The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.2.169
27(2), 169–183

Standardizing Unstructured Big Data and Visual Interpretation using MapReduce and Correspondence Analysis

Joseph Choi[a] · Yong-Seok Choi[a,1]

[a]Department of Statistics, Pusan National University

(Received October 16, 2013; Revised January 13, 2014; Accepted February 17, 2014)

Abstract

Massive and various types of data recorded everywhere are called big data. Therefore, it is important to analyze big data and to find valuable information. Besides, to standardize unstructured big data is important for the application of statistical methods. In this paper, we will show how to standardize unstructured big data using MapReduce which is a distribution processing system. We also apply simple correspondence analysis and multiple correspondence analysis to find the relationship and characteristic of direct relationship words for Samsung Electronics and The Korea Economic Daily newspaper as well as Apple Inc.

Keywords: Big data, unstructured data, MapReduce, correspondence analysis, direct relationship words, The Korea Economic Daily.

1. 서론

2010년을 기준으로 디지털 공간에 축적된 정보의 규모는 12억TB(terabyte)에 육박하는 것으로 추정된다 (Gantz와 Reinsel, 2010). Special Report(2010.02.25)에 의하면, 세계 최대의 소매 체인 월마트(Wal Mart)에서는 시간당 100만 건 이상의 거래 기록이 저장되며, 2008년까지 약 2,500TB의 정보가 축적되었다고 한다. 또한, 2011년 1월을 기준으로 트위터에서는 매일 약 1억 1,000만 개의 트위트가 발신되며 (Chiang, 2011), 2020년, 관리해야할 데이터량이 50배 급증할 것으로 전망된다고 한다 (Jeong, 2011).

최근 구글 자동 번역기와 지역별 검색어 빈도를 통한 독감 유행 정보, 각종 인터넷 포탈 사이트들의 품질 높은 검색어 기능, 링크 분석(link analysis)을 통한 키워드간의 연결 분석, 유권자들의 트윗(tweet)을 분석하여 맞춤형 캠페인을 펼친 미국 오바마 대통령의 선거 운동 등 비정형 빅 데이터(unstructured big data)에 대한 연구와 활용이 활발해지고 있다. 특히, Kim과 Cho (2013)는 빅 데이터 분석과 관련이 있는 통계적 방법론으로 고차원 회귀분석, 분류분석, 다중비교, 앙상블, 치적화

This study was supported by the Research Fund Program of Research Institute for Basic Sciences, Pusan National University, Korea, 2013, Project No. RIBS-PNU-2013-305.
[1]Corresponding author: Professor, Department of Statistics, Pusan National University, 2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan, 609-735, Korea. E-mail: yschoi@pusan.ac.kr

알고리즘, 차원축소, 네트워크분석, 군집화, 시각화, 온라인분석, 병렬계산, Rhive인 R 프로그램을 들고 있다.

본 연구에서는 이러한 빅 데이터를 분석하기 위하여, 분산 처리 시스템(distribution processing system)인 맵리듀스(MapReduce)를 활용하여, 비정형 빅 데이터(unstructured big data)를 정형화하고, 이를 분석하고 시각화하기 위하여 대응분석(correspondence analysis)을 활용하려 한다. 이에 먼저 2장에서는 빅 데이터의 출현배경과 개념에 대해서 설명하고, 3장에서는 빅 데이터를 처리하는 맵리듀스의 개념과 처리방법 및 대응분석의 이론에 대해 설명하고, 추가적으로 비정형 데이터를 정형화하여 분석하고 시각화하는 전체적인 작업흐름에 대하여 소개하려 한다. 4장에서는 3장의 기법을 활용하여 한국경제신문 지면에 실린 삼성전자와 애플에 대한 기사를 분석하고 시각화하여, 두 기업에 대한 분기별 이슈와 기업의 동향을 비교하며, 끝으로 5장에서 본 본문을 정리·요약하겠다.

2. 빅 데이터

2.1. 개념 및 출현 배경 및 특성

오늘날의 빅 데이터는 단순한 데이터량의 증가를 나타내는 것이 아니라 다양한 데이터의 형식, 데이터의 발생 속도, 대량의 데이터로부터의 가치 창출을 모두 포함하는 말이다. Manyika 등 (2011)은 빅 데이터란 기존의 데이터베이스 관리도구의 데이터 수집, 저장, 관리, 분석의 역량을 넘어서는 데이터를 빅 데이터라 정의하였고, Gantz와 Reinsel (2011)은 업무수행에 초점을 맞춘 정의로 빅 데이터란 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고, 데이터의 빠른 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처라 정의 하였다. 즉, 빅 데이터란 오늘날 발생하는 데이터로부터의 새로운 가치를 창출해내기 위한 모든 행위를 말한다.

빅 데이터의 특성은 크기(volume), 다양성(variety), 속도(velocity) 3개의 키워드로 나타낼 수 있다. 크기는 앞서 언급한 것과 같이 시대와 데이터를 다루는 대상에 따라 상대적으로 다를 수는 있지만, 현재는 수십 테라바이트에서 수십 페타바이트(petabyte) 이상 크기의 데이터량을 말하며, 일반적으로는 데이터를 관리하는 기존의 기술역량을 넘어서는 크기의 데이터량을 말한다. 다양성은 데이터의 구조에 따라 정형, 반정형, 비정형 구조로 나눌 수 있다. 마지막으로 속도는 데이터의 업데이트(update)를 나타내는 빅 데이터의 중요한 특징이다.

3. 비정형 빅 데이터의 정형화와 대응분석

3.1. 맵리듀스

맵리듀스는 대용량의 데이터를 처리하기 위한 분산 처리 시스템의 프레임워크 구조이다. 맵리듀스는 단어의 의미 그대로 매핑(mapping)을 하는 맵(map) 단계와 데이터를 줄이는 리듀스(reduce) 단계로 구성되어 있다. 맵 단계에서는 입력데이터(input data)로부터 분리된 데이터(separated data)의 <키(key), 값(value)>을 입력 받아 맵 함수를 통해 매핑된 <키(list(key)), 값(list(value))>의 쌍으로 내보내며 이렇게 생성된 쌍은 리듀스 단계를 거쳐 맵 단계에서 매핑된 키를 기준으로 집계 연산된 새로운 쌍의 <키(key), 값(value)>을 생성한다.

