Journal of the Korea Industrial Information Systems Research Vol. 20 No.2, Apr. 2015 :113-124
http://dx.doi.org./10.9723/jksiis.2015.20.2.113 ISSN:1229-3741

Data Dictionary 기반의 R Programming을 통한 비정형 Text Mining Algorithm 연구

(A study on unstructured text mining algorithm through R programming based on data dictionary)

이 종 화[1], 이 현 규[2]*
(Jong Hwa Lee and Hyun-Kyu Lee)

요 약 미리 선언된 구조를 이용하여 수집·저장된 정형적 데이터와는 달리 웹 2.0의 시대에서 일반 사용자들이 평상시에 사용하는 자연어 형태로 작성된 비정형 데이터 분석은 과거보다 훨씬 더 넓은 응용범위를 가지고 있다. 데이터 양이 폭발적으로 증가하고 있다는 특성뿐 만 아니라 인간의 감성이 그대로 표현된 특성을 가진 텍스트에서 의미 있는 정보를 추출하는 빅데이터 분석 기법을 텍스트마이닝(Text Mining)이라 하며 본 연구는 이를 주제로 하고 있다. 본 연구를 위해 오픈 소스인 통계분석용 소프트웨어 R 프로그램을 이용하였으며, 비정형 텍스트 문서를 웹 환경에서 수집, 저장, 전처리, 분석 작업과 시각화(Frequency Analysis, Cluster Analysis, Word Cloud, Social Network Analysis)작업 등의 과정에 관한 알고리즘 구현을 연구하였다. 특히, 연구자의 연구 영역 분석에 초점을 더욱 높이기 위해 Data Dictionary를 참조한 키워드 추출 기법을 사용하였다. 실제 사례에 적용한 R은 다양한 OS 구동, 일반적 언어와의 인터페이스 지원 등 통계 분석용 소프트웨어로써 매우 유용하다는 점을 발견할 수 있었다.

핵심주제어 : R program, Big data, Text Mining, unstructured data

Abstract Unlike structured data which are gathered and saved in a predefined structure, unstructured text data which are mostly written in natural language have larger applications recently due to the emergence of web 2.0. Text mining is one of the most important big data analysis techniques that extracts meaningful information in the text because it has not only increased in the amount of text data but also human being's emotion is expressed directly. In this study, we used R program, an open source software for statistical analysis, and studied algorithm implementation to conduct analyses (such as Frequency Analysis, Cluster Analysis, Word Cloud, Social Network Analysis). Especially, to focus on our research scope, we used keyword extract method based on a Data Dictionary. By applying in real cases, we could find that R is very useful as a statistical analysis software working on variety of OS and with other languages interface.

Key Words : R program, Big data, Text Mining, unstructured data

* Corresponding Author
본 논문은 부경대학교 경영대학 간접연구경비 2015년도 우수논문 지원 사업으로 수행된 연구임.
Manuscript received January 20, 2015 / Revised March 18, 2015 / Accepted March 21, 2015
1) 부경대학교 일반대학원
2) 부경대학교 경영대학, 교신저자(hyunqlee@pknu.ac.kr)

1. 서 론

패턴을 읽을 수 있는 통찰과 천리안을 가지고 있다면 불확실성과 리스크를 줄이기 위한 대응을 할 수 있을 것이며 복잡한 대용량의 데이터라 하더라도 패턴을 읽는다면 데이터에 숨겨진 이야기를 이해할 수 있다.

빅데이터는 모든 종류의 정보이며 데이터 처리기술과 통신기술의 발달로 그 양이 폭발적으로 증가하고 있다. 네트워크 기술의 발달로 기존 데이터베이스 내의 정형적 데이터보다 문자나 사진, 동영상 같은 정형화되지 않은 정보들이 훨씬 많은 양을 차지하고 있다. 각종 센서, 보안카메라, 서버 log 파일 등 컴퓨터에 의하여 자동 생산되는 데이터나 트위터, 블로그, 사진, 게시판 글 등 실생활에 생산되는 데이터, 그리고 페이스북, 링크드인처럼 관계에 의한 데이터 등 모든 매체들의 네트워크가 진행되면서 광범위한 데이터의 수집이 가능하게 되었다[1,2]. 그런 정보에서 통찰과 가치를 얻을 수 있다면 사회나 기업 등 모든 영역에서 활용할 수 있을 것이다.

데이터는 세상에서 일어나는 일을 이해할 수 있기 때문에 강력한 도구로 등장하고 있다. 실시간 세계 통계 자료를 제공하는 Worldometers에 따르면 전 세계 인구가 약 72억 7천명에 이르고, 매일 발행된 신문은 21억부, 발송된 이메일 902억건, 발생한 트윗 2억9천 건으로 집계를 하고 있다. 데이터 자체는 어떤 의미를 지니지 않을 수 있지만 데이터에서 패턴을 찾아낼 때 사회현상에 관한 정보를 추출해 낼 수 있을 뿐만 아니라 사물이나 현상에 대한 새로운 시각이나 법칙을 발견할 가능성을 매우 높여 준다[3].

빅데이터 가운데 숨겨져 있는 유용한 정보 즉, 자료들의 상관관계를 발견하여, 통계 분석과 유용한 패턴을 찾아내는 방법을 데이터 마이닝이라 한다. 정형화된 데이터를 대상으로 처리하는 데이터 마이닝은 서로의 연관성을 탐색하거나 패턴 인식 등 다양한 알고리즘들이 개발되고 있다. 분류, 군집화, 연관성, 연속성, 예측 등 다양한 분야에 적용하고 다가오는 미래를 현실로 연결하여 의사 결정에 이용되고 이는 데이터베이스 마케팅의 핵심기술이다[1,4].

