빅데이터 분석도구 R을 활용한 성경 데이터의 분석

김용수, 반재훈

고신대학교 인터넷비즈니스학과

Analysis of the Bible Data using Big Data Analytics Tools R

YongSu Kim · ChaeHoon Ban

Dept. of Internet Business, Kosin University

E-mail : ehyerisin@gmail.com · chban@kosin.ac.kr

요 약

빅 데이터가 정보통신기술 분야의 핵심 이슈로 부각되면서 관련 기술에 대한 관심이 증가하고 있다. 빅 데이터 분석 도구인 R은 통계 기반의 정보 분석을 가능하게 하는 언어와 환경이다. 본 논문에서는 이를 이용하여 성경데이터를 분석한다. 분석을 통해 신구약, 모세오경, 사복음서별로 어떠한 텍스트가 분포되어 있는지를 빈도 조사를 수행한다.

키워드

Big Data, R, Text Mining, Bible, Analysis

Ⅰ. 서론

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.

성경은 하나님의 말씀이고, 성경 전체는 유기적으로 연결되어 있으며, 총 66권의 책으로 구분되어 구약 39권, 신약 27권으로 나누어 볼 수 있다. 성경 전체에 분포되어 있는 텍스트와, 구약전서와 신약전서, 구약과 신약의 모세오경과 사복음서를 통하여 성경 데이터를 분석하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 빅 데이터 기법에 관련된 연구를 기술한다. 3장에서는 본 논문에서 구현한 워드 클라우드 형태의 그림을 표현하기 위해 R 프로그램 활용 방법을 설명한다. 4장에서는 워드 클라우드 형태의 그림으로 표현한 성경 데이터 분석에 대한 결과를 설명하고, 마지막 5장에서는 결론 및 향후 연구에 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 부산지역지인 국제신문과 부산일보의 기사들 중 제목에 '부산'과 '교통'을 동사에 포함한 기사의 기사 내용의 관계 또는 관련 있는 데이터에 내재되어 있는 의미 있는 패턴을 찾는 사회네트워크 분석을 실시하여 정형화된 빅 데이터를 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 장소는 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 분석도구인 R을 이용하여 텍스트 데이터를 워드 클라우드 형태의 그림으로 표현한다. 성경 데이터는 '컴퓨터전문인선교회(CTM)'의 성경타자통독에 있는 개역개정판을 기준으로 한 텍스트(txt)파일의 데이터를 수집했으며, 성경 데이터를 분석하기 위해 성경을 각각 성경의 전체부분과 구약 및 신약성경 그리고 구약의 모세오경과 신약의 사복음서로 총 다섯 부분으로 나누었다. 데이터의 분석과정은 그림1과 같다.

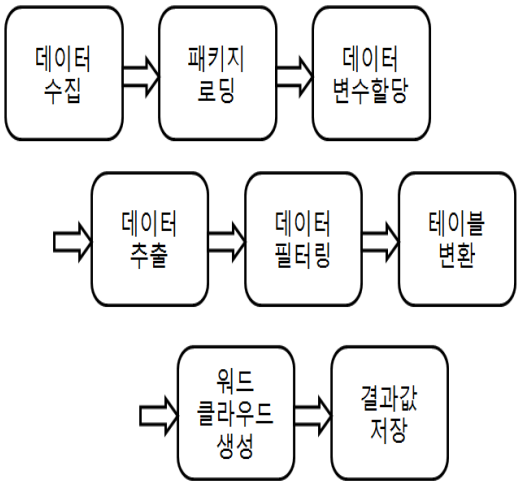

그림 1. 데이터 분석 과정

데이터 분석도구인 R을 설치하고 한글 데이터 분석에 필요한 패키지("KoNLP"), 워드 클라우드 생성에 필요한 패키지("wordcloud")를 설치하고 R 소스에 로딩한다. 성경 데이터를 성경전체, 구약성경, 신약성경, 모세오경, 사복음서의 그룹으로 구분하여 각 그룹의 성경 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. 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, 상위30위의 결과를 워드 클라우드 형태의 그래픽으로 출력한다. 출력 결과물을 이미지파일(JPGE, BMP, PNG 등)으로 저장한다.

IV. 성경 데이터 분석 결과

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

그림 2. 성경전체 키워드

표 1. 성경전체 키워드 빈도

여호와	사람	아들	하나님	이스라엘
5938	5118	2547	2453	1950
백성	말씀	자손	예수	마음
1839	1755	1712	1335	1121
이름	아버지	성읍	모세	제사장
1009	1007	867	863	771
명령	하나	사랑	왕	형제
728	718	641	625	588
성전	소리	구원	족속	조상
583	560	549	531	529
거룩	지혜	애굽	영광	머리
505	479	457	451	415

그림 2은 성경 전체 데이터에 대한 키워드 분석을 실시하였다. 키워드를 분석한 결과 '여호와' 빈도 수가 가장 많았고, 그 다음으로 '사람', '아들', '하나님', '이스라엘' 등으로 나타났다.

