

디지털 퍼블리싱에 대한 포털사이트 사용기록을 통한 동향 연구

빅데이터의 LDA 토픽 모델링 기법을 중심으로

Trend Research Based on Usage of Portals on Digital Publishing
Centered on LDA topic modeling technique of Big Data

주 저 자 : 이하나 (Lee, Ha Na)

서일대학교 커뮤니케이션디자인학과
swak21@seoil.ac.kr

<https://doi.org/10.46248/kids.2020.4.191>

접수일자 2020. 11. 26. / 심사완료일자 2020. 12. 17. / 게재확정일자 2020. 12. 22.

Abstract

Currently, the launch of services bundles such as Apple News+ and Apple One demonstrates that shift of the global publishing industry from the concept of ownership to a way of access. In order to revitalize Korea's stagnant publishing industry, we need to redefine digital publishing, which is in transitional stage, by reviewing and analyzing the latest trends in the industry. Thus, we conducted a big data analysis solely on the trends of digital publishing. The purpose of this study was to identify directions for the future and areas for improvement in the Korean digital publishing market through the big data analysis. The research methodology was based on unstructured data analysis such as TF-IDF and LDA topic modeling. According to the analysis, the four main topics of digital publishing were platform, digital publishing and industry, pattern of service usage, and content access depending on the media landscape. We hope that the findings of this study will contribute to the sustainability of the entire publishing industry by presenting future directions of the digital publishing market which has been under rapid developments with the launch of a variety of content and streaming subscription services.

Keyword

Digital Publishing(디지털 퍼블리싱), Big-Data(빅 데이터), Topic Modeling(토픽 모델링)

요약

현시점에서 Apple News+와 Apple One과 같은 번들 서비스의 시작은 소유의 개념에서 접속의 방식으로 변화하고 있는 글로벌 출판 산업의 동향을 보여준다. 한국의 침체된 출판 산업의 활성화를 위해서 디지털 퍼블리싱의 최근 동향을 고찰하고 분석하여 과도기적 상황에 처한 디지털 퍼블리싱에 대한 새로운 정의가 요구된다. 이에 본 연구는 디지털 퍼블리싱에 대한 트렌드 및 동향에 한정해서 빅 데이터 분석을 실시하였다. 빅 데이터 분석을 통해 디지털 퍼블리싱의 현 상황을 진단함으로써 국내 디지털 퍼블리싱의 발전을 위해 지향해야 할 부분과 개선되어야 할 부분들이 무엇인지를 파악하고자 하였다. 빅 데이터 분석 방법은 TF-IDF 빈도 분석과 LDA 토픽모델링 분석을 중심으로 비정형 데이터 분석을 수행하였다. 분석 결과 현재의 디지털 퍼블리싱은 크게 플랫폼, 디지털 퍼블리싱과 산업, 서비스 이용 패턴, 미디어 환경에 따른 콘텐츠 접근의 4개 주제로 설명되었다. 본 연구의 결과로써 다양한 콘텐츠와 스트리밍 구독 서비스의 등장으로 급변하는 디지털 퍼블리싱 산업의 방향성을 제시함으로써 출판 산업 전체의 지속 가능성에 일조하기를 기대한다.

목차

1. 서론

- 1-1. 연구 배경 및 목적
- 1-2. 연구 범위 및 방법

2. 이론적 배경

- 2-1. 디지털 퍼블리싱에 대한 이해
- 2-2. 빅 데이터와 LDA 토픽 모델링

3. 실증연구

- 3-1. 데이터 수집 및 정제
- 3-2. TF-IDF분석
- 3-3. 디지털 퍼블리싱 트렌드 분석
- 3-4. LDA 토픽 모델링 분석

4. 결론

참고문헌

1. 서론

1-1. 연구의 배경 및 목적

초기의 디지털 퍼블리싱은 전자출판이라는 용어로 주로 통용되었으며 인쇄물을 위한 컴퓨터 편집 프로그램과 전자 시스템을 갖춘 인쇄 과정을 지칭하는 용어로 사용되었다. 그러나 스마트 디바이스 환경과 IT 기술의 발전으로 인해 ePub출판 표준을 시작으로 디지털 퍼블리싱의 의미는 다변화되었다. 또한 Apple의 아이패드 출시를 시작으로 스마트폰, 스마트패드와 같은 LCD스크린을 통해 다양한 콘텐츠가 등장하면서 디지털 퍼블리싱의 의미는 다양한 매체를 아우르는 개념으로 진화하여 독자에게 전달되기까지의 디지털로 전환되거나 생산되는 일련의 전 과정을 포함한다. 그러나 출판 업계에서는 디지털 퍼블리싱에 대한 동향 분석과 대응이 늦어 디지털 퍼블리싱을 단순히 종이책을 PC나 모바일을 통해서 볼 수 있는 형태로 인식하고 있다.

2019년 출판 산업에 대한 인식조사에서 전자책을 발행 안 하고 있다는 답변이 전체 출판사 중 67.2%로 높은 비중을 보였다. 매출액이 클수록 전자책 발행을 한다는 비중이 높은 것으로 집계되었다.¹⁾ 이와 같이 현재 출판 산업은 시대의 변화에 민첩하게 대응하지 못하고 있으며 이러한 이유로 우리나라의 출판 산업 매출액은 감소하고 있다.

전자책의 유통사에서 채택 중인 현재의 판매 및 마케팅 방식(중복응답)은 “일반 판매 모델(단권 다운로드)”(83.3%), “기간제 대여 모델(최대 90일 이내)”(41.7%), “정액제 구독 모델(무제한 이용)”(33.3% 순²⁾으로 나타났다.