미리 선언된 구조를 이용하여 수집·저장된 정형적 데이터와는 달리 구조화되지 않은 대규모 텍스트 집합으로부터 새로운 지식을 발견하는 과정을 텍스트 마이닝이라 한다. 다시 말해, 텍스트마이닝은 비정형화된 텍스트로부터 키워드 분류, 나열, 요약 등을 통하여 새로운 패턴을 찾아내는 과정이다. 자료 수집, 키워드 추출, 자료 검색, 자연어 처리, 자료 분류 및 요약 등에서 사용되는 기법들을 결합하여 사용한다[5,6]. 많은 양의 자료로부터 알려지지 않은 숨겨진 정보나 지식을 찾아내는 것이 텍스트마이닝(Text Mining)의 핵심 기법이며 본 연구는 이를 주제로 하고 있다.

R 프로그래밍 언어(이하 R)는 통계 계산과 분석도구이며 그래픽 표현을 위한 프로그래밍 환경이다. R은 통계 소프트웨어 개발과 자료 분석에 널리 사용되고 있으며, 패키지 개발이 용이하여 통계학자들 사이에서 통계 소프트웨어 개발에 많이 사용되고 있다. R은 사용자가 제작한 패키지를 추가하여 기능을 확장할 수 있다. 핵심적인 패키지는 R과 함께 설치되며, CRAN(the Comprehensive R Archive Network)을 통해 6,000개 이상의 패키지를 내려 받을 수 있다. R의 또 다른 강점은 시뮬레이션과 그래픽스를 제공한다. R은 윈도, 맥 OS 및 리눅스를 포함한 UNIX 플랫폼에서 이용 가능하다[4,7].

비정형 텍스트 문서를 웹 환경에서 수집, 저장, 전처리과정을 통하여 Frequency Analysis, Cluster Analysis, Word Cloud, Social Network Analysis 하는 등의 과정에 관한 텍스트마이닝 분석을 오픈 소스의 통계 분석용 소프트웨어 R을 통하여 알고리즘 구현을 연구하였다.

빅데이터는 우리가 살고 있는 많은 것들을 데이터화시킨 결과물이다. 데이터의 양도 크지만 다양한 형태의 데이터들을 분석하여, 고객들의 Needs를 정확하게 먼저 파악하는 것이 그 가치를 더할 것이다. R은 텍스트마이닝 분석에 필요한 적합한 소프트웨어라 판단한다.

2. 선행연구

웹 2.0의 시대에서 일반 사용자들이 평상시에 사용하는 자연어로 작성된 비정형 데이터의 분석은 과거보다 훨씬 더 넓은 응용범위를 가지고 있다. 데이터 양이 폭발적으로 증가하고 있다는 특성 뿐만 아니라 인간의 감성이 그대로 표현된 특성을 가진 텍스트에

서 의미 있는 정보를 추출하기 위하여 텍스트마이닝을 연구하였다.

문서, 신문기사, 논문, 문헌(eBook포함), SNS 게시글 등 ICT에서 표현할 수 있는 모든 텍스트 형태들을 수집한 후 텍스트마이닝 분석을 통해 분류한 결과를 시각화하여 빠른 정보를 전달하는데 목적을 두고 있다. 사실과 정보를 전달해 주는 뉴스를 분석하여 현 시점의 주요 이슈를 객관적 입장에서 빠르게 파악하는 것이 유용하다.

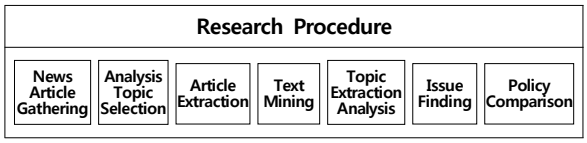

Fig. 1 Research Procedure [5].

원진용[5]의 텍스트마이닝 연구절차는 텍스트 형태의 뉴스 전체를 수집하고, 분석 주제를 선정하고 추출하여 토픽 분석, 가중치 계산 등의 텍스트마이닝을 연구하였고 사회 위험 관점에서 발생한 이슈의 데이터 마이닝 결과와 뉴스의 원문을 통하여 도출된 정책을 비교 검토하였다.<Fig. 1>참고 [5]

텍스트마이닝을 R로 구현하는 과정은 tm 패키지를 이용하여 문서별 Corpus(텍스트 묶음) 생성, 명사 추출, 불용어 및 기타 의미 없는 기호를 제거, 2자리 이상의 명사 추출, KoNLP 패키지를 사용하여 Corpus 내의 한글 형태소 단위를 인식한다. 그리고 명사들의 빈도를 이용하여 Matrix 구조의 유사도를 판단, 구현하고 유사한 문서의 그룹화를 거쳐서 키워드와 복합 명사들을 추출하였다[8].

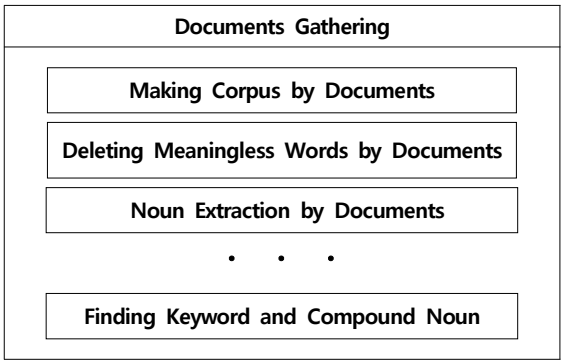

Fig. 2 Analysis process of the primary patent search results [8].

일반 텍스트마이닝 분석은 <Fig. 2>에서와 같이 비교 주체인 텍스트 데이터에서 R의 제공 KoNLP 패키지의 함수를 이용하여 단어를 추출하고 명사를 추출(extractNoun())하며 단어를 인식시키는 연구가 이루어지고 있다.

extractNoun()함수는 한글의 단어, 명사 추출에 매우 유용한 함수이며 수행 시간 또한 매우 짧은 시간에 단어들의 추출이 처리되어 대부분의 연구자들이 그들의 연구에서 보편적으로 많이 사용하는 것으로 나타났다. KoNLP 패키지의 편리성과 사용 용이함으로 인해 보편적인 추출 방법으로 사용하고 있는 것이다. 하지만 비교 주체인 텍스트 데이터의 단어들이 연구자가 원하는 단어들로만 이루어져 있는 것은 아니다. 또한, KoNLP 패키지의 역할이 연구 방향의 단어만을 선별할 수 있는 것은 아니다 라는 것도 확인하였다[8,9].

