http://dx.doi.org/10.15813/kmr.2018.19.1.007

빅데이터 분석을 위한 비용효과적 오픈 소스 시스템 설계[1]

Designing Cost Effective Open Source System for Bigdata Analysis

이 종 화 (Jong-Hwa Lee) 부경대학교 경영학부 강사[2]
이 현 규 (Hyun-Kyu Lee) 부경대학교 경영학부 교수[3]

ABSTRACT

Many advanced products and services are emerging in the market thanks to data-based technologies such as Internet (IoT), Big Data, and AI. The construction of a system for data processing under the IoT network environment is not simple in configuration, and has a lot of restrictions due to a high cost for constructing a high performance server environment. Therefore, in this paper, we will design a development environment for large data analysis computing platform using open source with low cost and practicality. Therefore, this study intends to implement a big data processing system using Raspberry Pi, an ultra-small PC environment, and open source API. This big data processing system includes building a portable server system, building a web server for web mining, developing Python IDE classes for crawling, and developing R Libraries for NLP and visualization. Through this research, we will develop a web environment that can control real-time data collection and analysis of web media in a mobile environment and present it as a curriculum for non-IT specialists.

Keywords: Big Data, Raspberry Pi, Real-time Processing, Web Mining, Text Mining

1. 서론

빅데이터(Big Data), 사물인터넷(IoT), 인공지능(AI), 클라우드 컴퓨팅(Cloud Computing), 시뮬레이션(Simulation), 자율주행자동차 등은 대표적인 제 4차 산업 혁명을 이끌고 있는 데이터 기반 산업들이다. 이들은 미래의 먹거리 산업에서 경쟁우위를 획득하기 위한 국가 간, 기업 간의 총성 없는 전쟁이 시작되고 있는 것이다(김윤경, 2017; 서새남, 2017). 이런 산업들의 기반이 되는 데이터들은 ICT기술의 빠른 발전 덕분에 더욱 상세하게 기록하고 있다. 또한, 비정형의 다양한 데이터, 영상 데이터(CCTV, 동영상), 문자 데이터(SMS,

[1] 논문접수일: 2017년 12월 29일; 1차 수정: 2018년 2월 1일; 게재확정일: 2018년 2월 20일
[2] 주저자 (newjwcom@daum.net)
[3] 교신저자 (hyunqlee@pknu.ac.kr)

검색어), 위치 데이터 등 민간 분야, 공공 분야의 모든 업종에서 데이터를 양산 중에 있다. 이러한 거대한 데이터는 기업, 정부 할 것 없이 모든 기관에서 수집과 분석이 이루어지고 있다(6. 김정선 외, 2014; 서새남, 2017). 또한, 사용자 정보와 관계 정보 그리고 소비자 형태에 따른 고객 데이터 관계 분석, 페이스 북, 트위터, 언론에서의 이슈 정보를 분석하는 SNS 비정형 데이터 분석, 이미지나 동영상의 의미 분석과 콘텐츠 소비 형태 또는 선호도 분석 등의 대용량 멀티미디어 분석, 실시간 사물 센서 데이터 분석이나 RFID, 원격 헬스 모니터링과 같은 M2M 센서정보 분석 등과 같이 데이터 분석 기술은 빠르게 발전되고 있다(조성룡, 2012). 빅데이터 시대에는 단순히 관계형 데이터베이스에 잘 정리된 정형 데이터뿐 아니라 인터넷을 통해 매일 3억 부 이상 발행되는 신문, 1400억 건 이상 발송되는 이메일, 4억 건 이상 발생되는 트윗 등의 비정형 빅데이터를 효과적으로 분석하는 것이 무엇보다 중요해졌다(http://www.worldometers.info/kr/). ETRI의 보고에 따르면 빅데이터는 2020년에는 40 Zettabyte로 급격히 증가할 것이며, 그 중 20%는 정형화 된 데이터, 나머지 80%는 비정형화 된 데이터가 될 것으로 예상된다고 한다(https://www.etri.re.kr/).

