R을 이용한 성경 데이터의 빈도와 소셜 네트워크 분석

반재훈[1] · 하종수[2]

[1]고신대학교 IT경영학과 · [2]경남정보대학교 방송영상과

Frequency and Social Network Analysis of the Bible Data using Big Data Analytics Tools R

ChaeHoon Ban[1] · JongSoo Ha[2]

[1]Dept. of IT Management, Kosin University ·

[2]Dept. of Broadcasting & Image, Kyungnam College of Information & Technology

E-mail : chban@kosin.ac.kr / hajs@eagle.kit.ac.kr

요 약

데이터를 저장하고 분석하여 새로운 지식을 얻을 수 있는 빅데이터 처리기술은 사회의 여러 분야에서 중요성이 강조되고 있으며 정보통신기술 분야의 핵심 이슈로 부각되면서 관련 기술에 대한 관심이 증가하고 있다. 이러한 빅데이터를 분석할 수 있는 도구인 R은 통계 기반의 정보 분석을 가능하게 하는 언어와 환경이다. 본 논문에서는 이를 이용하여 성경데이터를 분석한다. R을 이용하여 어떠한 텍스트가 분포되어 있는지를 빈도 조사를 수행하며 소셜 네트워크 분석을 통해 성경을 분석한다.

ABSTRACT

Big datatics technology that can store and analyze data and obtain new knowledge has been adjusted for importance in many fields of the society. Big data is emerging as an important problem in the field of information and communication technology, but the mind of continuous technology is rising. R, a tool that can analyze big data, is a language and environment that enables information analysis of statistical bases. In this thesis, we use this to analyze the Bible data. R is used to investigate the frequency of what text is distributed and analyze the Bible through analysis of social network.

키워드

빅데이터, R, 텍스트 마이닝, 성경, 분석

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. 본 논문에서는 이전 연구를 바탕으로 신약전서의 4복음서를 데이터의 빈도와 소셜네트워크 그래프를 통하여 분석하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 빅데이터 기법에 관련된 연구를 기술한다. 3장에서는 본 논문에서 구현한 워드 클라우드와 소셜 네트워크 그래프로 시각화하기 위해 R 프로그램 활용 방법을 설명한다. 4장에서는 워드 클라우드와 소셜 네트워크 그래프로로 표현한 4복음서 분석에 대한 결과를 설명하고, 마지막 5장에서는 결론 및 향후 연구에 대해 기술한다.

Ⅱ. 관련 연구

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 사회네트워크분석을 실시하여 정형화된 빅 데이터를 시각화하고 해석했다[4].

III. 데이터 분석 방법

데이터 분석도구인 R을 이용하여 텍스트 데이터를 워드 클라우드 형태의 그림으로 표현한다. 성경 데이터는 '컴퓨터전문인선교회(CTM)'의 성경타자 통독에 있는 개역개정판을 기준으로 한 텍스트(txt) 파일의 데이터를 수집했으며, 성경 중에서 신약의 사복음서(마태, 마가, 누가, 요한)를 분석하였다. 먼저 단어의 빈도수를 분석하고 워드클라우드로 표현하였으며, 한 문장에서 단어간의 관계성을 분석하여 소셜네트워크 그래프를 그려 분석하였다. 데이터의 분석과정은 그림1과 같다.

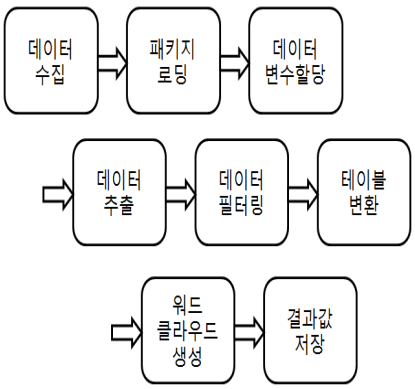

그림 1. 데이터 분석 과정

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 워드 클라우드와 소셜 네트워크 분석을 통해 성경 중에서 4복음서의 데이터를 분석하였다. 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.

워드 클라우드는 단순히 단어의 반복됨을 파악하는 것이기 때문에 전체적인 단어 간의 관계성을 알기 힘들다. 따라서 정확한 데이터의 분석을 위해서는 단어 간의 관계성을 분석하는 것이 필요하다. 본 논문에서는 단어 간의 관계성을 분석하기 위하여 한 문장에서 나오는 단어들을 쌍으로 표현하고 그 횟수를 분석하는 소셜 네트워크 분석을 실시하였다.