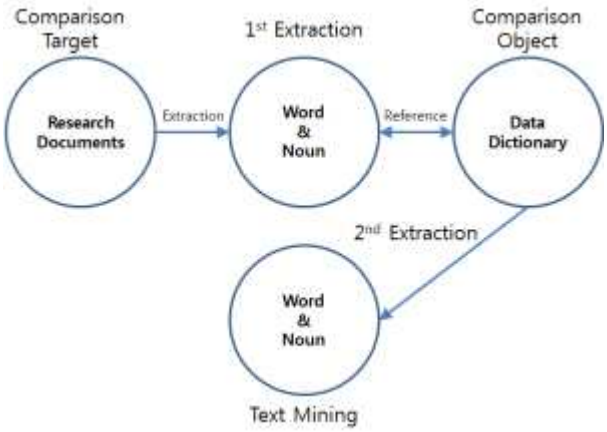

Fig. 3 Extraction of Words and Nouns Based on the Data Dictionary

정리하면, 비교 주체를 연구 문서로, 비교 대상은 비교 주체를 평가하는 기준으로 Data Dictionary를 적용하였다. 비교 주체인 연구 문서내의 단어들이 연구자의 연구 분야의 단어들로만 이루어져 있지 않기 때문에 비교 주체에서 단어 추출을 다시 비교 대상 즉, Data Dictionary를 만들어 비교 주체와의 텍스트마이닝을 연구하였다. 가령, 정보기술의 관련 연구에 비교 대상을 선정한다면 정보 통신 용어 관련 Data Dictionary를 만들어 해당 문서와의 텍스트마이닝을

구현함으로써 보다 현실성 있는 키워드 빈도 분석과 결과에 대한 높은 신뢰를 보일 것이다. 연구 방향이 마케팅 용어에 대한 텍스트마이닝을 통하여 알려지지 않은 흥미 있고 유용한 패턴을 찾는다 가정하자. 비교 주체인 연구 문서에서 1차 추출한 단어와 명사를 마케팅 용어 Data Dictionary와 참조하여 2차 추출을 진행하면서 연구자는 마케팅 영역의 분석에 더욱 초점을 두게 될 것이다.<Fig. 3 참조>

또한, 연구 문서와 Data Dictionary의 비교 분석 과정에 R을 통한 연구가 미진하여 <Fig. 4>와 같은 분석 모형을 제시한다.

일반적 절차는 데이터 수집, 데이터 처리, 키워드 추출, 정보 분석으로 구분되고, 키워드 추출 단계에서는 통계적, 수학적 모델을 이용하여 유용한 정보를 추출한다.

<Fig. 4> 분석 흐름도를 보면 Data Dictionary의 처리를 통하여 연구자의 연구 방향에 관련된 키워드들을 비교 분석할 수 있도록 리스트 생성 과정과 연구 문서와 Data Dictionary의 비교 과정을 삽입한 것이다. Data Dictionary와의 비교를 통해 연구 분야의 보다 정확한 분류를 위한 수단으로 사용한 것이다. 또한, 연구자가 의도하지 않은 단어들은 연구자가 관심 있는 키워드의 상관관계에 영향을 미칠 수 있다고 판단되어 <Fig. 4> 기준의 분석 흐름을 실시하기로 한다.

3. 연구 방법

3.1 연구 목적 및 연구 방법 개요

본 연구에서는 다양한 분야에서 사용되는 빅데이터 분석 기법인 텍스트마이닝(Text Mining) 기법을 R을 통하여 구현하였고 정보 추출, 통계 분석, 시각화를 활용한 알고리즘 구현 방법을 제시하고, 임의

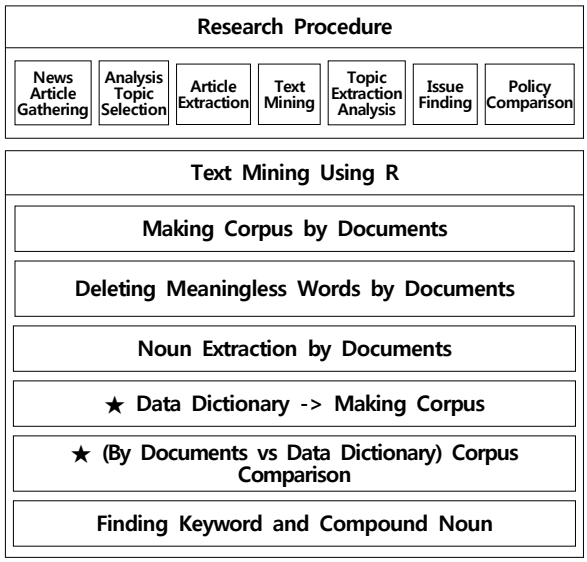

Fig. 4 Flow Chart of This Research

수집된 Sample 자료에 연구한 알고리즘을 적용해 본다.

우선 R의 구성과 진화 과정을 탐색하여 다양한 함수 및 Packages들을 확인할 수 있었다. R Script 기능을 통한 사용자 정의 함수(function)를 연구하여 반복적인 처리 과정을 매크로화 하고, 비교 및 처리 과정을 최적화된 알고리즘으로 구현하였다. 자료의 Import(Corpus)과정, 전처리(Transformations)과정, 처리 및 분석, 시각화 과정을 구현하였다.