이러한 배경하에 본 연구는 비정형 데이터의 비중이 커지는 웹 환경에서 텍스트 중심의 마이닝 연구를 하고자 한다(김은우·금득규, 2014; 이종화·이현규, 2017). 초소형 웹 서버 구축과 웹 프로그래밍 개발을 통하여 데이터 수집, 가공 및 처리, 통계 분석, 시각화 과정을 실시간으로 처리할 수 있는 처리 시스템을 구축하고, 현장연구 환경을 조성하고자 한다. 즉, 연구과정에 웹 서버 시스템의 하드웨어 설계 과정과 웹 프로그램 오픈소스 알고리즘인 소프트웨어 설계를 함께 제시하고자 한다.

본 연구를 통해 많은 연구자들의 데이터 수집 시스템에 관한 한계점을 일부 해결해 보고자 한다(안정국•김희웅, 2015; Lee and Lee, 2015; 임좌상•김진만, 2014; 이철성 외, 2013;). 이 과정에서 웹 서버 시스템 하드웨어 환경과 웹 마이닝 과정에 사용된 웹 프로그래밍, Python, R 등 소프트웨어 환경을 활용하여 최신 데이터를 반영한 이슈 분석 시스템인 빅데이터 분석의 과정을 시연하고자 한다. 또한, 웹 서버 구축의 비용과 운영의 편의성을 갖는 라즈베리 파이(Raspberry Pi) 보드를 활용하고자 한다. 보드를 이용한 웹 서버구축과 빅데이터 자료 수집과 분석 처리 시스템이 개발 된다면 빅데이터 연구 환경의 접근이 쉬워지며 대부분의 응용 프로그램들이 오픈 소스 소프트웨어로 이루어져 기존 개발자의 소스코드를 활용할 수 있으며 연구에 필요한 시간을 단축할 수 있다.

2. 기존문헌 연구

2.1 4차 산업혁명 시대

1784년 영국에서 시작된 증기기관의 동력을 활용한 기계화로 대표되는 1차 산업혁명과 1870년 전기를 동력으로 한 대량생산이 본격화된 2차 산업혁명은 인류가 쉼 없이 발전할 수 있는 동력과 노동 분업, 대량생산 체제가 가능하도록 만들어 인류의 역사적 성장을 이끌어 갔다. 또한, 1969년 인터넷이 주도한 컴퓨터 정보화와 생산 시스템 자동화를 이끈 3차 산업혁명은 전자, 정보, 통신 기술을 네트워크 기반으로 제조업 자동화에 공헌을 하며 인류를 더욱 변화 시켰다(박성원, 2017; 서새남, 2017; 서병조·나성욱, 2017).

정보통신기술(ICT)의 융합으로 이뤄지는 차세대 산업혁명인 4차 산업혁명은 빅데이터, IoT, 3D 프린팅, 인공지능의 주요 기술들을 의제로 2016년 다보스포럼에서부터 논의되기 시작하였다. 즉, 4차 산업혁명은 로봇이나 인공지능(AI)을 통해 실재와 가상이 통합되어 사물을 자동적, 지능적으로 제어할 수 있는 가상 물리 시

스템의 구축이 기대되는 산업상의 변화를 일컫는 것이다(서새남, 2017). 또한, 4차 산업혁명은 급속한 디지털 환경의 변화로 자동화가 가속화 되고 있다. 디지털화(digitalization), 플랫폼 노동자(Platform worker) 등 전 세계 모든 노동 환경을 변화시키며 4차 산업혁명 기술들이 일하는 방식에 많은 영향을 미치고 있다(박성원, 2017). 이 모든 변화의 중심은 네트워크의 혁신이며, 이 모든 4차 산업혁명을 가능하게 만드는 기반으로 자리 잡고 있다. 네트워크 지능화는 인공지능 기술과 소프트웨어 제어 체계를 활용하여 스스로 네트워킹을 구성하고 있다. 또한 네트워크로 인한 정보 격차가 사회, 경제, 문화 등 산업 전반에 걸쳐 격차가 해소되고 있다(서병조·나성욱, 2017).

새로운 산업혁명의 큰 이슈는 데이터 처리 기술이다. 빅데이터 분석 기술의 발달로 정형, 반정형, 비정형 데이터의 유형을 구분하여 패턴을 찾기 위한 노력들이 계속되고 있다. 또한, 사람이 생산하는 데이터, 컴퓨터가 생산하는 데이터, 소셜 네트워크로 인한 관계 데이터 등 생산주체에 따라 많은 연구들이 진행되고 있다(Chen et al., 2012; Wu et al., 2014).