표 1. 마태복음에서의 단어와 단어 쌍의 빈도

단어의 빈도		단어 쌍의 빈도	
예수	278	('예수', '제자')	46
사람	197	('예수', '사람')	44
제자	88	('예수', '대답')	31
말씀	59	('예수', '무리')	28
아버지	59	('예수', '말씀')	25
하나님	57	('하늘', '아버지')	22
대답	56	('나무', '열매')	21
하늘	49	('예수', '하나님')	17
아들	47	('아들', '하나님')	16
무리	47	('아들', '아버지')	14

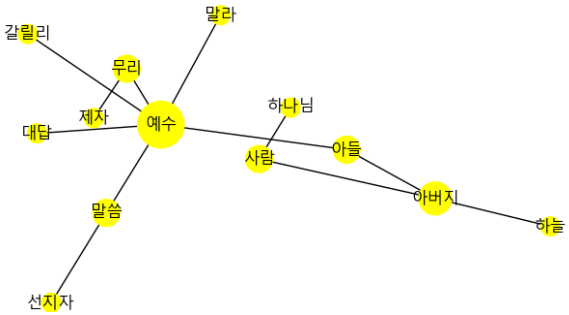

그림 2. 마태복음의 워드클라우드와 SNA 그래프

표 1은 4복음서중 가장 처음에 나오는 마태복음에서의 단어의 빈도와 한 문장에서 출현하는 단어쌍의 빈도를 나타낸다. 마태복음은 복음서에 비하여 매우 조직적이고 논리적이며 구약(舊約)과의 관계를 강조, 그 인용을 많이 하고 있다. 이 복음서의 특징은 예수를 구약 예언의 완성자이자 이스라엘의 왕, 즉 메시아(구세주)로 보고 유대인에 대한 예수의 사명을 강조한 점이다. 표와 같이 단어는 예수, 사람, 제자의 순으로 나타났으며 단어 쌍은 (예수, 제자), (예수, 사람)의 순으로 나타났다.

그림 2는 표 1의 결과를 시각화하여 표현한 것이다. 윗 그림은 워드클라우드 형태로 표현한 것이며 아랫 그림은 단어는 노드로 표현하고 단어와 단어의 관계는 엣지로 표현하였으며 그 관계의 빈도는 노드의 크기로 표현한 소셜 네트워크이다. 다른 복음서와는 달리 마태복음은 유대인들의 역사와 구약의 예언을 연결시켜 기술되었기 때문에 구약의 내용인 다윗, 선지자 등의 단어가 소셜 네트워크 분석에서 출현하였다.

표 2는 4복음서중 두 번째에 나오는 마가복음에서의 단어의 빈도와 한 문장에서 출현하는 단어쌍의 빈도를 나타낸다. 표와 같이 단어는 예수, 사람, 제자의 순으로 마태복음과 거의 유사한 순서로 나타났으며 단어 쌍은 (예수, 사람), (예수, 제자)의 순으로 나타났다.

표 2. 마가복음에서의 단어와 단어 쌍의 빈도

단어의 빈도		단어 쌍의 빈도	
예수	256	('예수', '사람')	60
사람	156	('예수', '제자')	44
제자	78	('무리', '예수')	33
하나님	54	('하나님', '예수')	26
말씀	44	('예수', '말씀')	23
무리	41	('제자', '사람')	22
귀신	36	('포도주', '부대')	16
요한	30	('하나님', '나라')	14
대답	23	('귀신', '예수')	14
대제사장	22	('사람', '하나님')	14

그림 3은 표 2의 결과를 시각화하여 표현한 것이다. 윗 그림은 워드클라우드 형태로 표현한 것이며 아랫 그림은 소셜 네트워크 그래프이다. 마가복음은 다른 복음서보다 내용이 적어 빈도가 적게 나타났으며 복음서의 특징이 예수탄생은 물론 그의 설교내용도 일체 쓰지 않고, 오직 '하느님의 아들'로서 예수의 업적만을 생생히 묘사한 데 있다. 따라서 예수-사람-하나님의 연관 관계가 다른 복음서보다 더 두드러지게 출현하였다.

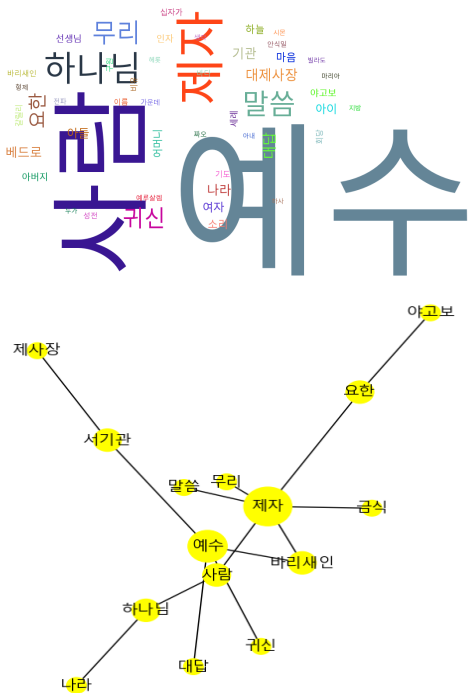

그림 3. 마가복음의 워드클라우드와 SNA 그래프

표 3은 4복음서중 세 번째에 나오는 누가복음에서의 단어의 빈도와 한 문장에서 출현하는 단어쌍의 빈도를 나타낸다. 표와 같이 단어는 예수, 사람, 하나님의 순으로 타 복음서와 거의 유사한 순서로 나타났으며 단어 쌍은 (예수, 사람), (사람, 하나님)의 순으로 나타났다.