3.2 데이터 Import(Corpus) 알고리즘

Corpus는 분석을 수행하는데 필요한 텍스트의 수

```
> news.docs <- Corpus(DirSource(directory, encoding="UTF-8"))
> summary(news.docs)
```

```
# 작업 폴더에서 데이터를 읽고 Corpus 명령을 이용하여 텍스트를 컴퓨터가 읽을 수 있는 형태로 저장하는 과정이다. ASCII 인코딩은 UTF-8의 부분 집합이다. 일반적인 ASCII 문자열은 올바른 UTF-8문자열이며, 하위 호환성이 보장된다.
# 판독된 데이터를 확인한다.
```

```
> for (j in seq(news.docs))
+   {
+     news.docs[[j]] <- gsub("[/@().,:]", " ", news.docs[[j]])
+     news.docs[[j]] <- gsub("[&]", "nnn", news.docs[[j]])
+     news.docs[[j]] <- gsub("[#]", "sss", news.docs[[j]])
+   }
> news.docs <- tm_map(news.docs, tolower)
> news.docs <- tm_map(news.docs, removeWords, stopwords("english"))
> news.docs <- tm_map(news.docs, stripWhitespace)
```

```
# gsub함수는 특정 문자열을 검색하여 지정된 문자로 바꾸는 함수로 [/@().,:], [&], [#] 등의 특수문자는 특정문자로 치환하는 과정이다. 또한 원하지 않는 내용은 필터링한다.
# news.docs 문서들의 글자들을 모두 소문자로 바꾸고 문장 부호 제거 등 불용어 처리, 문서 중간의 공백을 제거한다.
```

집, 저장을 담당한다. 문서 집합 전체에 대해 적용되는 메타데이터이며, 각 문서는 TextDocument로 표현된다. 지원되는 형식은 텍스트, PDF, 마이크로소프트 워드, 및 XML 등 을 포함한다[7,10,11,12,13,14].

3.3 데이터 전처리(tm_map, gsub) 알고리즘

문서 분류에 앞서 글에서 문장 부호를 제거하거나, 문자를 모두 소문자로 바꾸거나, Stemming을 적용하는 등 문서를 변환하는 과정을 구현하였다. 이 때 사용하는 함수가 tm_map ()이다.

tm패키지는 문서 변환을 위한 다양한 함수를 제공하며 함수들의 목록은 getTransformations()로 볼 수 있다. 다음은 news.docs 문서들의 글자들을 특정 단어로 치환하며 모두 소문자로 변환하고 문장 부호를 제거하는 예이다[7,10,11,12,13,14, 15,16].

3.4 데이터 처리 알고리즘

앞서 많은 양의 데이터 수집 과정과 불필요한 문자들을 제거하는 전처리과정을 구현하였다. 많은 연구자들은 연구 대상 문서에서 명사를 추출하는 처리 방식으로 R의 extractNoun()함수를 사용하며, JAVA 언어는 Lucene Korean을 활용하여 처리하고 있음을

```
>   for (j in seq(news.list)) {
+     for (i in seq(dic_ref)) {
+       is.kw <- str_count(news.docs[[j]][4], dic_ref[i])
+       if (is.kw) {
+         news.docs[[j]][4] <- gsub(dic_ref[i],"", news.docs[[j]][4])
+         id <- (id + 1)
+         tmp.date <- substr(news.list[j],1,6)
+         temp1 <- data.frame("ID" = id, "News.files" = as.character(news.list[j]), "News.ID" = j, "Date"
                  = tmp.date, "Keywords" = dic_ref[i], "Occurrences" = is.kw, stringsAsFactors =
                  FALSE)
+         outdata <- rbind(outdata,temp1)
+       }
+     }
+   }
```

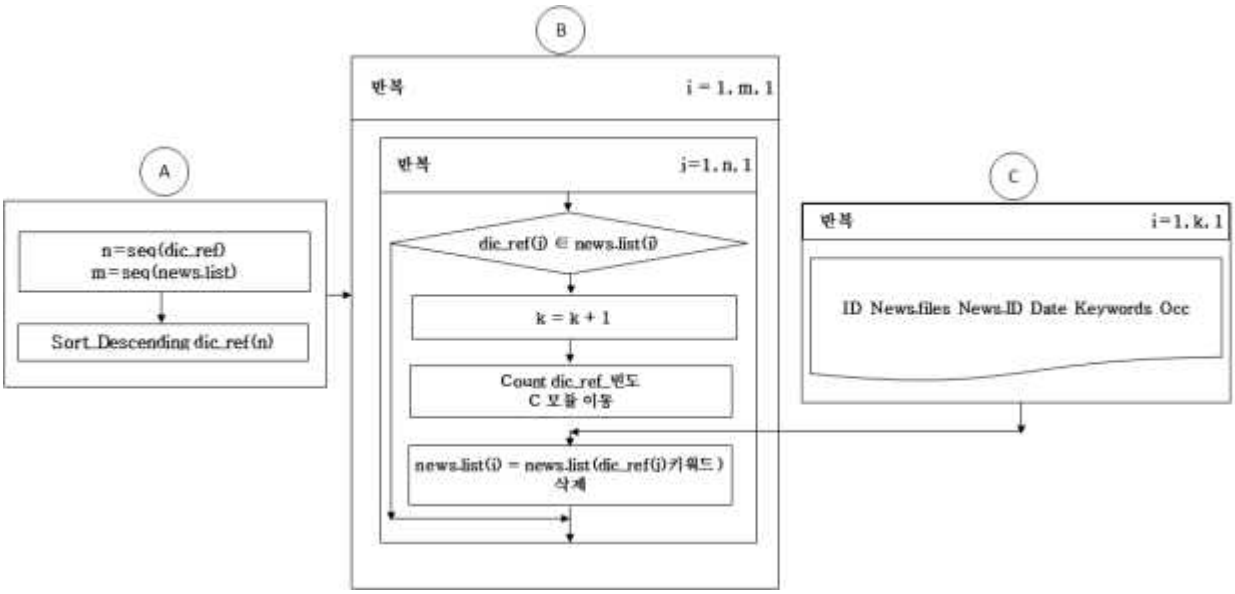

Fig. 5 Data Manipulation Algorithm

선행 연구에서 찾을 수 있었다[17]. 본 연구자는 연구 대상 문서에서 단어를 추출하는 방식이 아닌 비교할 키워드들을 Data Dictionary를 작성하여 연구 대상 문서를 참조하여 키워드 빈도를 도출하였다. 본 연구자는 비정형 텍스트인 연구 문서를 Data Dictionary와 대조하여 키워드 단순 빈도를 수집함으로서 정형 데이터로 변환하였다[18].