정형 데이터 마이닝은 의사 결정 나무(Decision Trees)와 같이 입력데이터를 통해 학습함으로써 새롭게 주어진 데이터에 대해 어떤 그룹에 속해있는지를 구분하는 분류분석(Classification Analysis)이 있으며, 연속형 값을 예측하는 것으로 시계열 변수를 이용해 예측하는 예측분석(Prediction Analysis), 특성에 따라 고객을 여러 개의 집단으로 나누어 가까운 데이터끼리 하나의 그룹으로 분류하는 군집분석(Clustering Analysis), 장바구니 분석이라 하며 구매 품목간의 관계를 알아보는 연관 분석(Association Analysis)등이 있다(조재혁, 2004; lee et al., 2016, Sivakumar, 2015).

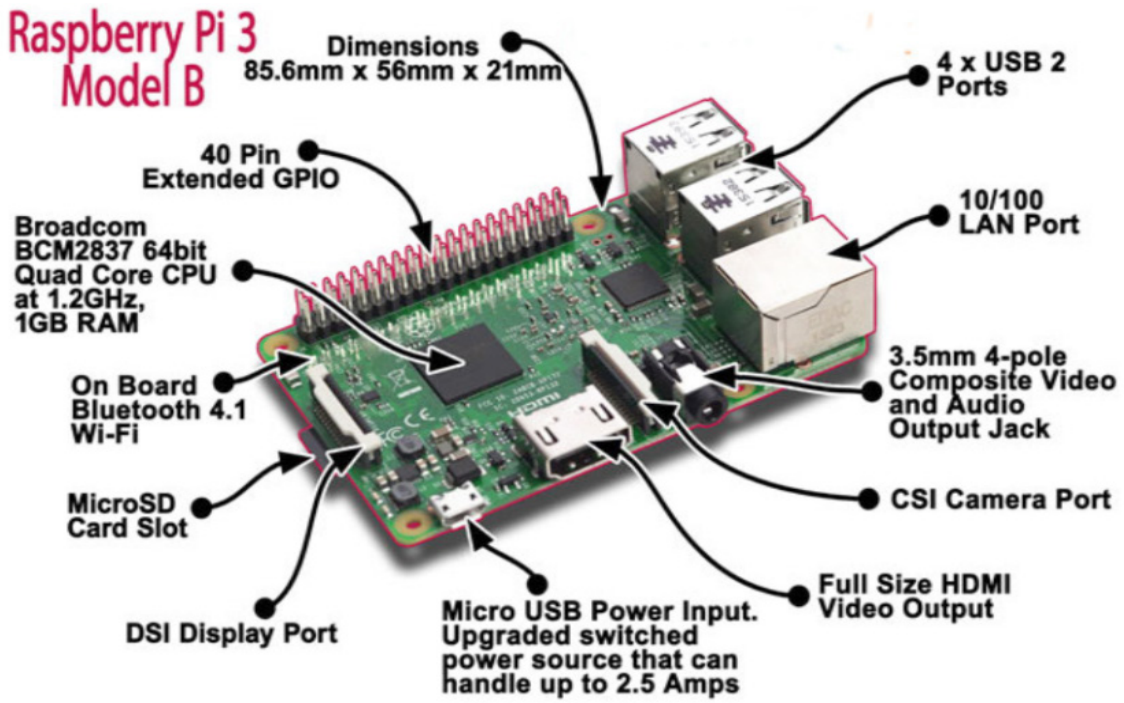
비정형 데이터 마이닝은 문서의 요약, 분류, 군집, 추출기능을 하는 텍스트 마이닝(Text Mining)과 서로의 관계가 확산되어 형성된 사람들 간의 네트워크 분석으로 사회 연결망 분석(Social Network Analysis)으로 구분된다(Lee and Lee, 2015, Fayyad et al., 1996).

데이터 마이닝은 대용량 데이터 내에서 의미 있는 패턴을 찾아 집단을 분류하고 예측하며, 유사집단으로 묶거나 동시 또는 순차적으로 발생하는 의미 있는 연관관계를 찾는 것을 목표로 하고 있다. 이는 마케팅 전략으로 보면 기업이 보유하고 있는 방대한 데이터에 존재하는 유용한 정보를 발굴하여 경영자의 의사 결정에 도움이 되는 지식을 제공할 수 있도록 하는 과정으로, 데이터가 아닌 지식 수준의 고객 통찰력을 기반으로 수행되어야 하는 고객관계마케팅 전략에 있어 핵심적인 기반 기술이다(Written et al., 2016; Liu, 2013).

