JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 24, No. 2: 166~171, Feb. 2020

빅데이터 분석도구 R을 이용한 성경 데이터의 빈도와 소셜 네트워크 분석

반재훈[1] · 하종수[2] · 김동현[3*]

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

ChaeHoon Ban[1] · JongSoo Ha[2] · Dong Hyun Kim[3*]

[1]Professor, Department of IT Management, Kosin University, Pusan, 49104 Korea
[2]Associate Professor, Department of Broadcasting & Image, Kyungnam College of Information & Technology, Pusan, 47011 Korea
[3*]Professor, Dept of Software, Dongseo University, Pusan, 46958 Korea

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. 성경 중에서 신약성경의 4복음서의 데이터를 분석한다. 먼저 성경데이터를 수집하고 분석을 위한 필터링을 수행한다. 이후 R을 이용하여 어떠한 텍스트가 분포되어 있는지를 빈도 조사를 수행하며 정확한 데이터의 분석을 위해 한 문장에서 나오는 단어들을 쌍으로 표현하고 단어 간의 관계성을 분석하는 소셜 네트워크 분석을 통해 성경을 분석한다.

ABSTRACT

Big data processing 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. the R, a tool that can analyze big data, is a language and environment that enables information analysis of statistical bases. In this paper, we use this to analyze the Bible data. We analyze the four Gospels of the New Testament in the Bible. We collect the Bible data and perform filtering for analysis. The R is used to investigate the frequency of what text is distributed and analyze the Bible through social network analysis, in which words from a sentence are paired and analyzed between words for accurate data analysis.

키워드 : 빅데이터, 성경, 소셜 네트워크 분석, 텍스트 마이닝, R

Keywords : Big Data, Bible, Social Network Analysis, Text Mining, R

Received 24 October 2019, **Revised** 6 November 2019, **Accepted** 6 November 2019

*** Corresponding Author** Dong Hyun Kim(E-mail:pusrover@dongseo.ac.kr, Tel:+82-51-320-1801)
Professor, Dept of Software, Dongseo University Pusan, 46958 Korea

Open Access http://doi.org/10.6109/jkiice.2020.24.2.166 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

IT 기술의 발전에 따라 실생활에서 발생하는 대규모의 비정형 데이터를 수집하고 수집된 데이터를 이용하여 미래를 예측할 수 있는 빅데이터 의 중요성이 강조되고 있으며, 다양한 산업에서 이를 활용하고 있다. 특히 빅데이터는 핵심 이슈로 부각되면서 중요성이 강조되고, 미래 경쟁력의 자원의 원천이 되며, 관련 기술의 발전, 자격증 등 다양한 분야에 활용됨으로 빅 데이터에 의미가 중요하다고 볼 수 있다. 이러한 빅 데이터를 분석할 수 있는 도구인 R은 통계 기반의 정보 분석을 가능하게 하는 언어와 환경이다. 본 논문에서는 이전 연구를 바탕으로 신약전서의 4복음서를 데이터의 빈도와 소셜 네트워크 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 빅데이터 분석 기술로는 데이터 마이닝, 텍스트 마이닝, 오피니언 마이닝, 웹 마이닝, 소셜 마이닝 등 다양한 기법을 통한 빅 데이터 분석연구가 있었다[1-3]. [4]에서는 정보통신의 발달과 소셜 미디어의 급속한 확산으로 생산된 빅 데이터를 분석하는 기법과 인프라 기술에 대해 기술하고 한글 텍스트 데이터를 R 프로그램을 이용하여 usesejongdic() 이라는 함수를 이용하여 명사만 추출하는 방법으로 비정형 데이터를 분석하였다.

[5]에서는 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 분석하였다.

[6]은 빅데이터 분석 도구인 R을 이용하여 빠른 시간 안에 사용자가 목적으로 하고 있는 특허검색 결과를 효율적으로 도출할 수 있는 검색어 추출에 관한 연구를 진행했다. [7]에서는 성경의 텍스트 데이터를 성경전체, 구약성경, 신약성경, 모세오경, 사복음서 데이터 분석결과를 각각의 워드 클라우드 형태 그림으로 표현하여 성경데이터를 분석하여 성경을 읽는 독자에게 주는 메시지가 무엇인지에 대한 연구를 제시하였다.