Figure 3.1은 맵리듀스를 이용하여 문장을 구성하고 있는 단어의 빈도수를 계산하는 과정이다. 우선, 입력데이터(input data)에 입력된 텍스트는 분리된 데이터 과정을 거쳐 두 개의 텍스트로 분리된다(separated text). 분리된 두 개의 텍스트는 각각의 맵 작업 공간에서 공백을 기준으로 단어들을 분리하고, 각각의 단어에게 숫자 1씩을 부여한다. 맵 단계를 거쳐 생성된 키와 값으로 이루어진 쌍은 다시

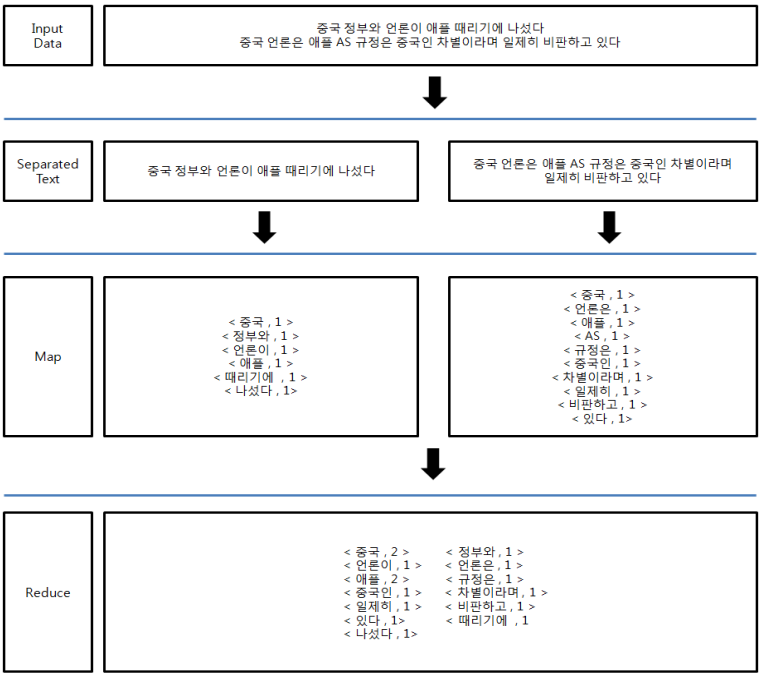


Figure 3.1. Procedure of MapReduce

리듀스 단계를 거쳐 키를 기준으로, 즉, 같은 단어를 기준으로 값을 합하고 새로운 키와 값을 가진 쌍을 생성하게 된다.

본 연구에서 사용된 맵리듀스 처리 과정의 코딩 파일(MapReduce.txt)과 예제파일(Apple.txt)의 내용은 저자의 홈페이지(yschoi.pusan.ac.kr)의 Researches>Systems 메뉴에서 받아갈 수 있다.

3.2. 대응분석

맵리듀스 과정을 거쳐 정형화된 빅 데이터는 분석 목적에 따라 행과 열 범주를 가진 범주형자료(categorical data)로 생성된다. 특히, 이들 범주간의 관계와 특징을 구체적으로 파악하고 시각화하기 위하여 다변량 기법인 대응분석을 적용할 수 있다. 대응분석이란 범주에 따른 빈도로 구성된 범주형자료의 행과 열 범주를 저차원 공간상의 점들로 나타내어 그들의 관계를 시각적으로 파악하는데 목적을 두는 탐색적 자료분석 기법이다 (Choi, 2001).

행 범주의 크기가 p이고, 열 범주의 크기가 $q\,(\leq p)$인 이원분할표 자료행렬을 $\mathbf{O} = o_{ij},\ i = 1, 2, \ldots, p;\ j = 1, 2, \ldots, q$라고 하자. 그러면 o_{ij}는 i번째 행 범주와 j번째 열 범주의 결합빈도수(joint frequency)이고, $\mathbf{O}$의 i번째 행 $(o_{i1}, \ldots, o_{iq})$는 q개의 범주를 갖는 빈도수가 $o_{i+} = \Sigma_{j=1}^{q} o_{ij}$인 다항표본(multinomial sample)이라고 할 수 있다. 이러한 이원분할표로부터 대응행렬(correspondence matrix)을

$$\mathbf{F} = (f_{ij}), \quad f_{ij} = \frac{o_{ij}}{o_{++}},\ i = 1, \ldots, p;\ j = 1, \ldots, q \tag{3.1}$$

로 정의 한다. 여기서 $o_{++} = \Sigma_{i=1}^{p}\Sigma_{j=1}^{q} o_{ij} = \mathbf{1}_p'\mathbf{O}\mathbf{1}_q$는 이원분할표의 전체 빈도수이고, $\mathbf{1}_p$와 $\mathbf{1}_q$는 크

기가 $p \times 1$과 $q \times 1$이고 모든 원소가 1인 벡터이다. 다차원의 정보의 차원 축소를 위하여, 대응행렬의 행과 열 프로파일 중심 $\mathbf{r} = \mathbf{F}\mathbf{1}_q$과 $\mathbf{c} = \mathbf{F}'\mathbf{1}_p$을 고려한 중심화 대응행렬(centered correspondence matrix) $\tilde{\mathbf{F}} = \mathbf{F} - \mathbf{r}\mathbf{c}'$의 비정칙값분해를 적용하면

$$\mathbf{D}_r^{-\frac{1}{2}} \tilde{\mathbf{F}} \mathbf{D}_c^{-\frac{1}{2}} = \mathbf{U}\mathbf{D}_\lambda \mathbf{V}' \tag{3.2}$$

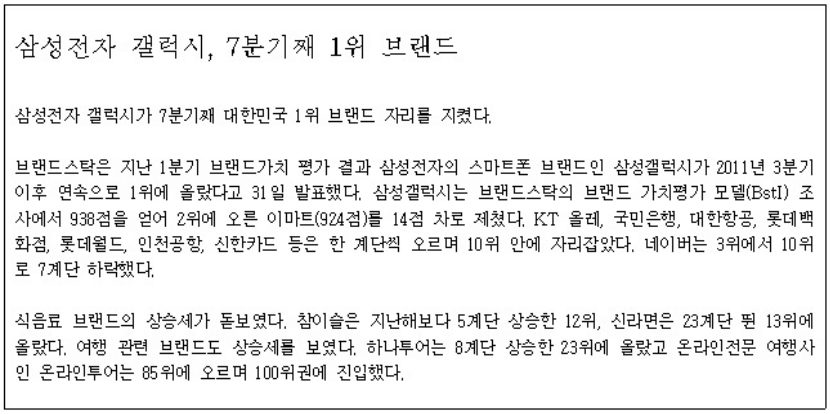
이고, 여기서 $\mathbf{D}_r = \mathrm{diag}(f_{1+}, \ldots, f_{p+})$과 $\mathbf{D}_c = \mathrm{diag}(f_{+1}, \ldots, f_{+q})$는 앞에서 언급한 행과 열 주변 비율을 대각원소로 하는 크기가 $p \times p$와 $q \times q$인 대각행렬이다. 또한, $\mathbf{D}_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$는 비정칙값을 대각원소로 갖는 크기가 $q \times q$인 대각행렬이며, $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_q$를 만족한다. 식 (3.4)로부터 차원의 차원축소에 대한 근사적합도(goodness-of-fit of the approximation)는 $\Sigma_{k=1}^{s}\lambda_k^2 / \Sigma_{k=1}^{q}\lambda_k^2$와 같다. 다음으로, 행과 열 범주의 크기가 p와 q이고 $j\,(= 1, \ldots, q)$번째 열 범주가 l_j개의 수준을 가지면, 다원분할표는 p개의 대상에 대하여 더미 변수를 이용하여 크기가 $p \times (l_1 + l_2 + \cdots + l_q)$인 표시행렬(indicator matrix)을 $\mathbf{Z} = [\mathbf{Z}_1|\mathbf{Z}_2|\cdots|\mathbf{Z}_j]$로 나타낼 수 있다. 여기서, $\mathbf{Z}_j$는 크기가 $p \times l_q$인 분할행렬이 된다. 또한, 식 (3.5) 표시행렬 $\mathbf{Z}$의 내적은 버트행렬(Burt matrix)이 된다. 다중대응분석의 대수적 알고리즘은 고유치체계를 이용한다.