웹 마이닝은 인터넷의 웹 문서 그 자체나 웹 환경의 로그와 같은 웹 데이터를 기반으로 데이터 마이닝 처리 기법을 활용하여 의미 있는 또는 특정 패턴을 찾기 위한 과정을 의미한다(Zhang and Segall, 2008; Kosala and Blockeel, 2000).

2.2 라즈베리 파이(Raspberry Pi)

영국의 라즈베리 파이 재단에서 아이들에게 값싸고 쉽게 접할 수 있는 교육용 컴퓨터 보급을 목적으로 개발되었다. 요즘 대부분의 가정에는 컴퓨터나 노트북, 태블릿 PC 등 한 두 대씩은 있지만 아이들 마음대로 삭제와 설치를 반복하면서 가지고 놀며 컴퓨터를 배울 수 있는 환경은 학교나 가정 모두 찾아보기 힘들다. 보통 PC급 서버 컴퓨터 한 대를 마련하려면 수백 만원이 드는 것을 생각하면 마음껏 가지고 놀 수 있는 컴퓨터 교육 환경을 마련하기 어렵고, 교육하려는 몇 가지 프로그램을 설치하여 사용법을 가르치는 정도로만 활용 가능하다. 컴퓨터를 배우려는 아이들이 자신의 컴퓨터를 한 대 또는 여러 대를 소유하여 마음껏 다양한 OS 를 설치해보고 다양한 설정들을 변경해보고 필요한 프로그램을 찾아 설치하고 상황에 따라서는 프로



<그림 1> 라즈베리 파이 3 구조

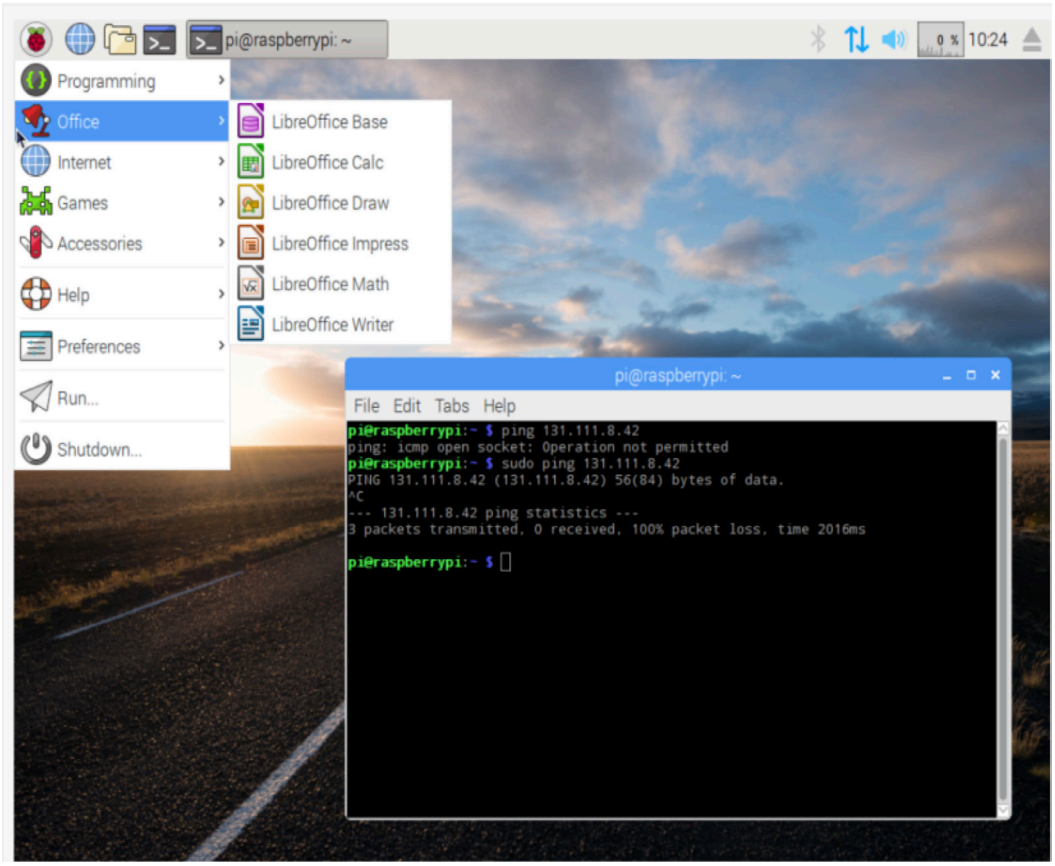
그래밍 공부를 하고 기존의 프로그램이나 운영체제를 변경해보는 등의 컴퓨터를 자유롭게 가지고 놀며 배울 수 있도록 하는 목적에서 만들어진 것이 라즈베이 파이(Raspberry Pi) 이다(김세민·최숙영, 2017; 김영근 외, 2016; 황보람·김성규, 2016; 김영근, 2014).