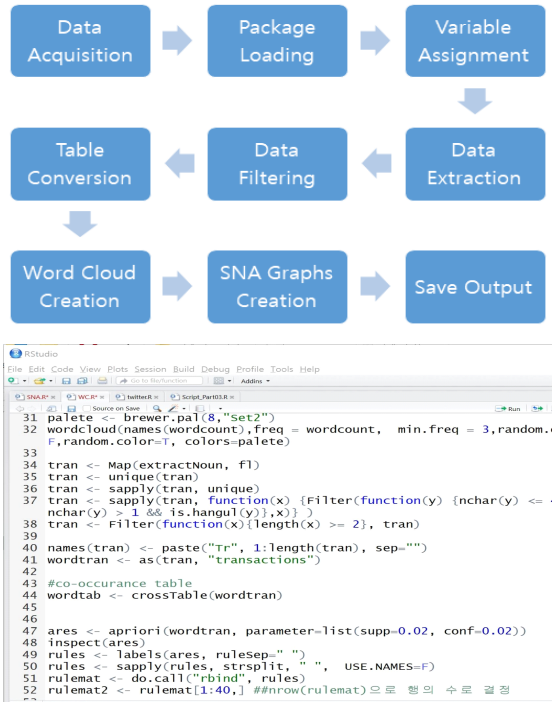
Ⅲ. 데이터 분석 방법

본 논문에서는 빅데이터 분석도구인 R을 이용하여 텍스트 데이터를 워드 클라우드 형태의 그림으로 표현하고 데이터 사이의 관계를 분석하여 소셜 네트워크 그래프를 생성한다. 이 논문에서 사용하는 성경 데이터는 ‘컴퓨터전문인선교회(CTM)’의 성경타자통독에 있는 개역개정판을 기준으로 한 텍스트(txt)파일의 데이터이며, 이중에서 신약의 사복음서인 마태, 마가, 누가, 요한복음을 분석하였다. 먼저 단어의 빈도수를 분석하고 워드클라우드로 표현하였으며, 한 문장에서 단어간의 관계성을 분석하여 소셜네트워크 그래프를 생성하였다. 데이터의 분석과정은 그림1과 같다.

한글 단어를 추출하기 위하여 KoNLP 패키지를 사용하였다. KoNLP는 한국어 텍스트 기반 연구를 위한 형태소 분석기 및 형태분석법을 제공하는 패키지이다.

이 패키지에서 제공하는 한글 명사를 추출 함수인 ‘extracNoun’ 함수를 사용하여 성경에서 명사를 추출하였다. 이후 원하지 않는 가비지 데이터에 대한 필터링을 수행하였다. 여기서는 2자리 이상의 명사만 추출하도록 프로그램을 구현하였으며 필터링 된 데이터를 텍스트 형식의 파일로 저장하여 테이블 형태로 변환하여 분석하였다.

상위30위의 결과를 워드 클라우드 형태의 그래픽으로 출력하였고 단어를 노드로 표현하고 단어와 단어의 관계를 에지로, 그 관계의 빈도를 노드의 크기로 표현한 소셜 네트워크 그래프를 생성하였다. 추후 연구에서는 보다 정확한 분석을 위하여 사용자 사전을 구축해 해당되는 단어를 추출 및 분석할 계획이다.




```
palete <- brewer.pal(8,"Set2")
wordcloud(names(wordcount),freq = wordcount,  min.freq = 3,random.
F,random.color=T, colors=palete)

tran <- Map(extractNoun, f1)
tran <- unique(tran)
tran <- sapply(tran, unique)
tran <- sapply(tran, function(x) {Filter(function(y) {nchar(y) <=
nchar(y) > 1 && is.hangul(y)},x)} )
tran <- Filter(function(x){length(x) >= 2}, tran)

names(tran) <- paste("Tr", 1:length(tran), sep="")
wordtran <- as(tran, "transactions")

#co-occurance table
wordtab <- crossTable(wordtran)


ares <- apriori(wordtran, parameter=list(supp=0.02, conf=0.02))
inspect(ares)
rules <- labels(ares, ruleSep=" ")
rules <- sapply(rules, strsplit, " ",  USE.NAMES=F)
rulemat <- do.call("rbind", rules)
rulemat2 <- rulemat[1:40,] ##nrow(rulemat)으로 행의 수로 결정
```