3.3. 작업흐름도: 빅 데이터의 정형화와 시각화

신문 기사나 SNS의 댓글은 텍스트 형식으로 저장되어 있는 대표적인 비정형 빅 데이터이다. 특히, 신문 기사의 경우에는 사회, 경제, 정치, 문화 등에 대한 상황적 특성을 잘 반영하고 있으며, 이를 포함하고 있는 기사의 내용을 분석함으로써 시대적인 변화와 이슈, 특정 기업의 경영방침, 경영전략, 국가정책 등에 대한 유용한 정보를 추출해 낼 수 있다. 또한, 급속도로 확산되고 있는 SNS 상에서의 댓글이나 메시지는 사용자의 감정이나 생각, 경험 등을 반영하고 있으며, 이를 분석함으로써, 사회적인 동향 혹은 관심 대상의 평판에 대한 정보를 추출해 낼 수 있다. 이러한 사회적, 경제적, 정치적, 문화적 변화와 동향 혹은 사용자들의 감정이나 생각, 경험들을 포함하고 있는 수많은 텍스트가 가지고 있는 정보들로부터 가치 있는 정보를 얻기 위해서는 어떻게 의미 있는 단어들을 추출할 것인지 그리고 이러한 비정형 빅 데이터를 통계적 방법론에 적용하기 위하여 어떻게 정형화를 할 것인지가 필요하다.

첫 번째 고려해야 할 것은 텍스트 파일을 구성하고 있는 단어들의 빈도수이다. 예를 들어 박근혜 대통령의 취임 연설문에서는 국민이라는 단어가 57번, 행복이라는 단어가 20번, 경제라는 단어가 19번 언급되었다. 이처럼 단순히 특정 단어들에 대한 빈도수만으로도 우리는 박근혜 대통령의 정치적 행보가 국민행복과 경제에 많은 비중을 둘 것으로 예상할 수 있다. 이와 같이 그 맥락과 내용이 일관성이 있으며, 목적성이 뚜렷한 연설문과 같은 텍스트는 언급된 단어들의 빈도만을 고려하여도 충분히 의미 있는 정보를 추출해 낼 수 있다.하지만, 신문 기사와 같이 한 기사에 여러 목적을 나타내는 텍스트는 단순히 단어들의 빈도만을 고려하여 유용한 정보를 추출해내기는 어렵다.

Figure 3.2는 2013년 4월 16일 한국경제신문 지면의 21면 2단의 기사이며 전체적인 맥락은 삼성전자에 관한 것이다. 하지만 두 번째 문단을 살펴보면 삼성전자와 관련이 없는 KT, 올레, 국민은행 등에 관련된 단어가 언급되어 있고, 세 번째 문단에서는 식음료, 브랜드, 상승세와 같이 삼성전자와 관련이 없는 단어들이 언급 되어 있다. 이처럼 목적성이 뚜렷한 연설문과 달리, 신문 기사와 같은 텍스트는 분석 대상에 대한 분석에 불필요한 단어들을 부가적으로 포함하고 있으며, 이러한 것들의 영향으로 잘못된 분석결과나 해석을 가지고 올 수 있다. 때문에 신문 기사와 같은 텍스트를 분석하기 위해서는 단순히 텍스트를 분해하는 것이 아니라, 어떠한 문장에 관심을 가지고 있는지 혹은 어떠한 문장들이 가치 있는 정보가 될 것인지의 여부를 잘 판단하여야 한다.



삼성전자 갤럭시, 7분기째 1위 브랜드

삼성전자 갤럭시가 7분기째 대한민국 1위 브랜드 자리를 지켰다.

브랜드스탁은 지난 1분기 브랜드가치 평가 결과 삼성전자의 스마트폰 브랜드인 삼성갤럭시가 2011년 3분기 이후 연속으로 1위에 올랐다고 31일 발표했다. 삼성갤럭시는 브랜드스탁의 브랜드 가치평가 모델(BstI) 조사에서 938점을 얻어 2위에 오른 이마트(924점)를 14점 차로 제쳤다. KT 올레, 국민은행, 대한항공, 롯데백화점, 롯데월드, 인천공항, 신한카드 등은 한 계단씩 오르며 10위 안에 자리잡았다. 네이버는 3위에서 10위로 7계단 하락했다.

식음료 브랜드의 상승세가 돋보였다. 참이슬은 지난해보다 5계단 상승한 12위, 신라면은 23계단 뛴 13위에 올랐다. 여행 관련 브랜드도 상승세를 보였다. 하나투어는 8계단 상승한 23위에 올랐고 온라인전문 여행사인 온라인투어는 85위에 오르며 100위권에 진입했다.

Figure 3.2. The Korea Economic Daily(3.31.2013)

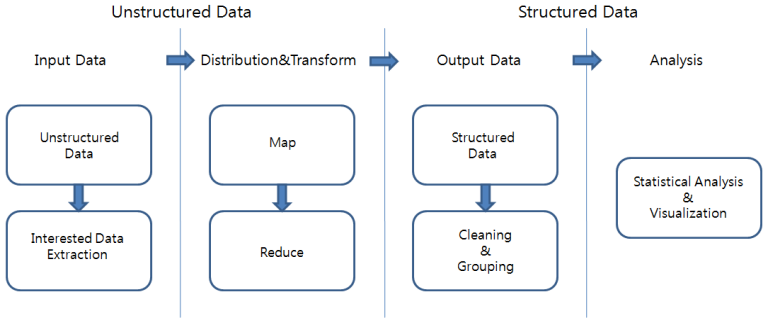


Figure 3.3. Work flow

Figure 3.2에서 관심 대상을 삼성전자로 정하고 신문 기사의 단어들을 분석하는 방법을 살펴보자. 첫 번째로, 전체 기사에 대해 언급된 단어들을 살펴보면, 식음료, 상승세, 참이슬 등 삼성전자와 관계가 없는 단어들로 인해 삼성전자에 대한 잘못된 분석이 될 가능성이 크다. 두 번째로, 삼성전자라는 단어가 포함된 문단들에 대해 단어들을 살펴보자. 전체 기사보다는 삼성전자를 분석하는 불필요한 단어들이 줄어들지만 여전히 KT, 올레, 국민은행 등과 같은 단어들의 영향으로 잘못된 분석이 될 가능성이 남아 있다. 세 번째로, 분석 대상이 되는 삼성전자를 포함하고 있는 문장에 대해서만 단어들을 살펴보면, 갤럭시, 7분기째, 대한민국, 1위 등 관계가 있는 단어들로 삼성전자가 언급되어 올바른 분석을 할 수 있다. 이와 같이 관심 대상이 되는 특정 단어가 포함된 문장의 단어들을 분석하게 되면, 불필요하게 언급된 단어들을 제외할 수 있고 관심 대상에 대해 명확하고 안정된 분석을 할 수 있다. 본 논문에서는 관심 연구 대상을 포함하고 있는 문장 속의 단어와 관심 연구 대상 사이의 관계를 직접관계(direct relationship), 연구 대상을 포함하고 있지 않은 문장 속의 단어와의 관계를 간접관계(indirect relationship)라 정의하겠다.