출판산업은 Apple News+와 Apple One과 같은 번들 서비스의 시작으로 소유의 개념에서 접속의 방식으로 변화하고 있는 글로벌 콘텐츠 시장의 변화를 대비해야 한다. 이러한 이유로 디지털 퍼블리싱 산업의 최근 동향을 고찰 및 분석하고 과도기적 상황에 처한 디지털 퍼블리싱에 대한 새로운 정의를 통해 출판 산업이 나아갈 방향을 모색할 필요성이 제기된다. 이에 본 연구는 디지털 퍼블리싱에 대한 트렌드 및 동향에 한정해서 빅 데이터 분석하였다. 본 연구의 결과로써 다양한 콘텐츠와 스트리밍 구독 서비스의 등장으로 급변하는 디지털 퍼블리싱 산업의 방향성을 제시하는 가이드라인으로 활용되기를 기대한다.

1) 김수영, 2019 출판산업 실태조사, 한국출판문화산업진흥원, 2020, p135

2) Ibid, p475

1-2. 연구의 범위 및 방법

디지털 퍼블리싱의 동향과 인식에 대한 분석을 위해 네이버 뉴스 데이터를 수집하였다. 빅 데이터의 표본은 2016년 1월부터 2020년 9월까지 약 4년 9개월간의 데이터이며, 빅 데이터 분석에 사용된 기사는 총 22,932건이다. 연구를 위해 1989년 귀도 반 로섬(Guido van Rossum)에 의해 개발된 프로그래밍 언어인 파이썬(Python)을 사용하였다. 파이썬은 빅 데이터 분야와 인공지능 분야에서 널리 쓰이고 있는 프로그래밍 언어이다. 플랫폼에 독립적이며 객체지향적이라는 특징을 가지고 있다. 파이썬으로 크롤링(Crawling)을 위한 프로그래밍을 제작하고 네이버 API를 연동시켜 네이버 뉴스 기사를 취합하였다.

수집된 빅 데이터의 분석은 Latent Dirichlet allocation 토픽 모델링(이하 LDA 토픽 모델링)에 앞서 INPUT 데이터를 만들기 위해 텍스트 마이닝(Text Mining)과정을 거쳤다. 또한 본문에서의 단어 중요도를 파악 할 수 Term Frequency - Inverse Document Frequency(이하 TF-IDF) 분석을 먼저 실시하였다. TF-IDF를 통해 추출된 단어들을 LDA 토픽모델링의 INPUT 데이터로 사용했으며, 토픽 모델링 분석을 위해 머신러닝 라이브러리인 Scikit-Learn과 Gensim, Word2Vec, TSNE 등의 라이브러리로 함께 사용하였다. 훈련을 통해 추출된 데이터는 Matplotlib과 pyLDAvis 라이브러리를 통해 시각화하고 각 토픽이 가진 의미를 파악하는데 활용하였다.

2. 이론적 배경

2-1. 디지털 퍼블리싱에 대한 이해

위키피디아(Wikipedia)에 따르면, 디지털 퍼블리싱은 아-퍼블리싱(e-publishing), 전자 출판(Electronic Publishing), 온라인 퍼블리싱(Online Publishing)이라고도 한다. 초기의 디지털 퍼블리싱은 전자출판이라는 용어로 주로 통용되었으며 인디디자인과 같은 컴퓨터 편집 프로그램과 전자 시스템을 활용한 인쇄 과정을 지칭하는 용어로 사용되었다.

e-퍼블리싱(e-Publisig), 전자출판(Electronic Publisig)으로도 표기되는 디지털 퍼블리싱이란 지식 정보산업의 콘텐츠가 디지털로 전환되거나 생산되어 소비자, 독자에게 제공되는 일련의 과정을 의미한다.³⁾

3) 김상욱, 권기정, 스마트 디바이스 환경의 디지털

디지털 퍼블리싱 1.0 시대가 디지털 콘텐츠를 구현하는 '기술'을 중심으로 확산된 것과 비교하여 디지털 퍼블리싱 2.0 시대는 기술을 배경으로 한 콘텐츠 커뮤니케이션의 '혁신성'을 그 축으로 하고 있다. 이러한 '혁신성'의 강조가 대두한 것은 미디어 융합기능을 제공하는 스마트 디바이스의 확산과 디지털 콘텐츠를 소비하는 일반 대중의 커뮤니케이션 트렌드에 기인한다.⁴⁾ 다변화되고 확장된 디지털 퍼블리싱에 관한 인식과 콘텐츠 형태에 관한 선행연구를 살펴봄으로써 현재 디지털 퍼블리싱의 제작형태와 구독 형태 등에 대해 고찰하고자 한다. 디지털 퍼블리싱에 대한 선행연구들을 살펴보면 다음과 같다.

김상욱, 권기정(2013)은 스마트 디바이스의 등장으로 읽기 문화의 변화에 맞춰 디지털 퍼블리싱의 새로운 디자인 프로세스를 방법론을 연구하였다. 연구 결과 기술을 배경으로 한 콘텐츠 커뮤니케이션의 혁신성과 디자인 융합이라는 방향성을 제시하였다.

박혜진(2017)은 미디어 환경의 변화에 따라 출판 미디어 영역에서의 스마트 미디어에 적합한 디지털 콘텐츠에 대한 연구를 진행하였다. 디지털 퍼블리싱 애플리케이션 제작을 위한 효과적인 디자인 체계를 위한 결론을 도출하였다.

