미디어에 나타난 직업 관련 데이터의 분석

반재훈 · 정윤성 · 정동민

고신대학교 IT경영학과

Analysis of Job Data on Media

ChaeHoon Ban · YoonSeung Jung · DongMin Jeong

Dept. of IT Management, Kosin University

E-mail : chban@kosin.ac.kr, vision8510@naver.com, jdmin96@naver.com

요 약

과거와는 비교 할 수 없을 만큼 방대한 양의 데이터가 생산되는 정보화 시대에서 과거와 현재의 데이터를 비교 분석하는 것이 매우 중요하다. 이러한 데이터를 분석하는 도구인 R은 통계 기반의 정보분석을 가능하게 하는 언어와 환경이다. 본 논문에서는 R을 이용하여 미디어에 나타난 직업 관련 빅데이터를 분석한다. 다양한 미디어에서 직업 관련 데이터를 수집하고 어떠한 텍스트가 분포되어 있는지 빈도 조사를 수행한다.

키워드

Big Data, R, Text Mining, Job, Analysis

I. 서 론

정보기술과 디지털 경제의 확산으로 대규모의 데이터가 생산되는 정보화시대에 내포되어 있는 빅 데이터의 시대에 도래했다. 최근 핵심 이슈로 부각되면서 빅데이터의 중요성이 강조되고, 미래 경쟁력의 자원의 원천이 되며, 관련 기술의 발전, 자격증 등 다양한 분야에 활용됨으로 빅 데이터에 의미가 중요하다고 볼 수 있다.

직업이란 생계를 유지하기 위해 자신의 적성과 능력에 따라 일정 기간 동안 계속하여 종사하는 일을 말한다. 이번 논문에서는 각종 미디어와 신문사 등에서 나타난 직업에 관한 데이터를 분석 및 조사하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 빅데이터 기법에 관련 연구 및 수집 방법을 기술한다. 3장에서는 본 논문에서 구현한 워드 클라우드 형태의 그림을 표현하기 위해 R 프로그램 활용 방법을 설명한다. 4장에서는 워드 클라우드 형태의 그림으로 표현한 직업 관련 데이터를 시각화하였고 수집 데이터의 가장 많이 검색된 데이터를 나열하였다. 5장에서는 결론 및 향후연구에 대해 기술한다.

II. 관련연구

기존의 연구에서는 데이터 마이닝, 텍스트 마이닝, 오피니언 마이닝, 웹 마이닝, 소셜 마이닝 기법 등 다양한 기법을 통한 빅 데이터 분석연구가 있었다. 정보통신의 발달과 소셜 미디어의 급속한 확산으로 빅 데이터가 경제적으로 자산이 되고 있는 시대를 맞이하는 데 필요한 데이터 분석기법과 인프라 기술에 대해 알아보고, 한글 Text 데이터를 R 프로그램을 이용하여 usesejongdic() 이라는 옵션을 이용하여 명사만 추출하는 방법으로 비정형 데이터를 분석하였다[1]. 데이터 시각화 도구 통계 패키지인 R을 이용하여 대기오염의 자료를 여러 가지 방법의 데이터 시각화를 통하여 나타내었고, 데이터 시각화 방법별로 통계적인 방법을 활용한 분석과 연계하여 어떤 특징이 있는지를 나타냈다. 2차원의 히스토그램과 선점도, 상자그림, 3차원 산점도와 투시도 등 다양한 방법의 그래프를 구현하여 오존농도와 설명 변수들 간에 어떠한 관련성이 있는지를 분석했다[2]. 빅데이터 분석 도구인 R을 이용하여 빠른 시간 안에 사용자가 목적으로 하고 있는 특허검색 결과를 효율적으로 도출할 수 있는 검색어 추출에 관한 연구를 진행했다[3]. 데이터 마이닝의 일부인 텍스트 마이닝의 기법을 이용하여 직업 관련 키워드를 국민일보, 중앙일보, 세계일보 등 각 신문사에서 데이터를 추출하였다. 그리고 관계 또는 관련 있는 데이터에 내재되어 있는 의미 있는 패턴을 찾는 사회네트워크분석을 실시하여 비정형화된 빅 데이터를 정형화 및 시각화하고 해석했다[4]. 구글, 야후, 네이버 등 주요 포털의 지도에는 POI(Point of interest)가 서비스 되고 있다. 지도의 위치 데이터 즉, 현재 이용자가 위치한 장소

는 인문학적인 스토리텔링의 시작점을 주목하여, POI는 카페, 레스토랑, 병원, 식당 등의 정보만이 서비스되는 한계점을 지적하고, 더 나아가 대안으로 POI 정보와 결합된 소위 '인문융합 지도 서비스'를 제안 했다[5].

