벤처창업 관련 뉴스 및 SNS 빅데이터 분석

반재훈, 이예찬, 안대중, 곽윤혁

고신대학교 IT경영학과

The Venture Business Starts News and SNS Big Data Analytics

ChaeHoon Ban, YeChan Lee, DaeJoong Ahn, YoonHyeok Kwak

Dept. of IT Management, Kosin University

E-mail : chban@kosin.ac.kr ycl0188@naver.com spotforce17@gmail.com power8059@naver.com

요 약

대규모의 데이터가 생산되고 저장되는 정보화 시대에서 현재와 과거의 데이터를 바탕으로 미래를 추측하고 방향성을 알아갈 수 있는 빅데이터의 중요성이 강조되고 있다. 정형화 되지 못한 대규모 데이터를 빅데이터 분석 도구인 R과 웹크롤링을 통해 분석하고 그 통계를 기초로 데이터의 정형화와 정보 분석을 하도록 한다.

본 논문에서는 R과 웹크롤링을 이용하여 최근 이슈가 되고 있는 벤처창업을 주 키워드로 하여 뉴스 및 SNS에서 나타나는 벤처창업 관련 빅데이터를 분석한다. 뉴스기사와 페이스북, 트위터에서 벤처창업 관련 데이터를 수집하고 수집된 데이터에서 키워드를 분류하여 효율적인 벤처창업의 방법과 종류, 방향성에 대해 예측한다. 과거의 벤처창업 실패요인을 분석하고 현재의 문제점을 찾아 데이터 분석을 통해 벤처창업의 흐름과 방향성을 제시하여 창업자들이 겪을 수 있는 어려움을 사전에 예측하고 파악함으로써 실질적인 벤처창업에 크게 이바지할 것으로 보여 진다.

키워드

Big Data, R, Text Mining, Transportation, Analysis

Ⅰ. 서론

미디어의 발전과 확산으로 대규모의 비정형 데이터가 생산되는 정보화 시대에서 데이터를 수집하고 수집된 데이터를 이용하여 미래를 추측하고 방향성을 찾아가는 빅데이터 분석이 강조되고 있으며 다양한 산업에서 이를 활용하고 있다. 빅데이터 수집 사이트인 빅카인즈(Bigkinds)는 입력한 키워드의 데이터 수집을 가능하게 하는 사이트이다. 빅 데이터 수집 도구인 파이썬(python)은 데이터의 수집을 가능하게 하는 언어와 환경이다. 빅 데이터 분석 도구인 R은 통계기반의 정보 분석을 가능하게 하는 언어와 환경이다. 벤처창업의 데이터를 활용하여 창업을 어떻게 준비하고 진행하여야 할지 모르는 예비 창업자들에게 벤처창업의 방향성을 제시함으로서 벤처창업 데이터 분석의 중요성을 볼 수 있다.

'벤처창업'과 관련된 키워드를 중심으로 최근 5년간 데이터를 기준으로 주요 언론사(경향신문 외 40개 언론사)와 벤처창업관련 져널과 트위터에서 데이터를 수집하고, 키워드의 빈도를 도출하여 어떠한 벤처창업 관련 정보가 최근 5년간 노출되었는지 분석하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 '벤처창업'의 빅 데이터와 관련된 연구를 기술한다. 3장에서는 본 논문에서 구현한 워드 클라우드 형태의 그림을 표현하기 위해 R 프로그램을 활용한 데이터 분석 방법에 대해 기술한다. 4장에서는 벤처창업 관련 키워드 중 상위 30개의 빈도수를 보여주고 워드 클라우드 형태의 그림으로 표현한다. 마지막 5장에서는 결론 및 향후 연구에 대해 기술한다.