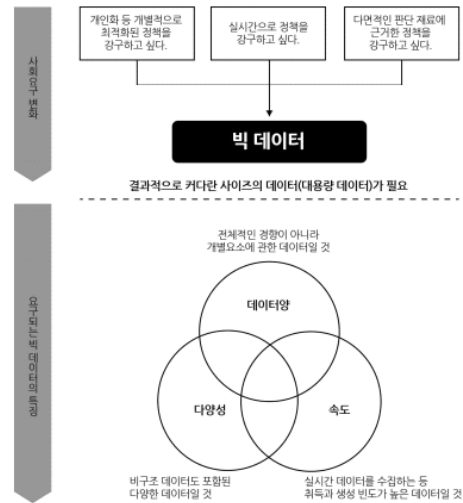
공병훈, 조정미(2018)는 2000~2018년까지 디지털 퍼블리싱 연구 흐름을 통합적으로 분석하였으며 분석 결과, 주제 유형에서는 산업·수용자·미디어에, 대상 미디어에서는 전자책·웹툰에 집중되었다는 결론을 도출하였다.

2-2. 빅 데이터와 LDA 토픽 모델링

오늘날 디지털 기술과 기기의 발전은 수많은 데이터를 쉽고 빠르게 생산하고 공유할 수 있는 환경으로 변화시켰고, 이는 곧 빅 데이터의 시대를 만들었다. 빅 데이터는 데이터의 크기, 속도, 다양성을 특징으로 거대하고 다양한 종류와 형태를 갖춘 데이터들을 빠르게 생성한다. 또한 전파하는 속도를 비롯해 수집 분석 처리에 걸리는 시간 또한 매우 빠른 속성을 가지고 있다. 빅 데이터를 사용한 분석연구는 수집한 데이터를 정제하고 추출하는 과정을 거쳐 도출된 키워드와 그에 해당하는 수치적 결과값을 분석함으로써 특정 단어 혹은 현상에 대한 인식과 예측이 가능하다.⁵⁾

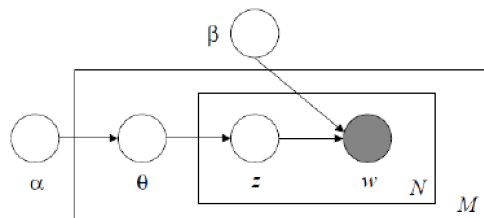
퍼블리싱 커뮤니케이션 연구, 한국디자인트렌드학회, 2013, p369

4) Ibid, p475



[그림 1] 사회 요구 변화에 대처하는 빅 데이터의 세 가지 특징⁶⁾

토픽 모델링은 문서를 각 토픽들의 혼합으로 정의하며, 각 토픽들은 단어의 분포로 나타나고 이를 통해 문서에 사용된 단어들을 바탕으로 문서를 분류하는 방법론이다. 토픽 모델링에서 주로 사용되는 방법론은 LDA로 입력받은 문서들의 단어 분포와 빈도를 통해 파라미터(α , β)를 추정하고, 이 파라미터들을 통해 각 문서의 토픽 비율(θ)과 각 단어의 할당 토픽(z)을 통해 각 주제별로 단어(w)가 결정된다.(그림 1 참조)⁷⁾



[그림 2] Latent Dirichlet Allocation(LDA)⁸⁾

- 5) 김소라, 빅데이터 분석을 이용한 아노락 패션의 인식에 관한 융합 연구. 한국과학예술융합학회, 38(4), 2020, p47
- 6) 이재원, 원하연, 최경란, 정도성, SNS 빅데이터의 집단감성정보 제공이 디자인 집단의 창의성에 미치는 영향 연구. 디지털디자인학연구, 15(3), 2015, p923
- 7) 민혜중, 김영정, 토픽모델링을 이용한 매체 별 빅데이터 동향 다각적 분석. 대한산업공학회 추계학술대회 논문집, 2015, p2448
- 8) Ibid, p2448

LDA 분석은 데이터를 읽어 들일수록 분포에 반영을 하면서 분류를 해 준다. 따라서 데이터 분석의 의도와 관련 없는 토픽들로 구성될 수도 있다. 따라서 모델을 설정할 때 다양한 하이퍼파라미터 값을 조정해 가면서 의도에 맞는 토픽 모델링을 수행해야 한다.

3. 실증연구

3-1. 데이터 수집 및 정제

2016년부터 2020년까지 ‘디지털 퍼블리싱’에 대한 동향과 인식을 분석하고자 네이버 개발자 센터에서 빅 데이터를 이용할 수 있는 API를 제공받고, 이를 파이썬 프로그래밍 언어를 활용하여 크롤러를 제작하였다. 크롤링 방식은 파이썬과 함께 사용할 수 있는 셀레니움(Selenium), 뷰티풀썬(Beautifulsoup) 라이브러리를 사용하였다. 셀레니움 라이브러리를 통해 네이버 뉴스, 블로그, 카페의 HTML 문서를 화신 받아 데이터 분석에 필요한 부분을 태그 기반으로 Parsing해 최종 데이터를 추출하였다.

네이버 API로 수집할 수 있는 데이터의 양은 1회에 1,000건으로 제약이 걸려있기 때문에 30회 정도에 걸쳐 데이터 수집과정을 거쳤다.

디지털 퍼블리싱에 대한 빅 데이터 분석을 위해 네이버의 뉴스의 비정형 데이터를 수집 대상으로 하였으며 데이터의 수집 기간은 2016년 1월부터 2020년 9월까지 약 4년 9개월간의 데이터를 년 단위로 수집하였다. 국내에서는 아직까지 디지털 퍼블리싱에 대한 개념적 정의가 명확히 확립되지 않아, ‘디지털 퍼블리싱’ 단어로만 데이터를 수집하는 데는 한계가 있었다.