III. 데이터 분석 방법

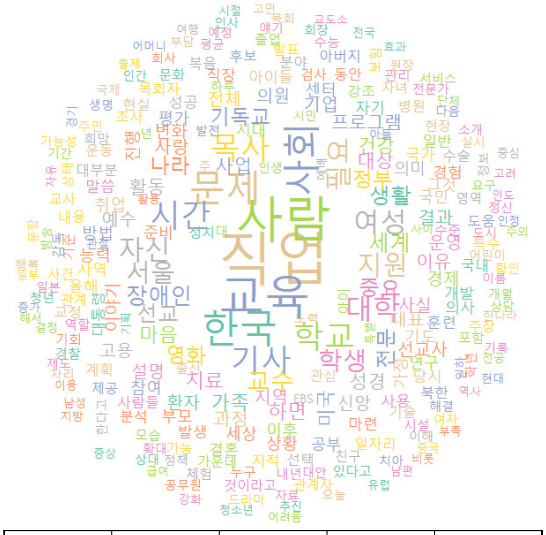
데이터 분석 도구인 R을 이용하여 중앙일보, 세계일보, 국민일보 등 여러 신문사를 통해 '직업'이라는 키워드를 가지고 검색을 했을 때 나오는 결과를 현 시점인 2018년 기준으로 2년 단위로 끊어 2010년부터 2018년 까지 텍스트(txt)로 된 파일 데이터를 수집하였다. 그리고 본 논문은 사설·칼럼을 포함한 데이터 이다. 그리고 직업이라는 데이터를 분석하기 위한 도구로 R을 사용 하였는데, 데이터의 분석과정은 그림 1과 같다.

그림 1. 데이터 분석 과정.

데이터 분석도구인 R을 설치한다. 설치에 필요한 파일은 R, R Studio, Java 이다. 설치가 완료된 후에는 한글 데이터 분석에 필요한 패키지("KoNLP"), 워드 클라우드 생성에 필요한 패키지("wordcloud")를 설치하고 R 소스에 로딩(다운로드)한다. 직업 데이터를 2011년, 2012년 등 각 년도 별로 구분하여 각 그룹의 직업 데이터를 변수를 할당하여 대입한다. 한글의 명사를 추출해주는 함수인 'extracNoun'함수를 사용함으로서 성경 데이터를 명사로 변환하여 변환된 데이터를 확인 후 원하지 않는 데이터에 대한 'gsub'함수를 이용하여 데이터를 필터링 한다. 또한 여기서는 nchar(word)를 이용해 2자리 이상의 명사만 추출하도록 프로그램을 구현하였다. 필터링 된 데이터를 텍스트 형식의 파일로 저장하여 테이블 형태로 변환하여 변수에 할당한다. 텍스트 형태로 각 명사에 대한 빈도수를 측정하여, 상위30위의 결과를 워드 클라우드 형태의 그래픽으로 출력한다. 출력 결과물을 이미지파일(JPGE, BMP, PNG 등)으로 저장한다.

IV. 직업 데이터 분석 결과

본 논문에서는 직업 관련 데이터를 각 신문사에서 수집하여 R을 이용한 워드 클라우드와 수집한 데이터 안에 들어가 있는 단어의 빈도수 에 대하여 표현하였다. 워드 클라우드란 문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 돋보이게 하는 기법이다. 다음에 나타날 것과 같이 텍스트가 많이 언급 될수록 단어의 크기가 커진다. 그로 인해 시각적으로 보기가 편할 뿐만 아니라 데이터 수집 후 분석, 통계에 용이하다.



직업	사람	교육	한국	사회
2328	1944	1841	1475	1414
문제	학교	기사	시간	목사
1366	1140	1133	1116	1081
대학	여성	필요	자신	지원
1015	959	951	917	867
서울	학생	교수	장애인	선교
830	819	777	698	682
중요	세계	생활	정부	기독교
655	653	652	640	627
영화	전문	나라	마음	치료
614	609	603	601	590

그림 2. 2010년~2011년 직업과 관련된 신문기사 분석 결과.

그림 2는 2010년에서 2011년의 신문기사를 워드 클라우드를 통해 수집한 데이터를 시각화 한 것이다. 아래의 표는 어떠한 단어가 가장 많이 도출이 되었는가를 나타내는 표이다. 가장 많은 단어로는 '사람'이고 다음으로는 '교육', '한국', '사회' 등으로 나타났다.

직업	교육	지원	여성	취업
3323	1895	1174	979	967
학생	사람	기업	대학	경우
886	855	805	805	798
이상	진로	일자리	프로그램	대상
779	723	714	655	654
결과	사회	과정	사업	생각
650	627	618	616	616
청소년	활동	학교	운영	진행
614	608	593	549	542
전문	다양	교사	시간	직장
526	522	504	504	499

그림 3. 2012년~2013년 직업과 관련된 신문기사 분석 결과.