Ⅱ. 관련연구

기존의 빅데이터 분석 기술로는 데이터 마이닝, 텍스트 마이닝, 오피니언 마이닝, 소셜 마이닝 기법, 군집분석 등 다양한 분석 기술이 있었다. 정보통신의 발달과 소셜 미디어의 급속한 확산으로 빅데이터가 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]

빅데이터를 활용한 스마트 스타트업 콘텐츠 및 서비스 플랫폼 개발 - 연구내용:

○ 주관기관: (예비)창업자를 위한 정보 체계 및 기반 지식 체계 설계, 전체 시스템 아키텍처(Architecture) 설계와 기술 개념 확립, 창업자 대상 설문 조사를 통한 창업콘텐츠 요구분석

○ 참여기관1: 빅데이터 처리, 소셜 미디어 데이터 마이닝, 스마트 스타트업 콘텐츠 큐레이팅 등의 기술 개념 설계

○ 참여기관2: 네트워크(동업자, 투자자 등) 추천 기술 개념 설계[2] 한국과 미국의 창업교육 비교를 통한 한국 벤처창업 교육전략 수립- 미국에서의 벤처기업의 확산과 성장은 1980년대 불경기의 미국을 지난 1990년대의 10년간 사상 최대의 호황경제로 만들었다. 미국 벤처기업의 성공은 대학과 벤처기업간의 산학협력은 물론 대학의 질높은 벤처창업 교육에 있었다. 미국의 벤처창업교육은 한국에 비하여 상당히 오래되었고 거의 모든 대학에서 장기적이면서도 적극적으로 실시하고 있다. 이에 반해 한국은 그 역사가 일천하고 아직도 피상적인 수준이다 이에 본 연구는 미국과 문화적 역사적 다른 한국에서 성공적인 벤처창업교육을 위한 5가지 기본적인 전략을 제시하였다.[3] 현재의 국내의 많은 자영업자들이 창업의 실패를 경험하고 있다. 이러한 점에서 무분별한 창업을 줄이고, 창업의 성공률을 높이기 위해 창업 준비 과정에서 명확하고 통합된 정보의 제공이 요구된다. 본 연구는 다양한 공공기관들이 분산되어 보유하고 있는 다양한 데이터를 통합한 빅데이터를 제안하고자 한다. 이를 위해 창업에서 요구되는 데이터의 유형을 분류하고 통합적 창업지원 정보 시스템 구축을 위한 데이터 통합, 분석 기술, 창업자를 위한 웹 또는 스마트 서비스의 유형을 제시하고자 한다.[4]

III. 데이터 분석 방법

빅데이터 분석도구인 R을 이용하여 텍스트 데이터를 워드 클라우드 형태의 그림으로 표현하였다. 신문기사의 데이터는 빅카인즈를 이용하여 '벤처창업' 관련 키워드에 접속하여 본문내용을 중심으로 텍스트 파일의 데이터로 수집하였으며, 데이터의 분석을 위해 경향신문, 한국경제, 헤럴드경제, YTN, 파이낸셜뉴스 등 총 41개의 신문사 및 언론사의 지면기사를 기준으로 약 7100여 건 이상의 기사를 분석하였다.

데이터 분석도구인 R을 설치하고 한글 데이터 분석에 필요한 패키지("KoNLP"), 워드 클라우드 생성에 필요한 패키지("wordcloud")를 설치하고 R 소스에 로딩한다. 수집한 데이터를 최근 1년, 최근 3년, 최근 5년으로 분류하여 각 그룹의 데이터를 변수를 할당하여 대입한다. 한글의 명사를 추출해주는 함수인 'extracNoun'함수를 사용함으로서 벤처창업 데이터를 명사로 변환하여 변환된 데이터를 확인 후 원하지 않는 데이터에 대한 'Fliiter'함수를 이용하여 데이터를 필터링 한다. 여기서는 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, 상위 30위의 결과를 워드 클라우드 형태의 그래픽으로 출력한다. 출력 결과물을 이미지파일(JPGE, BMP, PNG 등)으로 저장한다. 데이터 분석과정은 [그림 1] 과 같다.