출판을 근간으로 확장된 개념을 일컫는 단어임에도 이와는 상관없는 기사들이 수집되기도 했다. 디지털 퍼블리싱은 같은 의미지만 ‘전자출판’, ‘ePub’, ‘eBook’, ‘전자책’, ‘디지털 출판’이라는 다양한 이름으로 혼재되어 사용되고 있어, 이들과 관련된 데이터들도 함께 수집하여 분석할 필요가 있었다.

이와 같은 방식으로 수집된 데이터량은 총 23,012 건이었으나 이 중 중복된 글이나 오류가 있는 글들은 제거하고 총 22,932건을 최종 데이터로 분석에 활용하였다. 수집 데이터 정보는 [표 1]과 같이 정리였다.

[표 1] 수집데이터 정보

구분	내용
검색어	디지털 퍼블리싱, eBook, 전자출판, ePub
수집범위	네이버 뉴스
수집기간	2016년 1월 1일 ~ 2020년 9월 30일
수집단위	년 단위
수집도구	파이썬 + 네이버 API
수집량	22,932건

수집된 데이터는 연도 별로 CSV 파일로 저장하고, 판다스(Pands) 라이브러리를 이용해 하나의 데이터로 합친 후 이를 빅 데이터 분석이 용이한 데이터프레임(DataFrame)의 형태로 변환하였다.

텍스트들을 TF-IDF분석과 의미 연결방 분석, LDA 토픽 모델링 분석을 위해 형태소 분석기인 KoNLPy를 사용해 텍스트들을 토큰(Token)화 시킨 후 명사들을 추출하였다. 토픽 모델링 분석을 위해서는 하나의 독립된 명사로는 토픽이 담고 있는 주제의 의미를 명확하게 파악하기 쉽지 않기 때문에 n개 단어의 연쇄를 확률적으로 표현해 주는 N-gram(언어모델)을 활용하여 단어 묶음을 2단어의 조합인 Bigram으로 확장시켰다. 토큰 라يز어를 활용해 불용어, 특수문자, 한 음절 키워드 같이 빅 데이터 분석에 불필요한 텍스트들에 대해서도 제거 작업을 실시하였다. 정제 과정에서 디지털 퍼블리싱 관련 전문가의 자문을 얻어 객관성을 확보하고자 하였다.

3-2. TF-IDF 분석

의미 연결망 분석과 LDA 토픽 모델링 분석을 위한 INPUT 데이터를 가공을 위한 문서 내에서의 단어 중요도에 따른 단어들을 추출하고자 하였다. 이를 위해 유클리디언(Euclidean) 거리법을 적용해 문서 내에서의 인접 행렬 값을 구하였다. TF-IDF 분석은 TF(Term Frequency)와 IDF(Inverse Document Frequency)를 곱한 값으로 어떤 단어가 특정 문서 내에서 얼마나 중요한지를 나타내는 통계적 수치이다. 이는 다수의 문서에서 빈번히 등장하는 단어는 중요하지 않을 확률이 높다는 가정을 기초로 한다. 단어의 단순 빈도를 통해 중요도를 측정하는 것이 아니라, 문서 내에서 비중 있게 다루어지는 단어를 찾는 방법이다.⁹⁾

TF-IDF 분석을 위해서는 단어들에 대해 TF-IDF 가중

9) <https://ko.wikipedia.org/wiki/Tf-idf>

치를 적용해야 하는데 이는 TfidfTransformer 라이브러리를 활용했다. 또한 컴퓨터가 텍스트들에 대한 통계적 수치 계산을 수행하기 위해서는 이산화가 작업이 필요한데 이를 위해 넘파이(NumPy) 라이브러리를 사용하였다. TF-IDF 분석 결과를 상위 45위까지 정리해보면 [표 2]와 같다.

[표 2] TF-IDF에 따른 단어별 순위

순위	단어	TF-IDF	순위	단어	TF-IDF
1	전자책	3637.1156	41	전자도서관	113.5130
2	서비스	752.3368	42	시장	112.1604
3	디지털	581.4594	43	단말기	108.3712
4	도서관	454.6402	44	열람	105.5360
5	퍼블리싱	443.9108	45	기능	99.4801
6	콘텐츠	441.2602	46	경보	96.7943
7	eBook	436.7577	47	웹툰	96.2797
8	온라인	368.4684	48	기업	95.2768
9	도서	343.6424	49	자료	94.5331
10	오디오북	336.8660	50	스마트	92.5287
11	무료	328.5473	51	교육	92.5165
12	이용	326.7195	52	소셜	91.1860
13	홈페이지	323.5559	53	PC	90.7257
14	국내	250.7769	54	프로그램	90.2322
15	발간	249.1761	55	서점	90.1435
16	플랫폼	225.2354	56	확대	88.5329
17	종이책	215.6685	57	비대면	86.7683
18	스마트폰	207.2560	58	활용	84.7601
19	독서	204.9412	59	아마존	83.5647
20	형태로	198.4736	60	무제한	81.4380
21	모바일	182.9759	61	쉽게	79.1598
22	서재	182.9759	62	대비	63.0022
23	작가	165.2928	63	교보문고	62.9200
24	출시	162.0775	64	영상	62.6136
25	사업	161.4665	65	매출	59.8454
26	코로나	153.3181	66	디자인	59.0387
27	제작	131.3431	67	웹소설	58.3982
28	개발	129.3345	68	ePUB	55.4108
29	운영	129.0591	69	유통	55.3166
30	저작권자	128.3251	70	전자책문학	54.9780
31	새로운	123.6791	71	기간	54.9205
32	누구나	123.0952	72	오리지널	54.7754
33	앱	122.0252	73	작품	54.0405
34	베스트셀러	120.3919	74	투자	53.7688
35	구독	118.0238	75	공모전	53.6868
36	대출	115.3875	76	시리즈	53.4641
37	글로벌	114.4931	77	학습	52.5681
38	출판	114.0114	78	월경액	51.6193
39	분야	113.9608	79	후관	50.0312
40	스토어 원	113.5130	80	도서정가제	50.7851