Figure 3.3은 비정형 빅 데이터를 정형화하여 통계적으로 분석하고 시각화하기 위한 작업흐름도이다. 작업의 흐름은 크게 비정형 데이터(Unstructured Data)단계와 정형 데이터(Structured Data)단계로 나눌 수 있다. 비정형 데이터 단계의 입력 데이터(Input Data) 단계에서는 비정형 데이터 단계와 관심 데이터 추출(Interested Data Extraction) 단계를 거쳐 관심 연구 대상을 선정하고, 분석의 목적에 따라 직접관계(direct relationship)의 데이터를 추출한다. 여기서, 직접관계란 관심 연구 대상을 포함하고 있는 문장 속의 단어와 관심 연구 대상 사이의 관계를 말한다. 이렇게 추출된 데이터는 분산 처리와

Table 4.1. Frequencies of quarterly direct relationship words via MapReduce

	1분기	2분기	3분기	4분기
삼성전자	2769	2407	2395	2060
애플	2903	2745	5106	4288

Table 4.2. Data cleaning step 1: Removing words in parentheses

Obs	Word	New_Word	Count
1	결과	결과	1
2	결과가	결과가	1
3	디스플레이	디스플레이	1
4	디스플레이를	디스플레이를	2
5	신제품	신제품	1
6	신제품을	신제품을	2
7	아이라이프(iLife)	아이라이프	1
8	아이워크(iWork)	아이워크	1
9	아이폰과	아이폰과	3
10	아이폰이	아이폰이	1

변환(Distribution & Transform) 단계의 맵과 리듀스 과정을 통하여 데이터를 분해하고, 결합함으로써, 새로운 키와 값을 가진 데이터를 만들 수 있다. 정형 데이터 단계의 출력 데이터(Output Data) 단계에서는 비정형 데이터 단계를 거쳐 생성된 새로운 키와 값을 가지고 있는 데이터를 정형 데이터 단계를 거쳐 정형화된 데이터로 만들고, 가공 및 군집화(Cleaning & Grouping) 단계에서는 통계적 분석을 하기 위하여 데이터 가공 작업과 군집화 작업을 한다. 마지막으로, 분석(Analysis) 단계의 통계적 분석 및 시각화(Statistical Analysis & Visualization)단계에서 통계적 분석 기법을 적용함으로써, 비정형 빅 데이터를 분석하고 시각화 할 수 있다.

4. 사례 분석

4.1. 데이터 소개와 가공 및 군집화

본 연구는 대표적인 비정형 빅 데이터인 신문 기사에서 삼성전자와 애플을 언급하는 단어들을 토대로 두 기업의 분기별 동향과 이슈를 살펴보기 위하여, 2012년 1월부터 2012년 12월까지의 경제 중심 종합일간지인 한국경제신문사의 지면 기사를 사용하였다. 그 중 삼성전자와 애플이라는 단어가 제목으로 언급된 지면 기사만을 이용하였다. 제목으로 언급된 지면 기사만을 이용한 이유는 삼성전자와 애플을 주제로 한 신문 기사라 판단되어서이며, 두 기업을 언급하는 불필요한 단어들을 줄이고 삼성전자와 애플에 초점을 맞추기 위한 것이다. 더욱 세밀한 분석을 하기 위하여 3.3절에서 소개한 삼성전자와 애플이 제목으로 언급된 지면 기사들 중 직접관계에 있는 단어들을 추출하였다.

Table 4.1은 각각 분기별 삼성전자와 애플 제목의 분기별 지면 기사 중 직접관계 단어의 빈도이다. 맵리듀스의 맵과 리듀스 과정을 거쳐, 삼성전자를 포함하고 있는 문장을 구성하는 1분기 단어는 총 2769개, 2분기 총 2407개, 3분기 총 2395개, 4분기 총 2060개의 단어가 문장을 구성하고 있으며, 애플을 포함하고 있는 문장을 구성하는 1분기 단어는 총 2903개, 2분기 2745개, 3분기 5106, 4분기 4288개의 단어가 문장을 구성하고 있다.

Table 4.2는 맵리듀스 과정을 거쳐 나온 직접관계 단어 중 괄호 속의 단어를 제거하는 과정이다. 괄호 속의 단어를 제거하는 이유는 괄호 속 안의 단어는 괄호 밖의 단어와 동일한 의미를 가진 단어라 판단되

Table 4.3. Data cleaning step 2: Removing postposition

Obs	New_Word	Last_Word	New_Word1	Count
1	결과	과	결	1
2	결과가	가	결과	1
3	디스플레이	이	디스플레	1
4	디스플레이를	를	디스플레이	2
5	신제품	품	신제품	1
6	신제품을	을	신제품	2
7	아이라이프(iLife)	프	아이라이프	1
8	아이워크(iWork)	크	아이워크	1
9	아이폰과	과	아이폰	3
10	아이폰이	이	아이폰	1

Table 4.4. Data cleaning step 3: Recovering words

Obs	New_Word	New_Word1	New_Word2	Count
1	아이폰이	아이폰	아이폰	1
2	아이폰과	아이폰	아이폰	1
3	아이워크	아이워크	아이워크	1
4	아이라이프	아이라이프	아이라이프	3
5	신제품을	신제품	신제품	1
6	신제품	신제품	신제품	1
7	디스플레이를	디스플레이	디스플레이	1
8	디스플레이	디스플레	디스플레이	2
9	결과가	결과	결과	1
10	결과	결	결과	1

기 때문이다. Table 4.2의 Word는 애플의 기사 중 애플과 직접관계가 있는 일부 단어를 추출한 것이며, New_Word는 Word로부터 괄호 속 단어를 제거한 것이고, Count는 Word가 나타난 빈도수이다. 아이라이프(iLife), 아이워크(iWork)는 아이라이프, 아이워크와 같이 괄호 속 단어가 제거된 것을 확인 할 수 있다. 또한, Table 4.2의 New_Word를 다시 살펴보면, '결과', '결과가'와 '디스플레이', '디스플레이를'처럼 같은 단어를 나타내고 있지만, 조사를 포함하고 있기 때문에 그룹이 제대로 묶여지지 못한 것을 확인할 수 있다.