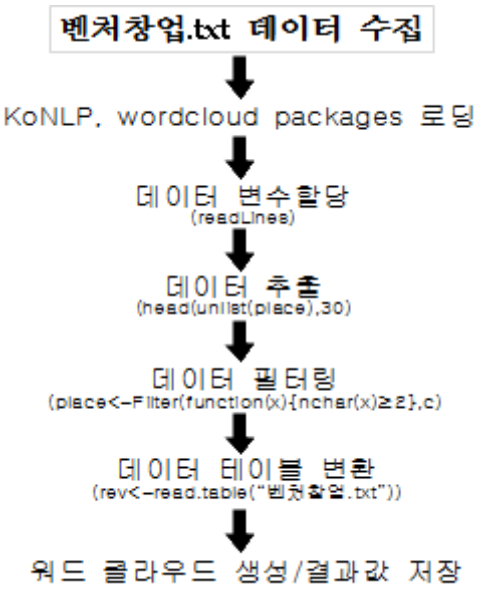

그림 1. 데이터 분석 과정

IV. 데이터 분석 결과 및 비교

본 논문에서는 '벤처창업' 관련 데이터 분석의 결과를 워드 클라우드와 키워드 빈도수에 대하여 표현하였다. 워드 클라우드란 문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 돋보이게 하는 기법이다. 예를 들면 많이 언급될수록 단어를 크게 표현해 한눈에 들어올 수 있게 하는 기법 등이 있다.

표 1. 언론사 1년 키워드 빈도수

창업	기업	지원	기술	청년
836	485	412	311	257
스타트업	벤처	창조	아이디어	혁신
255	202	157	148	126
성장	벤처기업	성공	일자리	창업자
124	112	109	100	87
계획	중소기업	전문가	진출	취업
84	76	72	63	63
창출	도전	드림	기회	대기업
62	61	59	55	55
가능성	정보	창업지원	창의	유망
45	45	45	41	35

[표 1]은 2016년 5월 1일부터 2017년 4월30일까지의 주요언론사 및 벤처창업 관련 전문지 총 40곳의 신문사 중 동아일보에서 최근 1년간 노출된 기사를 수집하여 총 단어 58945개 중 상위 500개의 키워드 추출 후 벤처창업과 유사한 순으로 30개 키워드 빈도를 표로 나열하였다.

표 2. 언론사 3년 키워드 빈도수

창업	기업	지원	경제	기술
2144	1479	1234	910	876
창조	투자	청년	벤처	개발
825	771	757	717	639
혁신	대학	스타트업	정부	벤처기업
618	592	532	512	496
성장	아이디어	글로벌	성공	취업
416	414	333	316	256
일자리	중소기업	드림	창업자	경영
236	234	231	225	196
창출	기회	행사	경험	특허
178	132	132	128	127

[표 2]는 2014년 5월 1일부터 2017년 4월30일까지의 주요언론사 및 벤처창업 관련 전문지 총 40곳의 신문사 중 동아일보에서 최근 3년간 노출된 기사를 수집하여 총 단어 178200개 중 상위 500개의 키워드 추출 후 벤처창업과 유사한 순으로 30개 키워드 빈도를 표로 나열하였다.

표 3. 언론사 5년 키워드 빈도수

창업	기업	지원	경제	창조
3258	2179	1640	1416	1222
기술	사업	벤처	투자	청년
1219	1197	1145	1102	1021
개발	대학	벤처기업	혁신	스타트업
858	852	798	701	624
성장	성공	아이디어	운영	중소기업
610	600	590	514	436
계획	창업자	정책	일자리	취업
426	360	359	351	313
드림	실패	펀드	도전	경험
309	294	282	256	243

[표 3]은 2012년 5월 1일부터 2017년 4월30일까지의 주요언론사 및 벤처창업 관련 전문지 총 40곳의 신문사 중 동아일보에서 최근 5년간 노출된 기사를 수집하여 총 단어 265735개 중 상위 500개의 키워드 추출 후 벤처창업과 유사한 순으로 30개의 키워드 빈도를 표로 나열하였다.

표 4. 트위터 키워드 빈도수

성공	준비	사람들	투자	지원
1764	1756	1440	1382	1127
청년	안철수	문재인	퇴사	사무실
1002	955	947	897	858
카페	취업	반려동물	가능	실패
771	767	715	624	612
걱정	경험	자격증	커피	동물
591	577	574	539	536
프리랜서 취업,창업	순수익	공무원	정보	매출
519	515	455	444	441
시험준비	혁신	비용	투 잡	일자리
397	350	347	255	251

[표 4]는 트위터에서 벤처창업 관련 트윗들을 수집하여 총 단어 199,491개 중 상위 30개의 키워드 빈도를 표로 나열하였다.