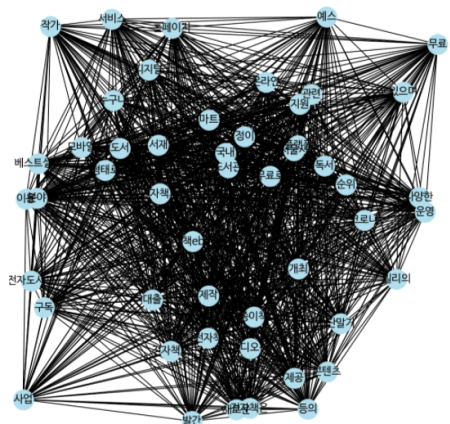
TF-IDF 분석 결과 ‘디지털’과 ‘퍼블리싱’을 대표하여

‘전자책’이라는 단어가 가장 많이 사용되고 있음을 알 수 있었다, 1위부터 21위까지는 ‘전자책’, ‘eBook’, ‘스마트폰’, ‘오디오북’과 같이 ‘디지털 퍼블리싱’의 서비스 방식과 형태에 대한 키워드들이 주를 이루고 있는데 이는 다양한 형태와 방식으로 서비스되고 있는 ‘디지털 퍼블리싱’의 특징이 반영된 결과라 할 수 있다, 22위부터 50위권 내에 분포된 단어들의 특징을 살펴보면 디지털 퍼블리싱을 통한 서비스 출시와 이에 대한 소비자의 이용과 관련된 단어들을 알 수 있다. 특히 주목해야 할 점은 26위의 ‘코로나’인데 이는 비대면 환경에서 ‘디지털 퍼블리싱’ 콘텐츠에 대한 소비자의 기대와 요구가 크다는 것으로 분석될 수 있다[그림 3]



[그림 3] TF-IDF 빈도에 따른 워드클라우드

TF-IDF 분석을 통해 추출된 단어 간 인접 행렬 값을 이용해 이를 의미 연결망으로 시각화하면 [그림 4]와 같다. 의미 연결망은 노드(Node)의 거리를 통해 단어 간 관계 밀접도를 파악할 수 있는 동시에 노드 수에 따라 해당 단어가 얼마나 많은 단어들과 서로 연관되어 있는지를 파악할 수 있다.



[그림 4] Semantic Network Analysis

3-3. 디지털 퍼블리싱 트렌드 분석

본 연구가 학술적 의미의 연구를 넘어 산업 현장에서도 유용하게 활용되기 위해서는 단순히 이론적 연구에 그치는 것이 아니라 디지털 퍼블리싱에 대한 실질적 접근이 필요하다고 판단되었다. 따라서 사용자를 포함한 산업 현장에서 디지털 퍼블리싱이 활용되고 있는 범위와 기술 현황 및 트렌드를 파악하는 것이 우선시되었다. 이를 위해 '디지털 퍼블리싱'에 대한 연관어 분석을 수행하였고, 이를 통해 추출된 단어들은 [표 3]과 같다.

[표 3] '디지털 퍼블리싱' 트렌드 관련 주요 단어

No	단어	연관도	No	단어	TF-IDF
1	책북전자책	0.9969	41	오디오북	0.9134
2	점자전자책	0.9966	42	오더블	0.9062
3	정액제	0.9957	43	응용콘텐츠	0.9013
4	빅 데이터	0.9955	44	커넥티드	0.8906
5	인공지능	0.9945	45	무료	0.8955
6	회원제	0.9940	46	플랫폼	0.8911
7	오버드라이브	0.9936	47	애플 원	0.8870
8	스트리밍북	0.9846	48	원 스토어	0.8821
9	스마트테이블	0.9813	49	융복합	0.8789
10	바이백서비스	0.9798	50	블록체인	0.7802
11	북튜브	0.9763	51	전자도서관	0.8725
12	콘텐츠전자책	0.9729	52	App	0.8653
13	헬스케어	0.9718	53	월정액	0.8647
14	API	0.9602	54	증강현실	0.8638
15	채팅	0.9557	55	리드북	0.8476
16	북클럽	0.9513	56	유튜브	0.8599
17	아마존	0.9479	57	북큐레이션	0.8460
18	어도비	0.9420	58	아티스트	0.8317
37	구글북스	0.9309	77	전자교재	0.8295
38	elibrary	0.9266	78	넷플릭스	0.8295
39	부르크	0.9222	79	인스타그램	0.8284
40	스마트폰	0.9178	80	하이브리드	0.8173

디지털 퍼블리싱에 대한 현 상황과 트렌드를 연관어 분석을 통해 분석한 결과 다양한 IT 기술이 ePub, ebook 등과 융합되어 기존의 전자책에서 벗어난 새로운 개념으로 확장되고 있음을 알 수 있었다. 초기의 전자책은 종이로 출판된 서적을 PC나 모바일을 통해 읽을 수 있도록 파일 포매팅의 변환 정도만 인식되었으나, 현재 디지털 퍼블리싱은 이러한 개념을 넘어 출판 산업에 있어서의 공유 콘텐츠로 여겨지고 있다. 특히 코로나19 확산으로 인해 사람들이 도서관이나 서점에 직접 가기보다는 공간과 시간의 제약이 적고 멀티태스킹이 가능한 디지털 퍼블리싱 관련 서비스들에 대한 관심을 볼 수 있다. 특히 트렌드 조사에서 1위에