Table 4.3은 조사에 의해 그룹이 묶이지 않는 문제점을 보안하기 위하여 단어의 조사를 제거한 표이며, 이는 한국어 조사목록에 포함된 조사의 길이가 1인 경우와 비교하여 제거한 것이다. New_Word는 Table 4.3에서 괄호 속 단어를 제거한 단어이고, Last_Word는 조사목록에 포함된 조사의 길이만큼 New_Word의 끝에서부터 추출한 것이며, New_Word1은 한국어 조사목록에 포함된 조사를 제거한 것이다. 조사가 제거된 데이터셋을 얻기 위해서는 Table 4.3과 같은 작업을 조사의 길이가 가장 긴 7부터 가장 짧은 1까지의 작업을 반복하면 된다. Table 4.3을 다시 살펴보면, '결과'와 '디스플레이'와 같은 경우 Last_Word가 조사의 역할이 아님에도 불구하고, 조사목록의 조사와 일치하는 경우, Last_Word가 New_Word로부터 제거되는 것을 확인 할 수 있다.

Table 4.4는 조사의 역할이 아닌데 제거된 글자를 복원한 표이다. New_Word와 New_Word1은 앞서 데이터 가공 단계1과 2를 거쳐 각각 괄호 속 단어를 제거한 단어와 이로부터 조사를 제거한 단어이며, New_Word2는 New_Word1에서 제거된 글자를 복원하여 새롭게 생성한 단어이다.

다음으로, Table 4.5는 단어의 군집 과정을 보여주는 표이다. 단어 군집의 목적은 주어진 많은 수의 단

Table 4.5. Clustering words

Obs	New_Word	G_Word	Min_L	G
1	7인치	7인	3	7인치
2	7인치급	7인	3	7인치
3	7인치대	7인	3	7인치
4	변호	변호	2	변호
5	변호인	변호	2	변호
6	변호사	변호	2	변호
7	배심원	배심	3	배심원
8	배심원단	배심	3	배심원
9	증거물	증거	2	증거
10	증거	증거	2	증거

Table 4.6. Segmentation of clustered word of top 5%

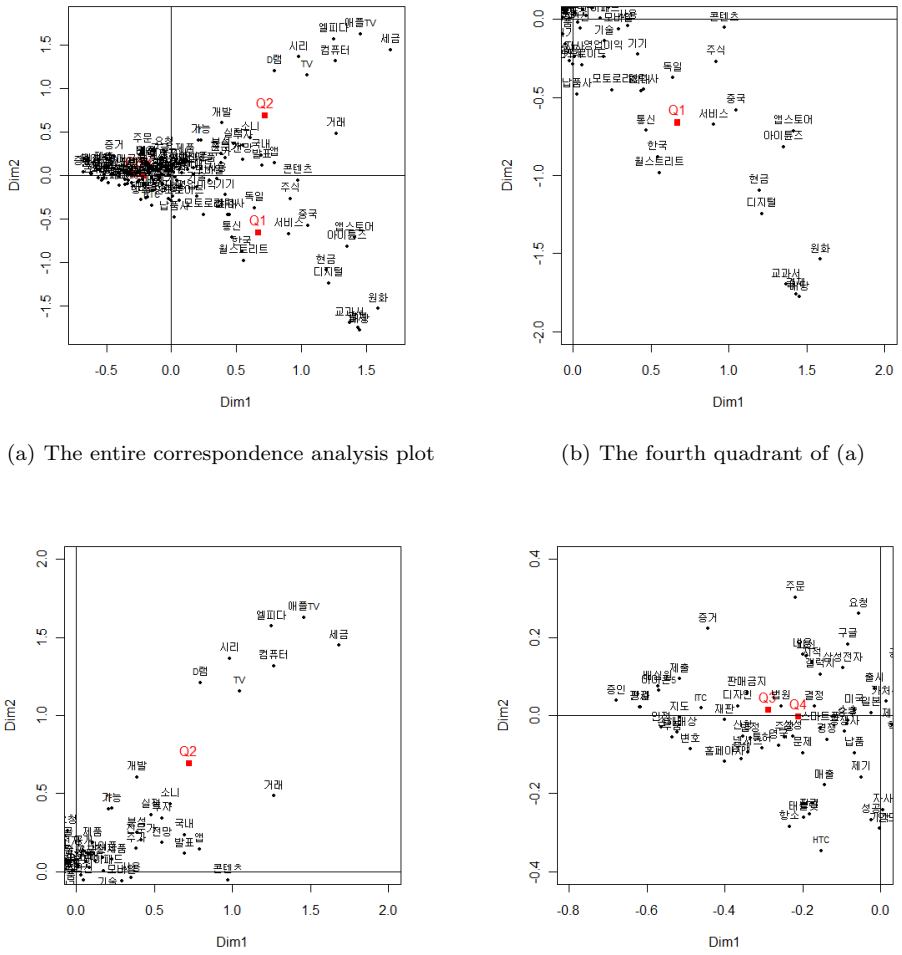
Obs	New_Word	G	New_G
1	판매금지소송	판매	판매금지
2	판매가격	판매	판매
3	판매규모	판매	판매
4	판매금지	판매	판매금지
5	판매대수	판매	판매
6	판매량	판매	판매
7	판매액	판매	판매
8	판매장	판매	판매
9	판매	판매	판매

어들을 비슷한 의미를 가진 군집으로 묶어, 흩어져 있는 빈도수를 비슷한 단어들의 묶음으로 합함으로써, 단어에 대한 특징을 더 잘 나타낼 수 있다고 판단되기 때문이다. 여기서, New_Word는 앞서 언급한 것과 같고, G_Word는 New_Word로부터 앞의 두 글자를 추출한 것이다. Min_L은 G_Word를 기준으로 그룹핑했을 때, New_Word의 가장 짧은 길이를 나타낸 것이며, G는 Min_L만큼 New_Word에서 추출하여 생성한 군집명이다. '7인치', '7인치급', '7인치대'는 '7인'으로 군집되었으며, 군집명을 G_Word를 기준으로 군집했을 때, New_Word에서 가장 짧은 단어명 '7인치'로 지정하여, 군집을 대표할 수 있는 단어의 의미를 살렸다. '변호', '변호인', '변호사'는 '변호'로 군집의 의미를 살렸으며, '증거' 또한 군집을 잘 대표하고 있는 군집명인 것을 확인할 수 있다.

4.2. 삼성전자와 애플의 분기별 특징 및 비교

4.1절의 비정형 납품 빅 데이터의 정형화와 군집화 과정을 거쳐 정형화된 신문 기사 데이터를 활용하여 삼성전자와 애플의 분기별 특징을 살펴보기 위해, 삼성전자와 애플에 대해 단순대응분석과 다중대응분석을 이용하여 보았다. 군집된 데이터로부터 각 분기별 이슈를 살펴보기 위하여 단어 출현 빈도수가 상위 5%인 군집단어를 추출하였으며, 조금 더 세밀하게 분석하기 위하여 군집을 세분화하는 작업을 하였다.

Table 4.6은 각 분기별 빈도수가 상위 5%에 해당하는 군집 단어 중 일부이며, New_Word와 G는 앞서 언급한 것과 동일하며, New_G는 군집되었던 G를 세분화시켜 군집한 것이다. New_G를 살펴보면, 처음 '판매'였던 군집이 '판매'와 '판매금지'로 '납품'은 '납품사'와 '납품'으로 특징이 세분화되어 군집된 것을 확인 할 수 있다.