<Fig. 5>의 A모듈 처리 과정은 Data Dictionary의 개수(n), 연구 문서의 개수(m)를 선언하고 Data Dictionary는 내림차순으로 정렬하여야 한다. 단어의 길이가 긴 순서대로 먼저 처리하기 위해서이다. B모듈 처리 과정은 다중 반복문으로 { 하나의 연구 문서내의 키워드(news.list(i))들을 모든 Data Dictionary (dic_ref(j))에 비교하여 추출한 키워드는 카운트하며 연구 문서내의 키워드에서 제거한다. } { ... }은 연구 문서의 개수만큼 반복되며 처리된다. 연구 문서내의 키워드를 제거하는 이유는 가령, social media, social, media 등이 같은 문서의 키워드라면 복합 단어인 Social media 키워드를 추출 후 연구 문서에서 제거한 다음 social과 media를 검색하여 복합 단어가 이중 카운트 되는 것을 제어할 수 있기 때문이다. C모듈 처리 과정은 같은 키워드를 찾았을 경우 data.frame에 추가하는 모듈이다.

3.5 데이터 Frequencies(Barplot) analysis 알고리즘

서로 관계가 있는 단어들의 빈도를 이용하여 상대

```
> termFrequency <- rowSums(as.matrix(myTDM))
> termFrequency <- subset(termFrequency, termFrequency >=1000)
> png("barplt.png", width = 960, height = 960, units = "px", pointsize = 12, bg = "white", res = 110, family =
      "", restoreConsole = TRUE)
> barplot(termFrequency, las = 2))
> dev.off()
```

```
# rowSums()함수는 matrix에서 각 행의 총합을 구한다.
# subset()함수는 빅데이터(전수조사)를 기반으로 키워드 분석 결과 1000 이상의 빈도수만을 대상으로 제한
  하고 있다.
# 이미지 파일은 PNG형식으로 저장한다.(그 외에도 BMP, JPEG, TIFF 등 지원한다.)
# barplot()함수는 카테고리의 빈도가 들어간 벡터로 막대그래프를 구현한다.
```

```
>  myTDM2 <- removeSparseTerms(myTDM, 0.97)
>  m <- as.matrix(myTDM2)
>  wordFreq <- sort(rowSums(m),decreasing = TRUE)
>  set.seed(375)
>  palete <- brewer.pal(9,"Set1")
> png("wordcloud.png", width = 1920, height = 1920, units = "px", pointsize = 17, bg = "white", res =
      250, family = "", restoreConsole = TRUE)
> wordcloud(word = names(wordFreq), freq = wordFreq, scale = c(4,.2), max.words = 500, random.order
      = FALSE, random.color = T, colors = palete)
> dev.off()
# removeSparseTerms()함수를 이용하여 빈도가 낮은 키워드들은 제거한다.
# as.matrix()함수는 matrix구조로 변환한다.
# rowSums()함수는 matrix에서 각 행의 총합을 구한다.
# set.seed()함수는 랜덤하지만 일정한 규칙이 있는 수를 얻는다.
# wordcloud(출력할 단어들, 빈도수, 글자 크기, 최대 빈도 횟수, 순서 랜덤, 색상 랜덤, 단어들의 색상)
```

값을 나타낸 그래프(Graph)들을 다양하게 표현할 수 있었다. pie()함수를 사용한 원형 그래프, barplot()함수를 사용한 가로, 세로 Bar 그래프, pie3D()함수를 사용한 입체형 원 그래프, abline()함수를 이용한 Line 그래프, ggplot()함수를 이용한 구간 데이터 분석 표현이 쉬운 Boxplot 등 비정형 데이터를 분석해서 다양한 그래프로 표현하는 방법을 실험하였다[10,11,12,13,14,19,20].

3.6 데이터 Word cloud analysis 알고리즘

데이터 분석 결과를 쉽게 이해할 수 있도록 도표라는 시각적 수단을 통해 정보를 효과적으로 전달하는 것이 데이터 시각화이다. 정보 디자인 관점에서 수많은 데이터를 한 장의 그림으로 표현한 인포그래픽스 방식으로 문서에 사용된 단어의 빈도와 중요도를 시각적으로 표현한 것을 Word Cloud라 한다. wordcloud패키지를 사용하여 글자 크기, 언급 횟수 제한, 단어 배치 각도, 색상, 순서 등 다양한 형태의 변형이 가능한 것으로 나타났다[21].

3.7 데이터 Cluster analysis 알고리즘

다양한 특성을 지닌 대상들을 유사성에 기초하여 동질적인 집단으로 분류하는데 사용되는 기법으로 자료를 보다 간단하게 축약하는데 많이 사용되고 있는 분석이다. R에서도 데이터가 저장된 행렬의 각 행렬 간의 거리를 구하는 dist()함수와 수형도 주위에 테두리를 그리는 rec.hclust()함수와 함께 Cluster 함수인 hclust()함수를 사용한다[10,11,12,13,14].

3.8 데이터 Social Network analysis 알고리즘

```
> myTDM2 <- removeSparseTerms(myTDM, sparse=1000)
> m2 <- as.matrix(myTDM2)
> distMatrix <- dist(scale(m2))
> fit <- hclust(distMatrix, method = "ward.D")
> plot(fit, cex = 0.6)
# removeSparseTerms()함수를 이용하여 빈도가 낮은 키워드들은 제거한다.
# as.matrix()함수는 matrix구조로 변환한다.
# scale함수는 matrix로 표준화되며 dist()함수는 데이터가 저장된 행렬의 각 행렬 간의 거리(간격)를 구하는
  함수이다.
# 글러스터분석은 hclust()를 사용한다.
# 산포도나 꺾은선 그래프를 표현할 수 있는 plot(데이터, 문자를 그릴 때 문자의 크기를 지정)함수이다.
```