위치해 있는 책북 전자책은 기존 컴퓨터가 책을 직접 읽어주는 오디오북을 넘어 채팅 형식으로 책을 설명해 주고 큐레이션 해주는 서비스로, 기존 오디오북 보다 진화된 모습을 보여주고 있다. 이외에도 스마트폰, 태블릿, 노트북 등 기기에 상관없이 언제 어디서나 볼 수 있고, 원하는 크기로 텍스트나 이미지, 레이아웃을 조정하여 사용할 수 있는 스트리밍북 서비스도 활성화되기 시작했다. 어도비, 아마존, 구글 등 글로벌 IT기업들은 이미 스트리밍북 시장에서 우위를 확보하기 위한 전쟁을 펼치고도 있다. 그만큼 디지털 퍼블리싱 시장은 4차 산업 혁명의 핵심 기술이라 할 수 있는 빅 데이터, 인공지능 등과의 융합을 통해 지금까지 예측하지 못했던 다양한 형태의 콘텐츠와 플랫폼 서비스를 창출할 것으로 보인다. 9위에 위치한 스마트테이블은 학교나 직장에서 사용하는 테이블에 초대형 태블릿 PC를 접목해 디지털화시킨 것으로, 이는 향후 교육과 업무가 디지털 퍼블리싱을 통해 생산되는 서비스들로 대체될 수 있는 가능성을 시사한다. 또한 시각장애인용으로 점자 출력기를 이용한 점자전자책까지 사용되고 있어 디지털 퍼블리싱은 인쇄물이 가지고 있는 다양한 활용성을 보여준다. 마지막으로 39위에 위치해 있는데 부르크는 온라인 출판 플랫폼으로서 누구나 부르크를 통해 출판사를 거치지 않고 바로 출판을 할 수 있는 1인 퍼블리싱 서비스이다. 다양한 기술들과 융합되어 빠르게 진화되어 발전하고 있는 디지털 퍼블리싱은 단순히 디지털 도구를 이용한 또 다른 출판 방식 정도로만 인식해 왔던 기존의 개념을 재정립해야 할 필요성이 요구된다.

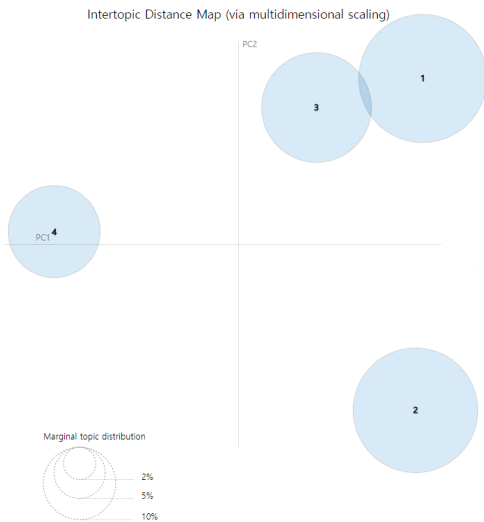
3-4. LDA 토픽 모델링 분석

문서에서 단어들의 동시 사용 패턴을 분석함으로써 해당 문서가 어떤 의미의 주제들을 설명하고 있는지를 추론하는 잠재 디리클레 할당(Latent Dirichlet allocation, LDA) 분석을 위해, TF-IDF 분석을 통해 산출된 인접 행렬 값을 LDA 모델의 INPUT 값으로 사용하였다. 또한 토픽 최적화를 위해 Coherence Score 산출을 위한 Gensim 라이브러리를 이용해 Coherence 모델을 만들고 Coherence Score 값을 [표 4]와 같이 구하였다.

[표 4] 토픽수 산출을 위한 Coherence Score

Topic	Coherence 값	Topic	Coherence 값
N=2	0.1850029725608	N=5	0.2946382283103
N=3	0.2842114675632	N=6	0.2873941664227
N=4	0.3845277894657	N=7	0.3491991253562

토픽의 수를 2개부터 7개까지로 설정하고 본 연구를 위해 가장 적절한 토픽의 수를 구한 결과 토픽 수를 6개로 설정했을 때 0.3845277894657로 가장 높은 수치를 나타냈으며 토픽 수를 2로 설정했을 때 0.1850029725608로 가장 낮은 수치를 보였다. 따라서 본 연구를 위한 LDA 토픽 모델링의 토픽 수를 4개로 설정하여 토픽 분석 모델에 적용하였다. 또한 보다 정밀한 분석을 위해 LDA 토픽 모델링의 하이퍼파라미터 값들을 조정하였는데 α 와 η 는 Auto로 설정하였으며, 머신러닝 훈련에 사용할 페이지를 설정하는 chunksize는 100개로 조정하여 LDA모델을 완성하였다. 생성된 머신러닝 모델에 학습시킨 결과 4개의 토픽에 대한 좌표값들이 [그림 5]와 같은 결과가 나타났다. LDA 토픽 모델링의 결괏값은 pyLDAvis라이브러리를 이용해 시각화하였다.



[그림 5] 토픽 좌표 시각화

각 토픽들의 좌표 상 위치를 보면 토픽 1과 3이 근접한 위치에 위치해 있고, 나머지 토픽들은 상대적으로 많이 떨어져 나타났다. 토픽 1과 3 역시 근접해 있기는 하나 교차하고 있는 범위가 매우 적기 때문에 각 토픽의 판별 타당성이 매우 높다 할 수 있다. 또한 각 토픽을 이루고 있는 원의 크기가 비슷한 크기로 나타나 같은 빈도를 가진 단어들로 구성되어 있음을 알 수 있었다.