(a) The entire correspondence analysis plot

(b) The fourth quadrant of (a)

(c) The first quadrant of (a)

(d) The second and third quadrant of (a)

Figure 4.1. Correspondence analysis plot of the newspaper about Apple

Table 4.7은 애플 기사에 관한 대응분석도의 결과를 분기별로 요약한 것이다. 사분면은 각 분기가 나타나 있는 위치를 나타내며, 단어는 각 분기별의 단어 일부 중 주요 특징을 나타내는 단어들이다. Figure 4.1의 (a)는 애플을 제목으로 언급한 신문 기사의 직접관계에 있는 단어들 중 분기별 빈도수가 상위 5%에 포함된 단어군집을 활용한 단순대응분석도이며, (b)는 (a)의 제4사분면, (c)는 제1사분면, (d)는 제2사분면과 제3사분면을 나타낸 그림이다. 먼저, 각 축에 대한 설명력은 제1축(Dim1)이 50.05%이고, 제2축(Dim2)은 34.32%로 총 84.37%의 설명력이 있는 것으로 나타났으며, Q1(1분기)은 제4사분면에, Q2(2분기)는 제1사분면에, Q3(3분기)과 Q4(4 분기)는 제2사분면과 제3사분면으로 나누어지는 것을 확인 할 수 있다. Table 4.7과 같이 1분기가 나타난 제4사분면의 단어들을 살펴보면, <중국, 한국, 서비스, 앱스토어, 아이튠즈, 원화, 결제 등>의 단어들이 위치하고 있어, 애플이 한국과 중국시장의

Table 4.7. Quarterly summary words of the Figure 4.1

분기	사분면	단어
Q1(1분기)	제4사분면	중국, 한국, 서비스, 앱스토어, 아이튠즈, 원화, 결제, 사용, 기기, 혁명, 콘텐츠, 모바일, 기술, 교과서, 디지털 등
Q2(2분기)	제1사분면	애플TV, TV, 컴퓨터, 개발, 신제품, 아이폰, 아이패드, 투자, 거래, 전망, 제품, IT, D램 등
Q3,Q4 (3분기, 4분기)	제2사분면 제3사분면	판매금지, 디자인, 갤럭시, 소송, 법원, 특허, 침해, 판결, 항소, 삼성전자, 주장, 제기, 재판, 요구, ITC 등

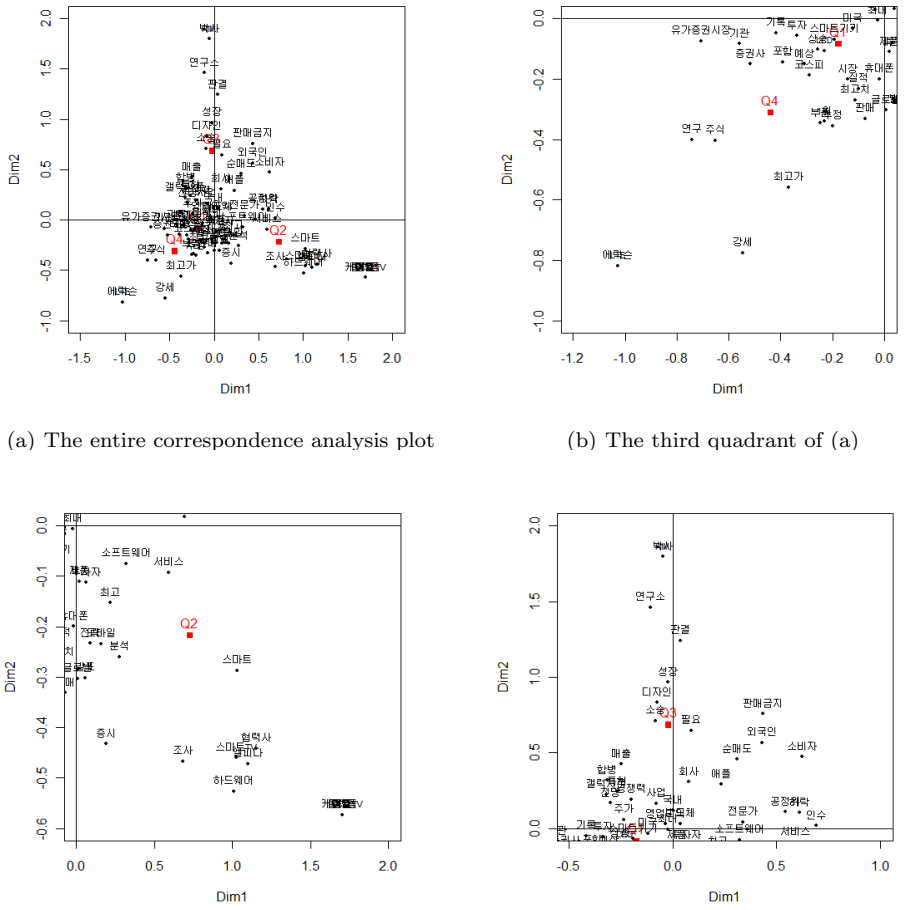
사용자에 대한 서비스를 강화할 것으로 예상된다. 또한, <콘텐츠, 모바일, 기술, 교과서 등>의 단어들이 위치하여 새로운 콘텐츠에 대해 힘을 기울이고 있음을 나타내고 있다. 2분기가 나타난 제1사분면은 <애플TV, TV, 컴퓨터, 개발, 신제품, 아이폰, 아이패드 등>의 단어들이 나타나 있다. 이는 애플의 주력 상품이었던 아이폰, 아이패드와 더불어 스마트TV나 신개념의 컴퓨터와 같은 새로운 시장 진출에 주력하고 있으며, 향후 애플이 스마트TV, 컴퓨터, 아이폰과 같은 신제품을 내놓을 것으로 예상할 수 있다. 3분기와 4분기가 나타난 제2사분면과 제3사분면에는 <판매금지, 디자인, 갤럭시, 소송, 법원, 특허, 침해, 항소, 삼성전자 등>이 위치하여, 애플이 특허 소송에 주력하고 있음을 나타내고 있다. 이를 토대로, 애플과 삼성전자의 관계는 악화될 것이라고 예상 가능하며, 애플과 삼성전자의 경쟁구도는 계속해서 치열해 질 것으로 예상된다.

Table 4.8은 삼성전자 기사에 관한 대응분석도의 결과를 분기별로 요약한 것이다. Figure 4.2의 (a)는 삼성전자를 제목으로 언급한 신문 기사의 직접관계에 있는 단어들 중 분기별 빈도수가 상위 5%에 포함된 단어군집을 활용한 단순대응분석도이다. (b)는 (a)의 제3사분면, (c)는 제4사분면, (d)는 제1사분면과 제2사분면을 나타낸 그림이다. 제1축의 설명력은 40.79%이고, 제2축의 설명력은 32.42%로, 총 73.21%의 설명력이 있는 것으로 나타났으며, Q1(1분기)과 Q4(4 분기)는 제3사분면에, Q2(2분기)는 제4사분면에, Q3(3분기)은 제1사분면과 제2사분면으로 나누어지는 것을 확인 할 수 있다.