수학의 그래프 이론에 따라 연결 구조와 연결 강도 등을 바탕으로 사람과 사람간의 관계를 연구하는 소셜 네트워크 분석을 키워드 간의 사용 빈도 분석을 통하여 서로의 영향력을 측정하여 시각화해 보았다[10,11,12,13,14,19,20].

```
> myTDM2 <- removeSparseTerms(myTDM, spr)
> m <- as.matrix(myTDM2)
> TDM.logic <- m
> TDM.logic[TDM.logic >=1] <- 1
> TermMatrix <- TDM.logic %*% t(TDM.logic)
> g <- graph.adjacency(TermMatrix, weighted = T, mode = "undirected")
> g <- simplify(g)
> V(g)$label <- V(g)$name
> V(g)$degree <- degree(g)
> set.seed(201409)
> layout1 <- layout.fruchterman.reingold(g)
> V(g)$label.cex <- 1.1 * V(g)$degree/max(V(g)$degree)+ .1
> V(g)$label.color <- rgb(0, 0,.2,.8)
> V(g)$frame.color <- NA
> egam <- (log(E(g)$weight)+.4) / max(log(E(g)$weight)+.4)
> E(g)$color <- rgb(.5,.5, 0, egam)
> E(g)$width <- egam
> png(paste("./output/", type, "/visual/sna-", gsub("[:]", "", Sys.time()), ".png", sep=""), width = 960,
      height = 960, units = "px", pointsize = 15, bg = "white", res = 110, family = "", restoreConsole =
      TRUE)
> plot(g, layout = layout1, edge.width = E(g)$weight/500, vertex.size = log(degree(g2))*25 /
      log(max(degree(g2))))
> dev.off()
```

```
# removeSparseTerms()함수를 이용하여 빈도가 낮은 키워드들은 제거한다.
# as.matrix()함수는 matrix구조로 변환한다.
# TDM.logic[TDM.logic >=1] <- 1
# t()함수는 전치행렬을 구현하며, matrix에 가로 첫 행과 세로 첫 열에 같은 키워드를 배치한다.
# graph.adjacency() matrix에서 무방향 그래프를 생성한다.
# simplify() 루프를 제거하여 간소화 시킨다.
# V(g)$label <- V(g)$name 높은 빈도의 키워드를 산출한다.
# V(g)$degree <- degree(g) 키워드의 빈도를 산출한다.
# set.seed()함수는 랜덤하지만 일정한 규칙이 있는 수를 얻는다.
# layout1 <- 레이아웃 세팅작업을 수행한다.
# V(g)$label.cex구문은 데이터의 글자 크기를 설정한다.
# V(g)$label.color구문은 데이터의 글자 색상을 설정한다.
# V(g)$frame.color <- NA frame 색상을 제거한다.
# E(g)$color구문은 연결선(edge) 색상을 설정한다.
# E(g)$width <- egam
# 이미지 파일을 PNG형식으로 저장한다.
# plot의 vertex.size 옵션을 이용하여 노드의 가중치를 조절한다.
```

4. 연구 알고리즘 실험과 결과

4.1 연구 알고리즘 실험 개요

본 실험에 사용하는 데이터로는 정보기술 관련 일간 신문인 전자신문(http://www.etnews.com)의 기사를 이용하였다. 2013년 한 해의 모바일 통신 환경과 관련된 기사를 발취하여, 전처리 과정을 통하여 획득한 단어들을 실험에 모두 이용하였다.

많은 연구자들이 이용하는 연구 문서 내 키워드 추출 방식이 아닌 IT용어리스트(TSS사전의 키워드와 Dic_Lee2013 사전의 키워드를 결합)를 생성하여 실험 데이터와 대조하는 방식으로 실험을 하였다.

4.2 Data Collection

전자 신문의 2013년 01월부터 12월까지 기재된 모든 신문 기사를 서버 컴퓨터가 검색하여 해당 주제어가 제목에 수록되어 있는 기사들을 저장하게 된다. 전자신문의 기사 링크 즉, URL 패턴을 분석하여 규칙성(http://www.etnews.com/201306
20xxxx)을 확인하였고 PHP 언어를 통하여 검색할 신문기사 URL을 생성하고 서버 컴퓨터가 임의 접속을 시도한다. 가상으로 만든 URL에 접속하여 기사가 존재하면 "모바일", "스마트", "휴대폰", "이동", "통신"의 키워드를 기사의 제목과 비교하여 하나 이상의 키워드가 발견되면 제목, 부제목, 내용, 발간일 등을 저장하는 알고리즘 구현, 자료 수집을 실시하였다. 각 뉴스 기사는 제목과 본문 외에 이미지나 영상 등을 포함하고 있으나, 본 연구의 목적에 적합하지 않은 데이터이기 때문에 제거하였다. 이렇게 추출된 2013년 한 해 동안의 기사가 4,467건으로 조사되었다. 이 기사들의 통계치를 살펴보면 2,243페이지, 단어 976,012개로 집계되며 자료 수집 소요 시간은 1개월당 1시간 소요 된 것으로 나타났다.

4.3 Data Dictionary

한국정보통신기술협회는 정보통신 용어 표준화 사업을 전개하고 있으며 선정된 표준화 용어와 기타 보급이 필요한 용어에 대하여 정보통신용어사전 및 정보통신표준용어집을 발간하고 이를 DB에 구축하여 TTA웹사이트를 통해 27,953개의 용어를 제공하고 있었다[22].

또한, 모바일 통신 환경과 관련된 11,156개 기사 내용에 등장하는 키워드를 정리한 Dic_Lee2013 사전을 정의하였고 4,252개의 용어를 확인할 수 있었다[23].

실험에서는 TTA용어사전의 27,953개의 용어와 Dic_Lee2013 사전의 4,252개의 용어들을 함께 통합하여 중복 용어는 제거하고 새로운 31,254개의 용어집을 생성하였다.

4.4 Frequencies analysis

간단한 명령으로 복잡한 그래프나 입체적인 그래프 작성이 용이하였고, 다양한 종류의 이미지 파일로 저장할 수 있었다. plot()는 산포도, hist()는 히스토그램, legend()는 막대그래프의 범례, abline()함수는 산포도의 회귀선, bmp(), jpeg(), pdf(), png() 등 이미지 파일 저장, curve()는 1차원 함수, persp()는 2차원 함수, 3차원 그래프 작성도 가능하였다.