머신러닝 모델에 의해 각 토픽을 이루는 단어들에 대한 분포는 [표 5]와 같이 나타났다. 각 토픽에 할당

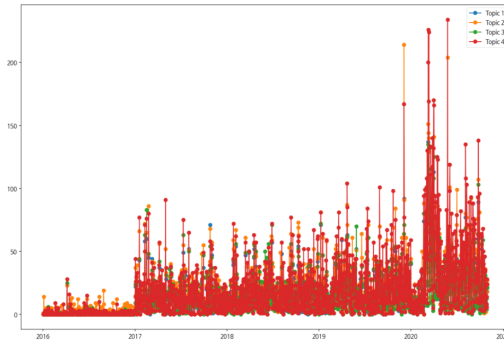
된 단어들을 추정하고, 단어들이 포함된 기사들을 추출하여 해석해 보았을 때 1번 토픽은 플랫폼, 2번 토픽 디지털 퍼블리싱과 산업, 3번 토픽은 서비스 이용 패턴 4번 토픽은 시대 환경에 따른 콘텐츠 접근에 대한 주제들을 설명하고 있는 것으로 해석된다.

[표 5] 단어별 토픽 가중치

토픽	단어별 토픽 가중치			
1	전자책	eBook	홈페이지	플랫폼
	0.0118415	0.00899616	0.00883565	0.00790218
	월정액	모바일	아마존	무제한
	0.0078700	0.00606738	0.00489199	0.00444726
	무료	오디오북	콘텐츠	구글북스
2	0.0042735	0.00390393	0.00359856	0.00356741
	디지털	전자책	출시	제공
	0.0506580	0.01900415	0.01816534	0.00974929
	사업	점자전자책	유통	오디오북
	0.0094120	0.00857169	0.00809749	0.00733181
3	퍼블리싱	ePub	개발	글로벌
	0.0069592	0.00660475	0.00539064	0.00524551
	디지털	퍼블리싱	모바일	콘텐츠
	0.0109534	0.00668012	0.00488599	0.00483949
	서비스	동영상	도서관	아마존
4	0.0046149	0.00428828	0.00360321	0.00325877
	베스트셀러	유튜브	스마트폰	홈페이지
	0.0031973	0.00311217	0.00310198	0.00274556
	전자책	콘텐츠	코로나	오디오북
	0.0341212	0.01351565	0.00654361	0.00600618
	비대면	무료	웹소설	스마트폰
	0.0058716	0.00583137	0.00531125	0.00477144
	웹툰	아마존	교보문고	무제한
	0.0039150	0.00378637	0.00362724	0.00352339

토픽 1은 전자책(0.0118%), eBook(0.0089%), 홈페이지(0.0088%), 플랫폼(0.0079%)순으로 토픽이 구성되어 있었으며 토픽의 주제를 구체적으로 판단하기 위해 해당 단어들이 포함된 기사 목록들을 별도 추출하여 주제에 접근하였다. 토픽 1의 추론 결과 구글북스, 어도비시스템즈에서 운영하는 e스트리밍 플랫폼과 플랫폼을 이용하기 위한 이용 방법 등에 대한 내용이 대부분을 차지하고 있는 것으로 나타났다. 토픽 2는 디지털(0.050%), 전자책(0.019%), 출시(0.018%), 제공(0.0097%)순으로 토픽이 구성된 것으로 나타났다. 디지털 퍼블리싱과 관련된 국내 기업들의 서비스 출시와 같은 내용이 주를 이루고 있었다. 토픽3은 디지털(0.010%), 퍼블리싱(0.0066%), 모바일과 콘텐츠(0.0048%)순으로 구성되어 있으며, 디지털 퍼블리싱의 서비스 소비자의 이용 패턴에 관한 주제를 설명하

고 있는 것으로 판단되었다. 마지막으로 토픽 4는 전자책(0.034%), 콘텐츠(0.013%), 코로나(0.0065%), 오디오북(0.0060%)의 단어들로 구성되어 있었으며 시대 환경에 따른 소비자들의 접근 내용에 대한 주제를 나타내고 있는 것으로 파악되었다.



[그림 6] 토픽 시계열 분석

토픽 모델링 분석을 위한 마지막 단계로 2016년 1월부터 2020년 9월까지 수집된 데이터를 시계열로 도식화하고 시기별로 디지털 퍼블리싱과 관련하여 어떤 주제들이 이슈화되었는지를 살펴보았다. 시계열에서 나타난 결과값은 2019년까지 큰 변화 없이 일정한 보도량을 보여주다 코로나19가 발생한 2020년 급격한 변화를 보이고 있었다. 그래프상에서의 급격한 변화는 토픽 4가 이끌고 있었는데, 이는 코로나19로 사회적 거리두기가 시행되면서, 소비자의 디지털 퍼블리싱 관련 서비스에 이용이 증가된 결과로 보인다. 이러한 변화가 일시적인 현상에 그칠 수도 있었으나, 코로나19 이후 4차 산업의 영향과 언택트 사회로의 전환으로 인해 디지털 퍼블리싱은 2020년을 기점으로 하여 확연한 성장세를 보일 것으로 예상된다.

4. 결론

본 연구 결과를 종합해 보았을 때, TF-IDF 분석 결과를 디지털 퍼블리싱의 서비스 방식과 서비스 형태에 대한 키워드들이 상위권을 차지하고 있음을 알 수 있었다. 또한 비대면 환경에서의 서비스 이용 관련 단어들이 많이 등장했음을 파악할 수 있었다. 연관어 분석을 통한 트렌드 분석에서는 전통적인 전자책의 개념에서 벗어나, 채팅을 통한 책보이나 장애인을 위한 점자

전자책, 딥러닝 등의 4차 산업과 관련된 다양한 서비스들이 조사되었다.