Table 4.8과 같이 1분기와 4분기의 제3사분면은 <기록, 최고치, 실적, 1위. 주식, 강세, 상승, 등>의 단어들이 나타나 있으며, 이는 2012년 1분기와 4분기 때에 삼성전자의 상승세를 나타낸다. 2분기가 나타나 있는, 제4사분면을 살펴보면, <스마트TV, 케이블TV, 하드웨어, 전략, 분석 등>의 단어들이 나타나 삼성전자의 주력상품인 갤럭시 시리즈와 더불어 스마트TV 시장 진출에 힘을 쓰고 있는 것을 확인 할 수 있다. 3분기가 나타난 제1사분면과 제2사분면에는 <판매금지, 소송, 애플, 디자인, 특허, 외국인, 순매도, 주가, 하락 등>의 단어들이 위치하여, 애플과의 특허 소송에 의하여 외국인들 순매도가 일어난 것을 확인 할 수 있지만, <사업, 합병, 인수, 매출, 성장, 연구소, RIM 등>의 단어들도 함께 나타나 삼성전자가 애플과의 특허 소송에 맞대응하면서도 지속인 연구와 합병, 인수 등의 사업 확장을 통하여 성장하고 있는 것을 확인할 수 있다. 이를 토대로 삼성전자는 애플과의 특허 소송으로 인해 외국인들의 순매도 현상은 일시적인 것으로 예상되며, 삼성전자의 성장은 지속될 것으로 예상된다. 이는 앞서 언급한 4분기에서 삼성전자의 상승세를 나타내는 단어들의 결과와도 일치한 것을 확인할 수 있다.

Table 4.9는 삼성전자와 애플 기사에 관한 다중대응분석도 결과를 요약한 표이다. 사분면은 각 분기가 나타나 있는 위치를 나타내며, 단어는 각 분기별의 단어 일부 중 특징을 나타내는 주요 단어들이다.

Figure 4.3의 (c)는 삼성전자와 애플을 언급하는 신문 기사의 직접관계 단어들과 삼성전자와 애플의 관계를 나타내는 다중분석도이며, (a)는 (c)의 제2사분면과 제3사분면, (b)는 제1사분면과 제4사분면을 나타낸 그림이다. 제1축의 설명력은 1.18%이고, 제2축의 설명력은 0.87%이며, 총 설명력은 2.05%로 나타났다. 제1축을 기준으로 Samsung(삼성전자)은 오른편 아래쪽에 Apple(애플)은 왼편 위쪽으로 나



(a) The entire correspondence analysis plot

(b) The third quadrant of (a)

(c) The fourth quadrant of (a)

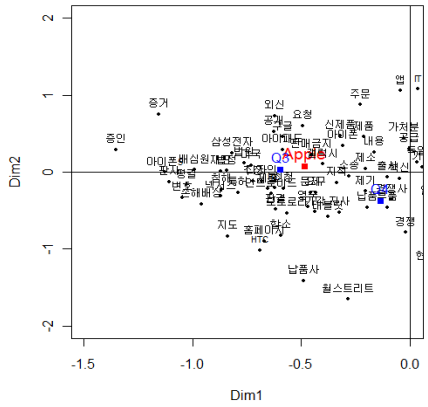
(d) The first and second quadrant of (a)

Figure 4.2. Correspondence analysis plot of the newspaper about Samsung

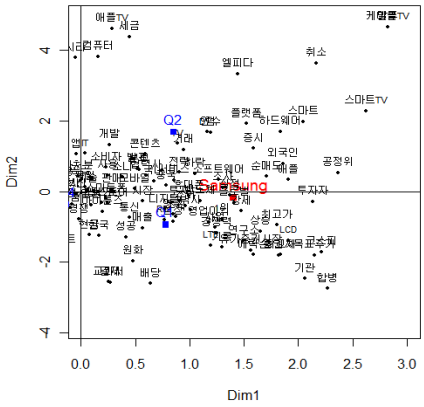
Table 4.8. Quarterly summary words of the Figure 4.2

분기	사분면	단어
Q1,Q4 (1분기, 4분기)	제3사분면	기록, 최고치, 실적, 1위, 주식, 강세, 코스피, 상승, 시장, 유가증권시장, 증권사, 최고가, 투자 등
Q2(2분기)	제4사분면	스마트, 스마트TV, 케이블TV, 하드웨어, 모바일, 전략, 분석, 서비스, 소프트웨어, 제품 등
Q3(3분기)	제1사분면 제2사분면	판매금지, 판결, 소송, 애플, 디자인, 특허, 외국인, 순매도, 주가, 하락, 경쟁력, 사업, 합병, 인수, 매출, 성장, 연구소, RIM, 소비자, 공장 등

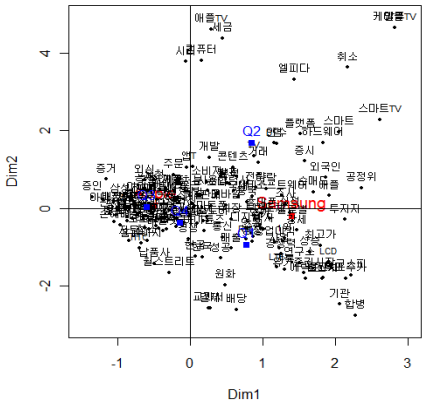
누어져 있으며, Q1(1분기)는 제4사분면, Q2(2분기)는 제1사분면, Q3(3분기)은 제2사분면, Q4(4분기)는 제4사분면에 나뉘어져 있는 것을 확인 할 수 있다. 특히, Table 4.9의 요약 결과와 같이 Samsung이 위치하고 있는 제4사분면과, 1사분면은 각각 1분기와 2분기의 단어들이 나타나 있으며, <목표주가, 기



(a) The second and third quadrant of (c)



(b) The first and fourth quadrant of (c)



(c) The entire multiple correspondence analysis plot

Figure 4.3. Multiple correspondence analysis plot of the newspaper about between Apple and Samsung

Table 4.9. Quarterly summary words of the Figure 4.3

기업	분기	사분면	단어
Samsung (삼성전자)	Q1(1분기)	제4사분면	매출, 원화, 성공, 상승, 최고가, 영업이익, 1위, 목표주가, 교과서, 경쟁력, 코스피, 한국, 현금 등
	Q2(2분기)	제1사분면	D램, 하드웨어, 스마트, 스마트TV, 애플TV, 콘텐츠, 개발, 모바일, 소프트웨어 등
Apple (애플)	Q3(3분기)	제2사분면	외신, 증거, 증인, 삼성전자, 요청, 제소, 소송, 판매금지, 배심원 등
	Q4(4분기)	제4사분면	지도, 홈페이지, 손해배상, 변호, 제기, 경쟁사, 모토로라, 판결, 경쟁 등

록, 상승, 최고치, 1위, 영업이익, 성장, 실적, 강세 등>의 기업의 성장에 관련된 단어들과 <스마트TV, 스마트, 인수, 분석, 콘텐츠, 플랫폼, 개발, 디지털, 연구, 사업 등>의 사용자를 위한 서비스, 제품 혹은

사업 확장 등에 관한 단어들이 주로 나타나 있는 것을 확인 할 수 있다. 반면, Apple이 위치하고 있는 제2사분면과 제4사분면의 3분기와 4분기에 대응하는 <가처분, 경쟁사, 소송, 판매금지, 기각, 판결, 항소 법원, 침해, 재판, 특허 등>은 주로 특허 소송에 관련된 단어들인 것을 확인 할 수 있다. 이를 토대로 2012년 애플은 특허 관련 소송에 주력을 하였으며, 삼성전자는 사업 확장과 성장에 주력을 하였다는 것을 확인할 수 있다. 또한, 이는 2013년 상반기 삼성전자는 지속적인 성장과 사업 확장으로 인한 상승세를 전망할 수 있으며, 애플은 계속되는 특허 관련 소송으로 인해 자사의 슬로건인 혁신적 기업 이미지에 타격받아 성장이 주춤할 것 할 것으로 전망할 수 있다.