직업	교육	진로	체험	학생
14127	4596	3671	2995	2289
지원	프로그램	취업	전문	대학
1774	1664	1558	1514	1435
운영	과정	학교	능력	분야
1319	1315	1286	1270	1211
일자리	미래	청소년	전문가	센터
1197	1191	1123	1073	1047
기업	사회	여성	개발	고용
1025	1014	1013	993	969
교사	제공	진행	활동	사람
946	926	891	884	870

그림 4. 2014년~2015년 직업과 관련된 신문기사 분석 결과.

그림 3은 2012년에서 2013년의 '직업'이라는 키워드를 가지고 데이터 수집하여 워드 클라우드를 통해 시각화 한 것이다. 가장 많이 나온 단어는 큰 단어로 표시 된다. 표는 '직업' 이라는 단어가 가장 많이 나왔으며, 다음으로는 '교육', '지원', '여성' 순으로 나타났다.

그림 4는 2014년에서 2015년의 신문기사에 나타난 '직업' 이라는 키워드를 가지고 데이터를 수집한 자료이다. 표에서 나타난 단어들은 '직업', '교육', '진로', '체험', '학생' 순으로 많이 나타났다.

그림 5에서는 다른 년도와는 다르게 직업이 아닌 '교육' 이라는 단어가 가장 큰 비중을 차지했다. 표에서도 알 수 있듯이 '교육', '대학', '학생', '직업' 순으로 가장 많은 단어가 나타났다.

마지막으로 그림 6에서는 2010년부터 2017년 전체의 데이터를 통합하였다. 그 결과 표에서 '직업' 그 다음으로는 '교육'이라는 단어가 가장 많이 도출된 것을 알 수 있다. 이는 우리나라에서 직업 교육을 많이 실시하고 있기 때문이다.

V. 결론 및 향후 연구

정보기술과 디지털 경제의 확산으로 대규모의 데이터가 생산되는 정보화시대에서 빅데이터의 중요성이 강조되고 있으며 다양한 분야에서 응용하고 있다. 본 논문에서는 데이터 분석도구인 'R'을 이용하여 '직업' 이라는 키워드를 가지고 현시점인 2018년부터 8년 전인 2010년 전까지의 데이터를 각 신문사의 신문기사를 통해 비정형 데이터를 정형화 하였다. 또한 정형화한 데이터를 가지고 워드 클라우드를 하여 데이터를 시각화함으로써 누구나 쉽게 해당 정보에 접근할 수 있다.

향후 연구 방향으로는 직업 관련 빅데이터를

교육	대학	학생	직업	사회
2065	1106	1099	1098	828
지원	기업	과정	학교	사람
681	662	641	629	620
서울	전문	취업	한국	사업
580	560	529	523	514
정부	생각	다양	문제	대통령
487	474	463	462	461
운영	분야	시간	프로그램	대표
461	448	446	446	444
필요	교수	인재	일자리	미국
429	425	381	381	369

그림 5. 2016년~2018년 직업과 관련된 신문기사 분석 결과.

직업	교육	사람	대학	사회
6749	5801	3419	2926	2877
학생	지원	학교	한국	여성
2804	2722	2362	2327	2279
문제	시간	기업	취업	서울
2188	2066	1964	1930	1864
과정	전문	사업	프로그램	활동
1762	1695	1667	1607	1524
정부	교수	운영	결과	일자리
1514	1482	1450	1438	1422
미국	세계	진행	분야	대표
1328	1316	1303	1290	1275

그림 6. 2010년~2018년 직업과 관련된 신문기사 분석 결과.

수집하고, 출제 단어 빈도수를 보아서는 앞으로 한국의 직업 및 진로 결정의 방향을 알 수 있을 것이며, 부족한 점을 보완 할 수 있을 것으로 보인다. 해당 데이터의 사용방향으로는 고등학교나 대학교 또는 국가의 직업교육 프로그램에 사용될 것으로 보인다. 이 외에도 다양한 분야에 도움이 될 것이다.

참고문헌

[1] 김현근. R을 이용한 빅 데이터 사례 분석. 호서대학교 일반대학원 정보통계학과 석사학위논문, 2014.

[2] 오영창, 박은식. 2015. R 소프트웨어를 이용한 대기오염 데이터의 시각화. 한국데이터정보과학회지, 26(2), 399-408.

[3] 장청윤, 장정환, 김석주, 이현군, & 이창호. (2013). 빅데이터 분석 도구 R을 활용한 효율적인 특허 검색에 관한 연구. 대한안전경영과학회지, 15(4), 289-294.

[4] 이경준, 노윤환, 윤상경, 조영석. 2014. 부산지역 교통관련 기사를 이용한 비정형 빅데이터의 정형화와 시각적 해석. 한국데이터정보과학회지, 25(6), 1431-1438.

[5] 이원태, 강장묵. 2015. 빅데이터 중 POI와 공간메타포를 활용한 인문 융합 지도 연구. 한국인터넷방송통신학회. 15(3), 43-50.