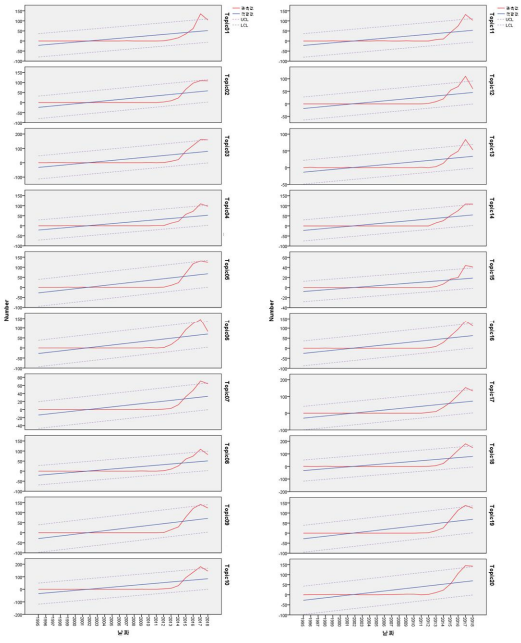
토픽id	추정 값	p-value	Hot/Cold
T01	3.193	.001	Hot
T02	3.580	.000	Hot
T03	4.870	.000	Hot
T04	3.200	.000	Hot
T05	4.143	.000	Hot
T06	4.231	.000	Hot
T07	2.023	.000	Hot
T08	3.113	.000	Hot
T09	4.372	.000	Hot
T10	5.160	.000	Hot
T11	3.271	.001	Hot
T12	2.772	.000	Hot
T13	2.059	.000	Hot
T14	3.359	.000	Hot
T15	1.160	.001	Hot
T16	3.847	.000	Hot
T17	4.416	.000	Hot
T18	4.915	.000	Hot
T19	4.210	.000	Hot
T20	4.214	.000	Hot

시간의 경과에 따른 빅데이터 분야의 핵심 토픽들의 변화를 분석하기 위하여, 토픽모델 결과를 기반으로 SPSS를 활용하여 시계열분석을 수행하였다. 이러한 시계열분석의 결과를 통해 각 토픽들의 지난 24년간 연도별 추세를 통계적으로 확인할 수 있다. 시계열분석 결과 회귀계수가 양수이고, 유의확률이 유의미하면 핫 토픽(Hot Topic), 회귀계수 값이 음수이고, 유의확률이 유의미하면, 콜드 토픽(Cold Topic), 회귀계수 결과 값이 유의미하지 않은 경우 중립토픽(Neutral Topic)으로 구분한다[1-5].

시계열분석 결과 단일 토픽과 다중 토픽 모두 상승하는 형태인 핫 토픽으로 나타났다.



<그림 2> 단일 토픽 트렌드 분석 결과



<그림 3> 다중 토픽 트렌드 분석 결과

Ⅳ. 결론

정보통신기술의 발전에 따라 여러 비즈니스 영역에서 다양한 서비스들이 도입되고 있으며, 특히 텍스트와 같은 비정형 빅데이터를 활용한 의미를 찾고자 하는 노력이 활발히 이루어지고 있다. 텍스트마이닝에서 중요한 기법 중의 하나인 토픽모델링을 활용하여 텍스트에서 시사점을 도출하기 위한 연구가 이루어지고 있다. 그러나 여전히 토픽모델링을 이용한 분석은 다양한 관점으로 진행되어야 한다. 특히 빅데이터를 활용하는 측면의 연구는 더욱 더 필요하다. 그러므로 본 논문에서는 빅데이터 연구의 주요 토픽별 연구동향을 분석 하였다.

본 연구에서는 1995년부터 2018년까지 게재된 논문의 초록들을 대상으로 토픽모델링과 시계열분석을 통해 기간별 연구동향을 분석하였으며, 도출된 주제들은 단일 토픽결과, 모델, 기술, 알고리즘 등이 있으며, 다중 토픽결과 기술아키텍처, 의사결정시스템, 학습 및 분류 알고리즘 순으로 중요하게 다루어지고 있음을 확인되었다. 한편 단일 토픽 및 다중 토픽을 1995년 이후 연도별로 분석한 결과 , 2015년 이후 급격한 증가세를 나타내고 있다.