5. 결론

빅 데이터 분석은 다양한 분야에서 빠른 속도로 증가하는 정형화 및 비정형화 데이터를 분석하여, 다양한 형태로 축적되어 있는 대용량의 데이터로부터 잠재 되어 있는 가치를 찾아낼 수 있다. 본 연구에서는 맵리듀스를 활용하여 비정형 빅 데이터를 정형화하고, 가공 및 군집화를 통하여 통계적 기법에 적용할 수 있도록 하였다. 더불어, 대응분석을 활용하여, 정형화된 빅 데이터를 시각화하고 해석하였다. 이에 삼성전자와 애플이 언급된 2012년 1월부터 2012년 12월까지의 신문 기사를 정형화하여 통계적 분석 기법에 활용할 수 있도록 하였고, 대응분석에 적용하여 분기별 삼성전자와 애플의 특징과 동향을 살펴보았다.

그 결과 삼성전자는 1분기와 4분기에 주가의 상승, 최고 실적 달성, 목표주가 경신 등을 확인 할 수 있었고, 2분기에는 스마트TV 시장, 시장 분석과 전략을 통해 새로운 시장 진출에 주력한 것을 확인 할 수 있었다. 3분기에는 애플과의 특허 소송으로 인하여 일시적으로 외국인의 순매도가 일어났지만, 합병과 기업 인수 계획 등의 사업 확장을 통하여 성장하였던 것을 확인 할 수 있었다. 또한, 애플의 1분기와 2분기는 주력 상품인 아이폰과 더불어 시리, 디지털 교과서, 애플리케이션 개발, 스마트TV 시장과 같은 새로운 시장 진출에 주력하였으나, 3분기와 4분기에는 삼성전자와의 특허 소송에 주력 하였던 것을 확인할 수 있었다. 뿐만 아니라, 이를 토대로 2013년 상반기 IT시장에서는 기업의 성장과 사업 확장에 주력했던 삼성전자의 강세를 전망할 수 있었으며, 혁신이 경영방침과 기업 이미지였던 애플의 성장은 주춤할 것이라고 전망할 수 있었다. 실제로도 2013년 상반기 삼성전자는 새로운 개념의 스마트 카메라 출시, 매출 200조, 1분기 순이익 7조, 갤럭시S4 출시, 스마트 시장 공략 강화 등 IT시장에서의 강세를 보이고 있다. 반면 애플은 애플쇼크, 10년 만에 순이익 감소, 고객충성도 하락 등으로 성장이 주춤하고 있다.

이처럼, 대용량의 텍스트를 정독하지 않고도, 빅 데이터 분석만으로 많은 정보를 얻을 수 있으며, 나아가 현재에도 폭발적으로 증가하고 있는 빅 데이터를 이해하고 분석함으로써, 다양한 분야에서 가치 있는 정보를 얻고, 활용 할 수 있을 것이라고 기대하여 본다.

References

Adrian, M. (2011). It's going mainstream, and it's your next opportunity, *Teradata Magazine*, AR-6309.

Choi, Y. S. (2001). *Understanding and Application of Correspondence Analysis using SAS*, Freedom Academy, Seoul.

Chiang, O. (2011). Twitter Hits Nearly 200M Accounts, 110M Tweets Per Day, Focuses On Global Expansion, *Forbes*, Available from: http://www.forbes.com/sites/oliverchiang/2011/01/19/twitter-hits-nearly-200m-users-110m-tweets-per-day-focuses-on-global-expansion/

Dean, J. and Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters, *OSDI*, **1**.

Gantz, J. and Reinsel, D. (2010). The digital universe decade-are you ready, *White Paper, IDC*.

Gantz, J. and Reinsel, D. (2011). Extracting value from chaos, *IDC iView*, 1–12.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. and Brilliant, L. (2008). Detecting influenza epidemics using search engine query data, *Nature*, **457(7232)**, 1012–1014.

Greenacre, M. J. (1984). *The and Applications of Correspondence Analysis*, Academic Press, New York.

Gruman, G. (2010). Tapping into the power of big data, *Technology Forecast*, **2010(3)**, 4–13.

Jeong, J. S. (2011). New value creation engine, new possibilities of big data and the corresponding strategy, *IT & Future Strategy*, **18**, National Information Society Agency.

Kim, Y. and Cho, K. H. (2011). Big data and statistics, *Journal of the Korean Data & Information Sciences Society*, **24(5)**, 959–974.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A. H. (2011). big data: The next frontier for innovation, competition, and productivity, *McKinsey Global Institute*, 1–137.

Special Report (2010.02.25). Data, data everywhere, *The Economist*, Available from: http://www.economist.com/node/15557443

맵리듀스와 대응분석을 활용한 비정형 빅 데이터의 정형화와 시각적 해석

최요셉[a] · 최용석[a,1]

[a]부산대학교 통계학과

(2013년 10월 16일 접수, 2014년 1월 13일 수정, 2014년 2월 17일 채택)

요 약

오늘날, 다양한 분야에서 다양한 형태의 빅 데이터들이 축적되고 있다. 이에, 빅 데이터를 분석하고 그 속에서 가치 있는 정보를 찾아내는 것은 매우 중요해지고 있다. 또한, 비정형 빅 데이터를 정형화하여 통계적 기법을 적용할 수 있게 하는 것은 매우 중요해지고 있다. 본 연구에서는 분산처리 시스템인 맵리듀스를 활용하여 비정형 빅 데이터를 정형화하고, 통계적 분석 기법인 단순 대응분석과 다중 대응분석을 적용하여, 한국 경제 신문의 지면에 실린 기사를 이용해 삼성전자와 애플을 언급하고 있는 단어들의 관계와 특성을 각각 파악하였다.

주요용어: 빅 데이터, 비정형 데이터, 맵리듀스, 대응분석, 직접관계 단어, 한국 경제 신문.

이 논문은 2013년도 부산대학교 기초과학연구원 기초과학연구지원사업비에 의하여 연구되었음 (RIBS-PNU-2013-305).

[1]교신저자: (609-735) 부산광역시 금정구 부산대학로 63번길 2, 부산대학교 통계학과.
E-mail: yschoi@pusan.ac.kr