표 3. 누가복음에서의 단어와 단어 쌍의 빈도

단어의 빈도		단어 쌍의 빈도	
예수	260	('예수', '사람')	69
사람	235	('사람', '하나님')	37
하나님	127	('하나님', '나라')	35
말씀	72	('예수', '무리')	31
아버지	53	('예수', '하나님')	30
제자	50	('예수', '말씀')	23
대답	47	('예수', '제자')	23
나라	47	('귀신', '사람')	22
무리	47	('아버지', '아들')	22
아들	41	('예수', '대답')	18

그림 4는 표 3의 결과를 시각화하여 표현한 것이다. 윗 그림은 워드클라우드 형태로 표현한 것이며 아랫 그림은 소셜 네트워크 그래프이다. 누가복음은 마태복음이 이스라엘 역사를 기초로 기술한 것과는 달리 이방인에 의한, 이방인을 위한 복음이자 가난한 자, 죄인, 약자에게 관심을 둔 사회적 복음이고 여성과 어린이의 복음으로, 인간의 개성과 그리스도의 인성(人性) 및 성령과 기도에 대하여

특별히 강조하고 있다. 이러한 단어들이 워드클라우드와 소셜네트워크 그래프에 출현하였다.

표 4는 4복음서중 마지막에 나오는 요한복음에서의 단어의 빈도와 한 문장에서 출현하는 단어 쌍의 빈도를 나타낸다. 표와 같이 단어는 예수, 사람, 아버지의 순으로 타 복음서와 거의 유사한 순서로 나타났으며 단어 쌍은 (예수, 사람), (제자, 예수)의 순으로 나타났다.

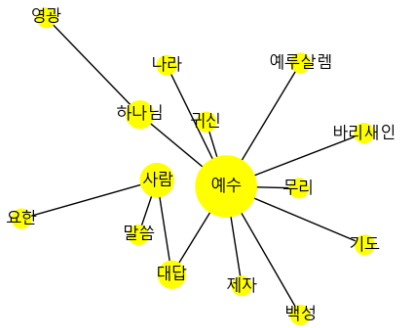

그림 4. 누가복음의 워드클라우드와 SNA 그래프

표 4. 요한복음에서의 단어와 단어 쌍의 빈도

단어의 빈도		단어 쌍의 빈도	
예수	309	('예수', '사람')	57
사람	190	('제자', '예수')	55
아버지	160	('예수', '대답')	49
하나님	84	('아버지', '세상')	37
제자	82	('아버지', '사랑')	32
말씀	79	('예수', '아버지')	31
세상	78	('유대인', '예수')	30
유대인	69	('예수', '말씀')	29
사랑	57	('사랑', '주님')	26
대답	55	('대답', '사람')	25

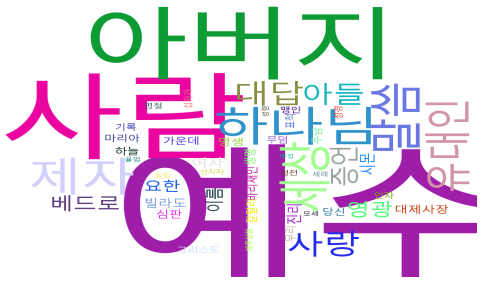

그림 5. 요한복음의 워드클라우드와 SNA 그래프

그림 5는 표 4의 결과를 시각화하여 표현한 것이다. 윗 그림은 워드클라우드 형태로 표현한 것이며 아랫 그림은 소셜 네트워크 그래프이다. 요한복음은 사랑의 교리를 유독 강조하여 '사랑의 복음서'라고 불리기도 한다. 따라서 다른 복음서와는 달리 사랑 등의 단어들이 워드클라우드와 소셜네트워크 그래프에 출현하였다.

Ⅴ. 결론 및 향후 연구

본 논문에서는 정보통신기술의 발전과 소셜네트워크 서비스가 급속한 속도로 확산함으로 빅 데이터라는 핵심 이슈를 나타나게 하였다. 빅 데이터 분석 도구인 R을 이용하여 성경 중에서 4복음서에서 나타는 단어들을 워드 클라우드와 소셜네트워크 그래프로 시각화하여 분석하였다.

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.

References

[1] H. Kim, "Big Data Case Study by Using R," M. S. theses, Hoseo University, Asan, Korea, 2014.

[2] Y. Oh, E. Park, "Data visualization of airquality data using R software," Journal of the Korea Data & Information Science Society, vol. 26, no. 2, pp. 399-408, 2015.

[3] C. Jang, J. Jang, S. Kim, H. Lee, C. Lee, "A study on the efficient patent search process using big data analysis tool R," Journal of Korea Safety Management & Science, vol. 15, no. 4, pp. 289-294, 2013.

[4] Y. Kim, C. Ban, "Analysis of the Bible Data using Big Data Analytics Tools R," in Proceeding of Korea Institute of Information and Communication Engineering 2015, pp. 349-352, 2015.