라즈베리 파이는 5년 전 2012년에 첫 출시되었으며 2015년 두 번째 버전, 2017년 연구 모델인 라즈베리 파이 3이 출시 되었다. B 타입 보드는 이더넷 칩이 내장되어 네트워크 기능을 강화한 보드이다(김영근 외, 2016).

<그림 1>와 같이 라즈베리 파이3의 구조를 나타난 것이며 보드기판의 사양을 간단히 설명하면 다음과 같다. 메인 프로세스는 64Bit 쿼드코어 1.2GHz CPU가 장착되어 있으며 1GB의 메모리와 온보드의 Video-Core IV MP2 400 MHz의 그래픽을 기본으로 제공하고 있다. 또한, 라즈베리 파이의 운영체제는 라즈비안을 사용하며 데비안 리눅스를 기반으로 만들어진 무료 배포판으로서 처음 사용하는 사람들이 일반적으로 설치하는 운영체제이다(www.debian.org).

운영체제 탑재된 화면은 다음 <그림 2>과 같다. 프로그래밍을 위한 통합 개발 환경인 컴파일러, 코드 편집기, 디버거 등 소프트웨어 어플리케이션 인터페이스를 제공하고 있다. 일반 문서를 만들 수 있는 오피스 프로그램, 인터넷 도구, 터미널, C/C++ Geany Editor, 텍스트 편집기인 Leafpad 등 사용자 편의를 제공하는 유틸리티도 함께 제공하고 있다.

라즈베리 파이는 오픈 소스 개발환경을 제공하며 아파치(Apache) 웹 서버를 설치하여 HTML, JavaScript, PHP 등 웹 프로그램을 작성할 수 있는 개발환경이다(김세민·최숙영, 2017). 본 연구는 라즈베리 파이를 활용한 최저가 휴대용 서버 시스템 구축과 웹 마이닝 처리를 위한 웹 서버 환경, 크롤링을 위한 Python 클래스 개발, 한글 자연어 처리와 시각화를 위한 R 라이브러리