Fig. 1 Data Analysis Process and Programming Code as R

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 1은 4복음서중 가장 처음에 나오는 마태복음에서의 단어의 빈도와 한 문장에서 출현하는 단어 쌍의 빈도

Table. 1 Frequency of words and word pairs in Matthew

Frequency of words		Frequency of word pairs	
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

를 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.

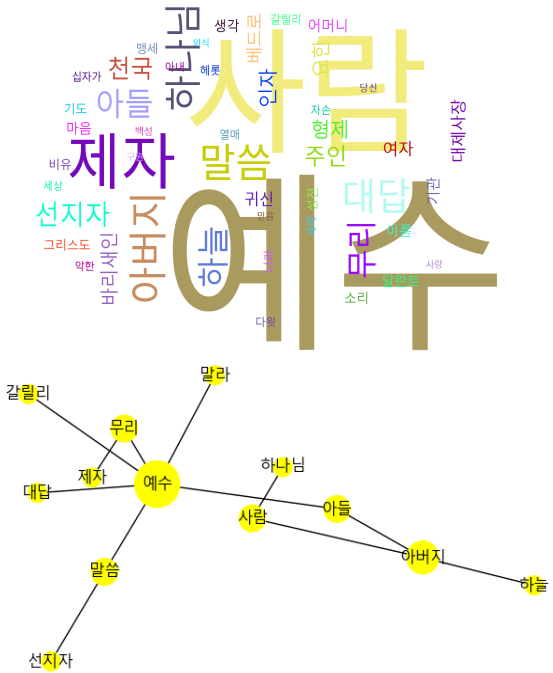


Fig. 2 Word cloud and SNA Graphs in Matthew

그림 2는 표 1의 결과를 시각화하여 표현한 것이다. 윗 그림은 워드클라우드 형태로 표현한 것이며 아랫 그림은 단어를 노드로 표현하고 단어와 단어의 관계를 에

지로 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 단어 쌍의 빈도를 나타낸다. 표와 같이 단어는 예수, 사람, 제자의 순으로 마태복음과 거의 유사한 순서로 나타났으며 단어 쌍은 (예수, 사람), (예수, 제자)의 순으로 나타났다.

Table. 2 Frequency of words and word pairs in Mark

Frequency of words		Frequency of word pairs	
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.

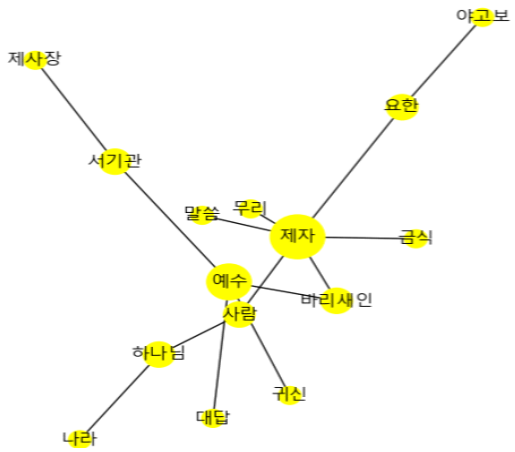


Fig. 3 Word cloud and SNA Graphs in Mark

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

Table. 3 Frequency of words and word pairs in Luke

Frequency of words		Frequency of word pairs	
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.