그림 3. 구약성경 키워드

표 2. 구약성경 키워드 빈도

여호와	사람	아들	이스라엘	백성
5938	3650	2285	1892	1687
자손	하나님	말씀	마음	성읍
1654	1587	1315	872	867

이름	모세	제사장	명령	왕
805	787	726	709	649
아버지	족속	하나	성전	조상
639	512	500	468	459
애굽	소리	제단	구원	거룩
447	422	391	388	387
지혜	형제	여호와여	기업	사랑
387	376	351	306	293

그림 3의 구약성경 분석 결과 '여호와'라는 키워드가 가장 많이 나타났고, 다음으로는 '사람', '아들', '이스라엘', '백성' 등으로 나타났다. 성경 전체 데이터의 키워드 분석 결과와 비교하여 보았을 때, 유사한 결과를 나타내는 것을 볼 수 있다.

그림 4. 신약성경 키워드

표 3. 신약성경 키워드 빈도

사람	예수	하나님	말씀	그리스도
1468	1335	866	442	400
아버지	사랑	제자	아들	마음
368	350	308	262	249
믿음	영광	하나	형제	세상
246	218	218	212	205
이름	성령	유대인	율법	구원
204	203	194	186	162
생각	백성	은혜	소리	기도
159	152	141	138	135
증언	선지자	복음	거룩	하나님을
129	125	123	122	121

그림 4의 신약성경 키워드 분석 결과는 '사람' 키워드가 가장 많이 나타나고, 그 다음으로 '예수', '하나님', '말씀', '그리스도' 등으로 나타났다. 구약성경과 신약성경은 예수님의 탄생을 기준으로 탄생 전인 구약에서는 '여호와', '아들' 그리고 구약에는 모세오경(창세기, 출애굽기, 레위기, 민수기, 신명기)이 포함되어 있어 이를 나타내는 키워드인 '이스라엘', '지혜', '모세', '성전' 등을 나타내고, 탄생 후인 신약성경에서는 예수님의 탄생 후인 '예수', '사람', '말씀' 등의 키워드가 나타나 있다. 구약은 옛 약속이라는 뜻에 아담, 아브라함 등 이스라엘의 조상들과 하나님이 맺은 약속을 말하며, 신약은 새로운 약속이라는 뜻에 예수님을 통하여 약속하신 것을 뜻한다.

그림 5. 모세오경 키워드

표 4. 모세오경 키워드 빈도

여호와	모세	사람	자손	아들
1714	667	642	606	599
하나님	이스라엘	백성	명령	아버지
495	483	396	312	285
제사장	말씀	애굽	이름	제단
281	275	252	211	195
숫양	종족	부정	조상	족속
166	161	156	148	146
거룩	성읍	하나	인도	속죄
144	140	140	136	135
형제	이삭	번제	거주	마음
133	132	130	123	122

그림 5는 구약성경의 첫 부분에 해당되는 창세기, 출애굽기, 레위기, 민수기, 신명기를 일컫는 모세오경의 키워드에 대한 결과로 '여호와'라는 키워드가 가장 많이 언급되었고, 그 다음으로 '모세', '사람', '자손', '아들' 등으로 나타났다.

그림 6. 사복음서 키워드

표 5. 사복음서 키워드 빈도

예수	사람	아버지	제자	말씀
1009	761	276	274	249
하나님	아들	하나	세상	귀신
209	150	138	103	102
사랑	이름	유대인	대제사장	주인
97	86	85	83	77
영광	마음	선지자	소리	성전
73	72	69	67	65

생각	형제	백성	기도	서기관
64	62	61	58	58
세례	어머니	증언	구원	비유
57	57	57	55	53

그림 6는 신약성경의 첫 부분에 해당하는 마태복음, 마가복음, 누가복음, 요한복음을 일컫는 사복음서의 키워드에 대한 결과로 '예수'의 키워드가 가장 많이 언급되고 있으며, 그 다음으로 '사람', '아버지', '제자', '말씀'의 순으로 키워드의 빈도가 나타났다. 이는 구약성경에 포함되어 있는 모세오경과, 신약성경에 포함되어 있는 사복음서는 각각 비슷한 분석 결과를 띄고 있다.

Ⅴ. 결론 및 향후 연구

본 논문에서는 정보통신기술의 발전과 소셜네트워크 서비스가 급속한 속도로 확산함으로 빅데이터라는 핵심 이슈를 나타나게 하였다. 빅 데이터 분석 도구인 R을 이용하여 성경 데이터에 접근하여 키워드를 분석하고 분포에 따라 워드 클라우드 형태의 그림으로 나타내어 성경의 전체 부분, 구약전서, 신약전서, 모세오경, 사복음서에서 나타는 키워드를 워드 클라우드로 시각화함으로써 빈도 수에 따른 키워드를 쉽게 알아볼 수 있었다. R 프로그램을 이용함으로써 누구나 쉽게 접근하여 다양한 데이터를 워드 클라우드, 다양한 그래프와 지도 데이터 등을 활용하여 시각화된 데이터를 구현할 수 있다고 본다.

향후 연구 방향으로서 성경을 세분화하고, 성경의 분석하여 배출되는 키워드를 중점으로 성경을 읽는 독자에게 주는 메시지가 무엇인지에 대하여 연구가 필요하고, 빅 데이터 분석을 통하여 가치와 의미가 있는 다양한 데이터를 활용하여, 다양한 분야의 정보를 얻을 수 있을 것이다.

참고문헌

[1] 김현근. R을 이용한 빅 데이터 사례 분석. 호서대학교 일반대학원 정보통계학과 석사학위 논문, 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 공간 메타포를 활용한 인문 융합 지도 연구. 한국인터넷방송통신학회. 15(3), 43-50