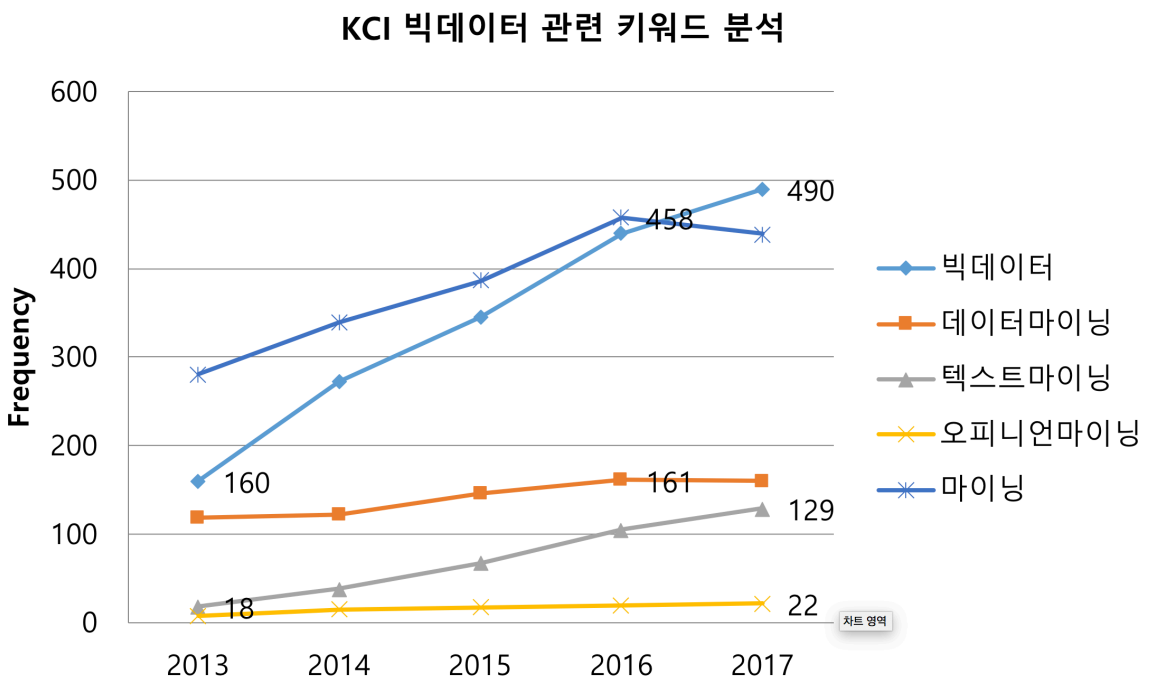


<그림 2> 라즈베리 파이 메인 화면

프로그래밍 등 빅데이터 분석 시스템을 구축하고자 한다. 또한 인터넷 신문 기사의 실시간 자료 수집과 분석, 시각화를 모바일 환경에서 제어할 수 있는 웹 환경을 개발하고자 한다.

3. 연구모델 및 방법

4차 산업혁명 시대의 도래를 통하여 많은 연구자들은 빅데이터에 관한 끊임없는 연구를 진행하고 있다 (Witten et al., 2016; 정민영, 2016). <그림 3>은 한국학술지인용색인을 이용하여 “빅데이터”, “데이터마이닝”, 텍스트마이닝”, 오피니언마이닝”, “마이닝”의 키워드 빈도를 조사한 결과이다. 한국연구재단의 등재학술지와 등재후보학술지 구분 없이 2013년 1월부터 2017년 12월까지 논문제목 및 키워드, 초록에 대하여 해당 키워드 빈도수를 나타내었다. 대부분의 키워드는 지속적으로 증가하는 것을 볼 수 있으며 특히, “빅데이터” 키워드는 2013년 160건, 2017년 12월 현재 490건으로 집계되어 4년 사이 3배 가까이 증가한 연구분야라 할 수 있다. “텍스트마이닝”이라는 키워드 또한 2016년 18건으로 시작하여 2015년 67건으로 3배, 2017년은 129건으로 무려 7배의 증가된 연구분야로 나타나고 있다. 국내 빅데이터 산업의 많은 연구가 이어지고 무수히 많은 텍스트 사이에 패턴을 찾고 통찰력을 키우기 위한 텍스트마이닝 연구 또한 활발히 진행되는 것을 볼 수 있다(www.kci.go.kr).



<그림 3> KCI 빅데이터 관련 키워드 분표도

<그림 3>을 통하여 빅데이터 분야의 지속적 연구 가치를 확인할 수 있었고 저비용과 실용성을 앞세워 오픈소스를 활용한 빅데이터 분석을 위한 컴퓨팅 플랫폼 개발 환경을 공유하므로 보다 많은 연구자들의 현장연구가 기대된다.

본 논문에서는 하드웨어 측면에서 신용카드 크기의 작은 컴퓨터로 빅데이터 자료 수집 및 분석 과정을 교육할 수 있는 초소형 서버용 PC 환경을 제공하는 라즈베리 파이(Raspberry Pi)를 활용하고자 한다. 소프트웨어 측면에서는 운영체제, 개발 툴, DB 환경 등 오픈소스 API를 활용하여 빅데이터 처리 시스템을 연구하고 있다. 또한 실시간 자료 수집과 분석을 모바일로 제어할 수 있는 웹 환경을 개발하여 빅데이터 학습 과정으로 새로운 커리큘럼을 제시하고자 한다.