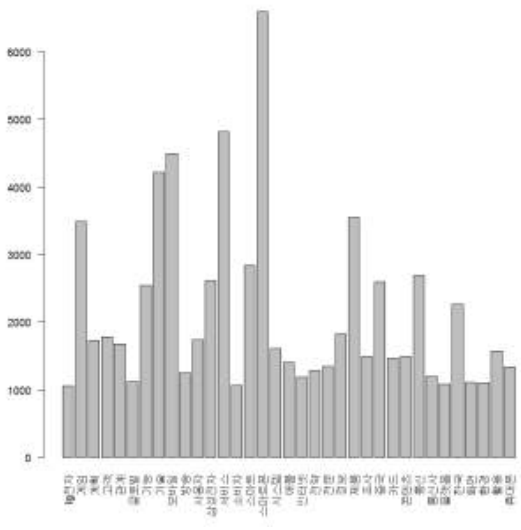

Fig. 6 Frequencies analysis using R

4.5 Word Cloud analysis

시각적으로 분석 결과를 이 보다 잘 표현할 수 있을까? 하는 의문을 해 보았다. 컬러를 자유롭게 사용할 수 있는 RColorBrewer 패키지, 빈도의 비율에 맞게 글자 크기 형태 및 색상, 각도 등을 조절하는

wordcloud 패키지가 Word Cloud 분석의 주인공이다. 보는 이의 눈을 즐겁게 하고 보다 쉽게 이해할 수 있도록 시각적으로 표현하는 분석 방법이다.

Fig. 7 Word Cloud Analysis using R

4.6 Cluster analysis

여러 개의 집단으로 유형화, 분류화하기 위한 통계해석 방법으로 몇 가지 키워드를 그룹으로 함께 클러스터 되었음을 보여준다. 즉, 비슷한 것 끼리를 하나의 그룹으로 전체를 몇 개의 계층적 클러스터로 분할하여 시각화 하는 기법이다. R에서는 dist()함수와 hclust()함수로 간단히 표현할 수 있었다.

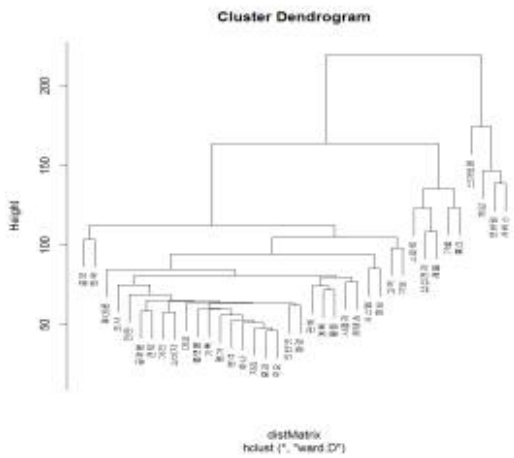

Fig. 8 Cluster Analysis using R

4.7 Social Network analysis

소셜 네트워크는 연결성의 종류 또는 사용자의 사회적 상호 작용을 파악하고 추론하는 기법이며 정보검색을 하면서 알게 된 우연한 타인과의 관계에서부터 매우 친밀한 대상과의 관계까지 다양한 사회적 양상을 확인할 수 있다. R의 matrix를 이용하여 가로, 세로에 키워드를 배치시키고 무방향 그래프를 그려 표현한다. vertex.size()함수를 통하여 노드의 차이를 조절하여 시각화 할 수 있었다.

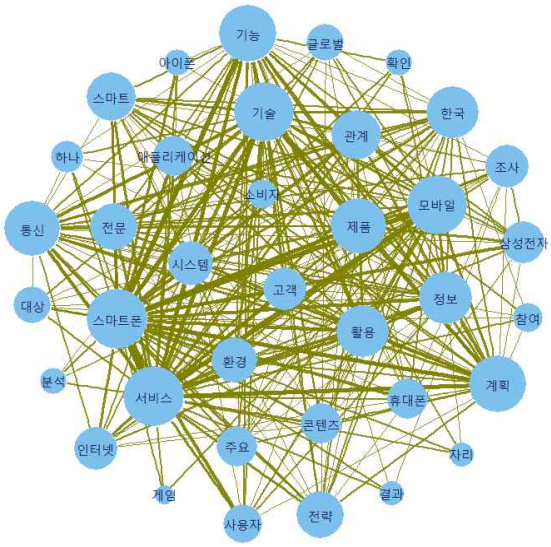

Fig. 9 Social Network Analysis using R

5. 결론 및 향후 과제

본 논문은 R을 통하여 텍스트마이닝 연구와 실험을 하였다. 그 결과를 종합적으로 정리해보면 다음과 같다.

빅데이터 분석을 위한 R은 다양한 통계적 분석 기법과 우수한 그래픽 기능을 지원하고 있다. 무엇보다 수많은 패키지들이 사용되고 있다는 것이다. 다수의 연구자들이 서로 협력을 통하여 얻게 된 지적 능력의 결과로 얻어진 R의 알고리즘을 패키지화로 개발하여 이후 연구자들이 실시간으로 함께 사용되며 Update되고 있는 것이다. 누구나 손쉽게 데이터를 생산하고 인터넷에서 공유할 수 있도록 한, 사용자 참여

중심의 인터넷 환경인 웹 2.0의 집단지성(Collective Intelligence) 원리로 선행 연구자가 함수 형식으로 개발한 패키지를 이후 사용자들이 사용할 수 있다는 것이다. 이러한 과정으로 R 구현은 편리함을 연구자들에게 제공하고 있었다.

특히, 텍스트마이닝에서 많이 사용되는 tm, KoNLP, wordcloud, stringr, ggplot2, igraph, RColorBrewer, SnowBallC등 그 수가 6,000개를 훌쩍 넘어선 것으로 집계되며 Freeware Software로써 인기도와 프로그램의 파워가 급등하고 있다. 그리고 빅데이터의 통계분석 도구인 R이 하둡과 병합하여 R_Hadoop, R_Parallel 등 빅데이터 분석 Tool로 자리 매김을 하고 있다.