[표 1]에서 1, 2번째로 높은 빈도를 차지하는 '창업'은 836회, '기업'은 485회로 분석되어 창업기업에 관한 기사들이 많이 시사되었음을 엿볼 수 있었다. 또한 '스타트업'은 255회, '청년'은 257회로 비슷한 수치에 있는 것으로 보여 청년들의 스타트업이 벤처창업과 관련됨을 알 수 있다.

[표 1]의 3번째로 높은 빈도를 차지한 '지원'은 412회로 [표 2]와 [표 3]에서도 높은 빈도를 차지하였다. 이는 벤처창업을 지원해 주는 정부의 정책이나 사회적인 지원이 꾸준히 증대되는 것으로 보여 진다. [표 2]와 [표 3]에서는 '투자'와 '경제' 키워드가 존재하지만 [표 1]에서는 주요 키워드로 선별되지 않았다. 따라서 창업에 대해 3년 전과 5년 전의 데이터에서는 투자나 경제에 관한 관심이 많았지만 최근 1년의 데이터에서는 그에 관한 관심도가 많이 감소했음을 알 수 있다.

[표 3]에서 '실패'에 관한 빈도수가 294회로 창업의 실패에 관련된 기사들이 많은 것을 알 수 있다. 그러나 [표 2]나 [표 1]에서는 중요 키워드에 포함 되지 않은 것으로 보아 최근 데이터 키워드로 넘어 올수록 창업의 실패 요소나 실패에 관한 기사가 많이 줄어들었음을 알 수 있다.

[표 4]에서는 [표 1], [표 2], [표 3]에서의 데이터와는 다르게 단기간의 데이터이기 때문에 '안철수', '문재인' 등과 같이 특정 사건이나 인물이 중요 키워드 빈도수에 영향을 미치는 것을 알 수 있다. 또한 '성공'이 1,764회 '준비'가 1,756회로 창업의 성공을 위한 준비에 관심이 많음을 알 수 있고 핵심 키워드는 [표 1], [표 2], [표 3]과 크게 다르지 않다는 것을 알 수 있다.

그림 2. 전체 키워드

그림 3. 트위터 키워드

V. 결론 및 향후 연구

본 논문에서는 정보기술과 디지털 경제의 확산으로 대규모의 데이터가 생산되는 정보화시대에서 빅데이터의 중요성이 강조되고 있으며 다양한 분야에서 응용하고 있다. 빅데이터 분석도구 R을 이용하여 미디어 매체에 나타난 벤처창업 관련 빅데이터를 워드 클라우드 형태의 그림으로 나타내고 신문기사에서 나타는 키워드를 워드 클라우드로 시각화함으로써 빈도수에 따른 키워드를 쉽게 알아 볼 수 있었다. R프로그램을 이용함으로써 누구나 쉽게 접근하여 다양한 데이터를 워드 클라우드 형태의 그림으로 시각화 표현을 구현할 수 있다고 본다.

향후 연구 방향으로서 더욱 많은 빅데이터, 벤처창업 관련 데이터를 수집하여 앞으로 벤처창업을 하려는 준비되지 않은 예비 창업자들에게 시대와 상황에 맞는 벤처창업의 방향성을 제시하여 무분별한 벤처창업의 도전을 방지하고 벤처창업의 성공가능성을 높일 수 있을 것이다.

참고문헌

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

[2] 이경익, “빅데이터를 활용한 스마트 스타트업 콘텐츠 및 서비스 플랫폼 개발” 영화조세통람 2015년

[3] 장대성, “한국과 미국의 창업교육 비교를 통한 한국 벤처창업 교육전략 수립“ 한국컴퓨터정보학회논문지 제8권 제1호 통권 제25호 (2003. 3) pp.129-139 1598-849X KCI

[4] 신성윤, 김도관, “창업지원을 위한 공공기관 빅데이터 통합” 한국정보통신학회논문지 제19권 제6호 (2015년6월) pp.1341-1346 2234-4772