시스템 개발을 진행하기 위한 재료를 살펴보면 <표 1>과 같다. 라즈베이 파이는 초소형 컴퓨터로 내장 주기판에 칩과 메모리, 그래픽 카드, 사운드 카드, 이더넷 카드 등이 함께 배치된 온 보드(On Board)구조이다. 즉, 하나의 보드만으로 사무용 본체 하나가 구성된 형태이며 보드의 크기는 명함 크기에 각종 인터페이스를 제공하고 있다.

이러한 보드에 외부기억장치인 하드디스크를 대신하여 SD카드를 장착하며 보드를 보호하기 위한 케이스로 서버용 본체를 조립하였다. 출력 장치 모니터와의 인터페이스는 HDMI(High Definition Multimedia Interface) 포트를 제공하며 4개의 USB(Universal Serial Bus) 포트를 제공하여 키보드, 마우스, USB 테더링, 화상카메라 등을 연결할 수 있다. 본체 전원은 5핀 USB케이블로 적용하여 일반 PC의 USB단자에서 전원을 제공받을 수 있고, 스마트폰 보조배터리의 USB 포트를 사용하여 전원공급을 받을 수 있다.

서버 구축에서 선행되어야 하는 것은 안정된 OS 탑재이다. 운영체제는 리눅스 데비안(Debian)을 기반으로 라즈비안(Raspbian)을 라즈베리 파이 공식 OS로 사용하고 있으며 라즈베리 파이 공식 홈페이지에서 ISO파일로 이미지 파일을 SD카드에 적재하는 별도 프

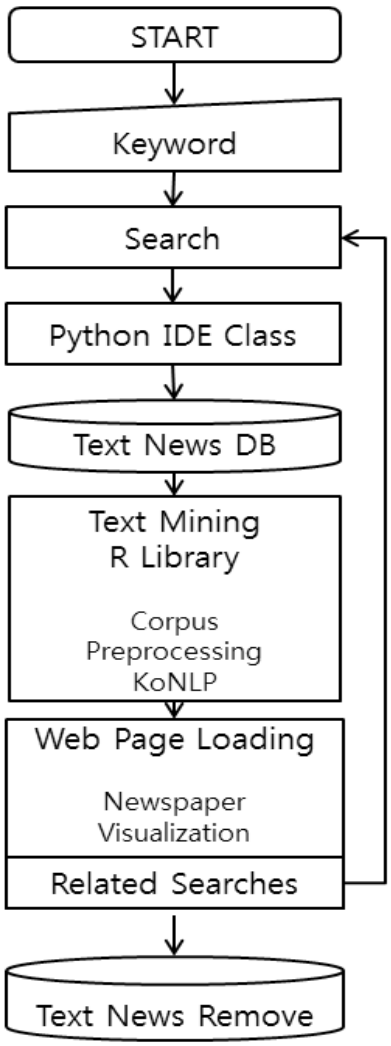
<표 1> 시스템 재료 목록

번호	품명	용도
1	라즈베리파이3	메인프로세스
2	32GB SD카드	보조기억장치
3	케이스	보드보호용
4	5핀 USB 케이블	보조배터리
5	모니터	개발용
6	키보드 마우스	입력용
7	스마트폰	Wifi 및 화면

로그램을 통하여 운영체제가 설치된다(www.raspberrypi.org).

웹 서버를 이용하기 위하여 아파치를 리눅스 환경에서 설치하고 리눅스 R 프로그램을 설치하므로 텍스트 마이닝 개발환경을 준비한다. 파이썬(Python)은 기본적으로 탑재되어 첫 부팅과 동시에 사용 가능한 개발환경을 제공한다.

본 연구는 최저가 휴대용 서버 시스템을 구축하여 웹마이닝 처리가 가능한 빅데이터 분석 시스템을 구축하고자 한다. 데이터 수집과정과 분석과정을 웹 환경에서 일원화하여 실시간 처리가 가능한 프레임워크를 연구하였다. 자료의 수집과정인 크롤링(Crawling)과 불필요한 단어, 불용어 제거 등 전처리 과정을 거쳐 마이닝 실제 재료인 명사 및 표준어를 추출한다. 명사의 빈도를 활용하여 워드 클라우드(Word Cloud) 분석을 진행하였다. <그림 4>는 본 연구의 실시간 이슈분석 시스템의 프로세스이며 모바일 웹 브라우저 앱에 최적화되어 설계되었다. 아파치 설치와 PHP(Hypertext Preprocessor) 언어를 사용하여 동적 웹 페이지를 구축하여 연구 시스템을 구성하였다. HTML, JavaScript, jQuery, Ajax 등의 스크립트 언어를 사용하여 사용자 인터페이스를 구축하였으며 인터넷 뉴스의 텍스트 분석엔 R program을 사용하였다.