본 연구의 이론적 및 실무적 시사점은 빅데이터 연구동향을 텍스트 마이닝 및 시계열분석 기법을 통해 확인하였다는 점이다. 연구 결과를 통해 학자들은 연구주제 선택 관점에서, 빅데이터 분야 실무자들은 향후 기술의 변화 관점에서 통찰력을 얻을 수 있을 것이다.

이와 같은 시사점에도 불구하고, 본 연구는 다음과 같은 한계점이 있다. 본 연구에서는 연구 설계 단계에서 다양한 분류체계를 고려하지 못했다. 추후 연구 설계에서 인문사회 분야인지 공학 분야인지를 구분하거나, 또한 경영학, 언론학, 정치학 등의 관점으로 분류가 가능토록 반영 한다면 보다 더 의미 있는 시사점이 도출될 수 있을 것이다.

참고문헌

[1] 김창식, 곽기영, 윤혜진, "관광분야 연구동향 분석: 토픽모델링과 시계열분석을 중심으로," 관광레저연구, 29(12), 2017, pp. 25-39.

[2] 김창식, 최수정, 곽기영, "토픽모델링과 시계열회귀분석을 활용한 정보시스템분야 연구동향 분석," 디지털콘텐츠학회논문지, 18(6), 2017, pp. 1143-1150.

[3] 김태경, 김창식, "텍스트마이닝을 이용한 정보보호 연구동향 분석," (사)디지털산업정보학회 논문지, 14(2), 2018, pp. 19-25.

[4] 박준석, 김창식, 곽기영, "텍스트마이닝과 소셜네트워크분석 기법을 활용한 호텔분야 연구동향 분석," 관광레저연구, 28(9), 2016, pp. 209-226.

[5] 윤혜진, 김창식, 곽기영, "Research Trends Investigation Using Text Mining Techniques: Focusing on Social Network Services," 디지털콘텐츠학회논문지, 19(3), 2018, pp. 513-519.

[6] 김남규, 이동훈, 최호창, William Xiu Shun Wong, "텍스트 분석 기술 및 활용 동향," 한국통신학회논문지, 42(2), 2017, pp. 471-492.

[7] Blei, D., Ng, A. and Jordan, M., "Latent Dirichlet Allocation," Journal of Machine Learning Research, 3, 2003, pp. 993-1022.

[8] 박자현, 송민, "토픽모델링을 활용한 국내 문헌정보학 연구동향 분석," 정보관리학회지, 30(1), 2013, pp. 7-32.

[9] 배정환, 한남기, 송민, "토픽 모델링을 이용한 트위터 이슈 트래킹 시스템," 지능정보연구, 20(2), 2014, pp. 109-122.

[10] 양승준, 이보연, 김희웅, "토픽모델링 기반 행복과 불행 이슈 분석 및 행복 증진 방안 연구," 지식경영연구, 17(2), 2016, pp. 165-185.

■ 저자소개 ■

박 종 순
(Park Jongsoon)

1993년 3월~현재
서일대학교 소프트웨어공학과
교수

2005년 2월 한국외국어대학교 경영학박사
1990년 2월 한국외국어대학교 경영학석사
1985년 2월 성균관대학교 행정학사
관심분야 : e-business, 기술경영,
시스템분석설계
E-mail : jspark@seoil.ac.kr

김 창 식
(Kim Changsik)

2018년 3월~현재
배화여자대학교 글로벌관광과
조교수
2015년 3월~ 2018년 2월
국민대 비즈니스IT전문대학원
BK21 플러스 사업팀 계약교수
2013년 8월 국민대 비즈니스IT전문대학원
비즈니스IT전공(경영정보학박사)
2002년 2월 경희대학교 산업정보대학원
경영정보학과(경영학석사)

관심분야 : 관광정보, 텍스트마이닝,
지식경영, 데이터 애널리틱스,
기술경영, 소셜네트워크 분석 및
응용
E-mail : solo21solo@naver.com

논문접수일 : 2018년 11월 30일
수 정 일 : 2018년 12월 17일
게재확정일 : 2018년 12월 26일