통계학 소프트웨어 관점에서의 R은 기존 SPSS나 SAS와도 호환이 우수하며 CUI방식의 장점을 살린 분석 옵션 용이성과 빠른 결과물 처리는 연구자의 흥미를 이끌기 충분하며, 기초 통계량, 상관계수, 분산분석, 회귀분석을 비롯한 로지스틱, 시계열, t검정 등 다변량 데이터의 처리까지 통계학에서 모든 검정 기법을 처리하고 있다.

개발 소프트웨어 관점에서의 R은 기존 C++, JAVA나 Python 등의 알고리즘 구현 기법들을 제공하고 있었다. 조건문, 반복문, 함수 정의를 비롯한, 재귀호출(Recursive Call), 배열, 객체 지향 처리까지 가능한 프로그래밍 언어로 확인되었다. 또한, 다양한 벡터의 지원으로 연산 속도 향상과 데이터 구조의 변환이 자유롭고, Sort와 Merger 기능을 기본으로 탑재한 matrix()함수는 빅데이터 처리에 큰 도움이 되는 것으로 나타났다.

R은 스크립트(Script)형태로 초보자의 경우 사용과 접근이 쉽지 않고 데이터 핸들링이 용이하지 않았지만 대화식 Command 형식과 R Script의 Function 정의 기능이 불편함을 보완해 주었다.

하둡(Hadoop), R_Parallel의 등장으로 R을 이용한 실시간 분석이 가능하게 되었다. 실시간으로 등록되는 텍스트 데이터들을 실시간 처리를 통하여 구현이 가능하다면 텍스트마이닝의 진정한 의미가 실현될 것으로 기대된다.

본 눈문의 실험에 사용된 4천여 개의 신문기사를 텍스트마이닝 기법으로 수집, 저장, 전처리, 빈도 분석과 Cluster Analysis, Word Cloud Analysis, Social Network Analysis 등의 시각화 과정을 R을 이용하여 알고리즘 구현을 수행하였다. 연구 문서내의 단어들을 1차 추출하였다. 비교 대상인 연구 분야 관련 Data Dictionary를 1차 추출된 연구 대상 키워드들에 참조하여 2차 추출 키워드가 텍스트마이닝의 정보가 된 것이다. Data Dictionary 기반의 키워드 추출을 이용한 텍스트마이닝 기법은 연구자의 연구 영역에 분석 초점을 더욱 높일 수 있을 것으로 기대 된다.

References

[1] Lee Ji Ho, “Big Data, Data Mining and Temporary Reproduction”, The Journal of Intellectual Property, Vol. 8, No. 4, pp. 93-125, 2013.

[2] Kang S. J., “Constructing a Large Interlinked Ontology Network for the Web of Data”, Journal of Korean Industrial Information Systems Society, Vol. 15, No. 1, pp. 15-23, 2010.

[3] URL http://www.worldometers.info/kr

[4] URL htt://www.wikipedia.org

[5] Won J. Y. and Kim D. G., “Deduction of Social Risk Issues Using Text Mining”, Korean Review of Crisis & Emergency Management, Vol. 10, No. 7, pp. 33-52, 2014.

[6] Kwon H. R., Na J. H., Yoo J. S., Cho W. S., “Text-mining Techniques for Metabolic Pathway Reconstruction”, Journal of Korean Industrial Information Systems Society, Vol. 12, No. 4, pp. 138-147, 2007.

[7] Feinerer I, “An introduction to text mining in R”. R News. Vol. 8, No. 2, pp. 19-22, 2008.

[8] Zhang J, Jang J, Kim S, Lee H, Lee C, Semicon L, “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.

[9] Yang S. and Ko Y., “Extracting Comparative Elements for Korean Comparison Mining”, Journal of KIISE, Vol. 38, No. 12, pp. 689-696, 2011.

[10] “THE R TIPS(THE SECOND EDITION)”, Nobuo Funao, 2009.

[11] Feinerer I, “Introduction to the tm package

text mining in R. nd)", n.pag.Web, 2014.
[12] Meyer D, Hornik K, Feinerer I, "Text mining infrastructure in R", Journal of Statistical Software, Vol. 25, No. 5, pp. 1-54, 2008,
[13] Zhao Y, "R and data mining: Examples and case studies", Academic Press, 2012.
[14] Williams G, "Data science with R text mining", 2014.
[15] Ingo F. and Kurt H., "tm: Text Mining Package". R package version 0.6., 2014. http://CRAN.R-project.org/package=tm
[16] Hadley W., "stringr: Make it easier to work with strings". R package version 0.6.2., 2012. http://CRAN.R-project.org/package=stringr
[17] Kam M. and Song M., "A Study on Differences of Contents and Tones of Arguments among Newspapers Using Text Mining Analysis", Journal of intelligence and information systems, Vol. 8, No. 3, pp. 53-77, 2012.
[18] Kurt H., "NLP: Natural Language Processing Infrastructure". R package version 0.1-5, 2014. http://CRAN.R-project.org/package=NLP
[19] Harley W., "ggplot2: elegant graphics for data analysis". Springer New York, 2009.
[20] Csardi G., Nepusz T., "The igraph software package for complex network research", InterJournal, Complex Systems 1695. 2006. http://igraph.org
[21] Ian F, "wordcloud: Word Clouds". R package version 2.5, 2014. http://CRAN.R-project.org/package=wordcloud
[22] Telecommunication Technology Association (http://www.tta.or.kr/)
[23] Lee H. K., "An analysis of mobile communication environment by a socio-technical approach", Journal of Korean Industrial Information Systems Society, Vol. 18, No. 2, pp. 59-69, 2013

이 종 화 (Jong Hwa Lee)

- 학생회원
- 부경대학교 경영학 석사
- 부경대학교 경영학 박사과정
- 관심분야 : Big Data, Mining, Content Analysis

이 현 규 (Hyun-Kyu Lee)

- 정회원
- 연세대학교 경영학 박사
- 부경대학교 경영학부 교수
- 관심분야 : 정보시스템 전략, 커뮤니케이션 전략, 기술전략