LDA토픽 모델링 분석을 통해서 총 4개의 토픽을 추출하였는데 토픽 1은 구글북스, 아마존, 어도비시스템즈에서 운영하는 e-스트리밍 플랫폼과 같이 디지털퍼블리싱 서비스를 위한 플랫폼에 관한 주제를 담고 있었다. 토픽 2는 디지털 퍼블리싱과 관련된 국내 기업들의 개발 및 유통 산업 현황을 주제로 담고 있었다. 토픽 3은 디지털 퍼블리싱 서비스 이용을 위한 이용 패턴에 관한 주제를 나타냈다. 마지막으로 토픽 4는 환경에 따른 디지털 퍼블리싱 콘텐츠에 대한 소비자들의 접근 특성에 대한 주제를 담고 있었다.

본 연구의 한계점으로 디지털 퍼블리싱이라는 단어가 아직까지 특정 산업을 대표하는 대표성이 부족한 관계로 디지털 퍼블리싱이라는 단일 키워드로서 검색할 수 있는 데이터양이 많지 않았으며 검색 데이터에서 직접적으로 연관성이 없는 글들도 다수 포함되어 있었다. 따라서 전자책, ePub, 전자출판, 디지털북 같이 디지털 퍼블리싱의 의미를 내포하고 있지만 제각각 다른 단어를 적용하여 데이터를 수집하였다.

빅 데이터 분석은 수집된 데이터를 정제하는 과정에서 연관성이 없는 단어를 정제해야 하는 과정에서 연구자의 판단이 필요하며 이로 인해 전문가의 자문을 통해 정제하였으나 완벽한 객관성을 확보하기에 부족함이 있다. 그러나 현 시대에 사용자와 구독자가 생각하고 요구하고 있는 디지털 퍼블리싱의 개념과 인식을 비정형 빅 데이터로 분석하여 의미 있는 토픽으로 그룹핑하여 시각화시켜 제시한 본 논문은 학술적인 연구로서 의의가 있다고 사료된다.

후속 연구에서는 본 연구에서 나타난 문제점을 보완하고 소비자의 인식에 대해 구체적으로 분석하기 위해 디지털 퍼블리싱의 구체적 분야인 디지털 매거진을 주제로 뉴스 자료뿐 아니라, SNS, 블로그, 카페 같은 소셜 네트워크 상에 나타난 비정형 데이터들까지 취합한 확장된 연구를 진행한다면 실증적인 결과를 얻을 수 있을 것으로 기대한다.

참고문헌

1. 김수영, 2019 출판산업 실태조사, 한국출판문화산업진흥원, 2020
2. 김소라, 빅데이터 분석을 이용한 아노락 패션의 인식에 관한 융합 연구. 한국과학예술융합학회, 38(4), 2020
3. 김상욱, 스마트 디바이스를 이용한 디지털 콘텐츠 커뮤니케이션, (33), 2011
4. 김상욱, 권기정, 스마트 디바이스 환경의 디지털 퍼블리싱 커뮤니케이션 연구, 한국디자인트렌드학회, 2013
5. 김중선, “텍스트마이닝을 통한 패션 스트리밍 서비스 인식에 관한 연구”, 한국패션디자인학회지, 18(1), 2018
6. 공병훈, 조정미, 국내 디지털 퍼블리싱 연구에 대한 메타분석, 한국출판학연구, 44권 3호, 2018
7. 류우중, 하종우, Md. Hijbul Alam, 이상근, 토픽 모델링 기법을 이용한 트위터 트렌드 추출, 한국정보과학회 학술발표논문집, 2013
8. 민혜중, 김영경, 토픽모델링을 이용한 매체 별 빅데이터 동향 다각적 분석. 대한산업공학회 추계학술대회 논문집, 2015
9. 박혜진, ‘시각디자인 교육 기반의 디지털 퍼블리싱 애플 디자인 체계 연구’, Journal of Integrated Design Research, 16(3), 2017
10. 이말레, 배한국. “TFIDF 를 이용한 키워드 추출 시스템 설계”, 인지과학, 13(1), 2002
11. 이윤주, 서지훈, 최진탁, “SNS 텍스트 콘텐츠를 활용한 오피니언마이닝 기반의 패션 트렌드 마케팅 예측 분석”, 한국정보기술학회논문지, 12(12), 2014
12. 이재원, 원하연, 최경란, 정도성, SNS 빅데이터의 집단감성정보 제공이 디자인 집단의 창의성에 미치는 영향 연구. 디지털디자인학연구, 15(3), 2015
13. 임해진, 최장식, 전진영, 변형기, 유클리디언 거리 기반의 단계적 소거 방법을 통한 화학센서 어레이 성능 최적화, 센서학회지, 2015
14. 장경남, 김민정, “토픽모델링을 이용한 국내 패션디자인 연구 동향”, 디지털융복합연구, 15(6), 2017
15. 전지현, 종이책 대신 전자책 읽는 사람 늘었다, 매일경제, 2020
16. 최영현, 이규혜, “텍스트 마이닝을 활용한 윤리적 패션 연구동향: 2009-2019 연구 네트워크 분석”, 한국의류산업학회지, 22(2), 2020
17. Abdullah Alfadda, Probabilistic Graphical Models Class Project, ECE, Virginia Tech, 2014
18. Blei, David M., Andrew Y. Ng, and Michael I. Jordan, “Latent dirichlet allocation”, the Journal of machine Learning research, 2003
19. ko.wikipedia.org
20. www.naver.com
21. www.epnc.co.kr