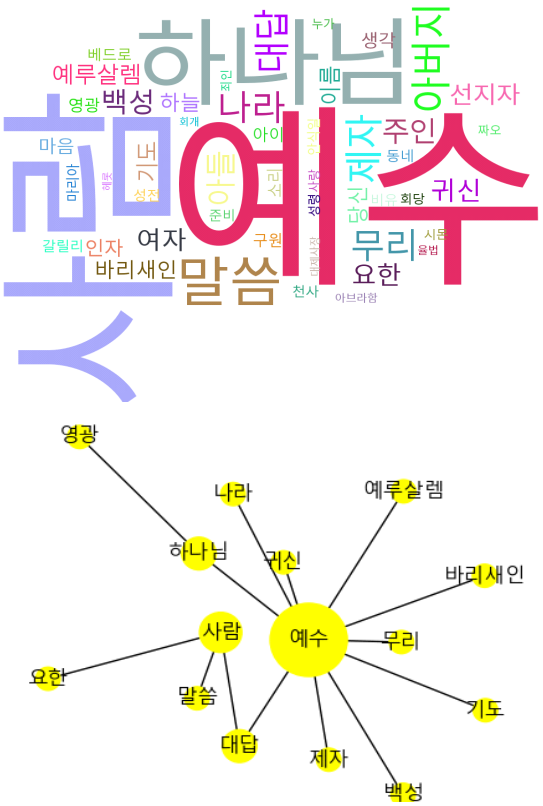


Fig. 4 Word cloud and SNA Graphs in Luke

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.

Table. 4 Frequency of words and word pairs in John

Frequency of words		Frequency of word pairs	
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

Fig. 5 Word cloud and SNA Graphs in John

그림 5는 표 4의 결과를 시각화하여 표현한 것이다. 윗 그림은 워드클라우드 형태로 표현한 것이며 아래 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. 결 론

본 논문에서는 R을 이용하여 성경데이터의 4복음서의 단어와 단어의 관계성을 분석하였다. 빅데이터 분석 도구 R을 이용하여 수집된 데이터를 분석하고 이를 워드클라우드 형태의 그림으로 나타내어 시각화함으로써 빈도 수에 따른 키워드를 쉽게 알아 볼 수 있도록 하였다. 또한 정확한 데이터의 분석을 위해 한 문장에서 나오는 단어들을 쌍으로 표현하고 그 횟수를 분석하는 소셜 네트워크 분석을 실시하였다.

4복음서를 분석한 결과 각 복음서에 출현하는 단어의 빈도와 단어 간의 관계성이 복음서 별로 가지고 있는 고유한 특징을 잘 반영하고 있음을 확인하였다. 향후 연

구 방향으로서 성경을 세분화하고, 성경의 분석하여 배출되는 키워드를 중점으로 성경을 읽는 독자에게 주는 메시지가 무엇인지에 대하여 연구가 필요하다.

REFERENCES

[1] C. Ban, Y. Lee, D. Ahn, and Y. Kwak, "The Venture Business Starts News and SNS Big Data Analytics," *in Proceeding of Korea Institute of Information and Communication Engineering 2017*, pp. 311-314, 2017.

[2] Y. Hwang, J. Park, I. Moon, K. Kim, and O. Kwan, "(The)Box-office Success Factors of Films Utilizing Big Data-Focus on Laugh and Tear of Film Factors," *Journal of Information and Communication Engineering*, vol. 20, no. 6, pp. 1087-1095, 2016.

[3] C. Ban, D. Kim, and J. Ha, "Analysis of University Department Name using the R," *Journal of Information and Communication Engineering*, vol. 22, no. 6, pp. 829-834, 2018.

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

[5] Y. Oh, and 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.

[6] C. Jang, J. Jang, S. Kim, H. Lee, and 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.

[7] Y. Kim, and 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.

반재훈(ChaeHoon Ban)

2006.2 부산대학교 공학박사
2008.9 ~ 고신대학교 IT경영학과 정교수
※관심분야 : 인터넷응용, 모바일, 빅데이터

하종수(JongSoo Ha)

2013.3 큐슈대학교 예술공학박사
2002.9 ~ 경남정보대학교 방송영상과 부교수
※관심분야 : 인터넷응용, 입체영상, 시지각

김동현(Dong Hyun Kim)

2003.2 부산대학교 공학박사
2004.3 ~ 동서대학교 소프트웨어학과 정교수
※관심분야 : 데이터베이스, 공간 데이터베이스, GIS, 센서데이터베이스