<그림 4> 본 연구의 프레임워크

```
from selenium import webdriver #selenium 웹드라이버 사용을 위한 모듈 임포트
class CrawlNews():)     # ------- ㉮의 해당 함수
  cplist = {A신문, B신문, C신문, ...}       #언론사번호, 테이블명을 인수로 받음
  def __init__(self, sch_text, fdate, edate):
    self.sch_text = sch_text       # 검색어
    self.fdate = fdate             # 검색 시작 날짜
    self.edate = edate             # 검색 끝 날짜
    self.driver = webdriver.PhantomJS(executable_path='/usr/bin/phantomjs') #PhantomJS로 사용
    self.newsurl=[] # 뉴스 링크 추가전 초기화
    self.crawl_url() #신문사별로 기사 전체 가져오기 함수 호출

  def crawl_url(self):    #신문사별로 기사 링크 전체 가져오기 # ------- ㉯ -------
    try:
      for press,pnum in self.cplist.items():
      #각 신문사별로 하나씩 기사 가져오기
        for i in range(1,4):
          self.search_url = "검색어와 날짜를 포함하여 검색할 뉴스 url"
          self.driver.get(self.search_url) #뉴스 url로 이동
          #검색결과에서 기사 리스트를 가지고 있는 ul의 li를 찾기
          links_ul = self.driver.find_elements_by_css_selector("ul li")
          #li 없으면(기사 리스트가 없으면) whlie 끝내기
          if not self.driver.find_elements_by_css_selector("ul#clusterResultUL li") :
            break
          #검색 결과 한 페이지당 기사 링크를 가져와서 newsurl리스트에 저장
          for i in links_ul:
            self.newsurl.append(i.find_element_by_css_selector('a').get_attribute("href"))
      self.crawl_articles(self.newsurl) #뉴스 내용 가져오는 함수 호출
    except:
      pass
                                                    # ------- ㉰ -------
  def crawl_articles(self, newslist):        #뉴스 링크 newsurl리스트를 인수로 받아서 기사 내용 가져옴
    f1 = open("abcd.txt", 'a') #기사내용 저장할 텍스트파일
    for nurl in newslist:
      self.driver.get(nurl)
      title = self.driver.find_element_by_xpath('h3').text #기사 제목
      contents = self.driver.find_element_by_id('contents').text #기사 내용
      f1.write(contents) #기사 내용 저장

CrawlNews(sch_text, fdate, edate)                    # ------- ㉮ -------
```

<그림 5> 파이썬을 이용한 뉴스 크롤링 알고리즘

<그림 4>의 프레임워크를 살펴보면 사용자의 니즈에 부합된 키워드를 선정하여 검색을 실행한다. 관련 기사들은 웹 마이닝을 통하여 크롤링 작업이 이루어진다. 크롤링은 Python 프로그램을 통하여 이루어졌으며, 인터넷 뉴스 웹 페이지의 “article_body” 클래스를 찾아서 “plaintext”만을 추출하여 데이터베이스에 관련 기사를 축적하였다. 같은 방법으로 관련 키워드 기사들을 반복하며 DB에 저장하여 자료 수집을 실시간으로 진행하였다. 크롤링 다음 과정은 기사 내용을 전처리 과정을 거쳐 명사 추출과정을 진행하며 빈도 분석을 진행하였다. 1차 분석된 빈도 분석을 통하여 워드 클라우드 분석을 통하여 시각화 작업을 진행하였으며 워드 클라우드 결과 이미지가 해당 웹 페이지로 업로드 되면 모든 작업이 완료된다. 텍스트 마이닝(Text Mining)은 비정형 텍스트를 기반으로 의미 있는 명사를 추출하는 기술이며 분석 결과를 나타내기 위한 한가지 기법으로 워드 클라우드를 많이 활용한다. 워드 클라우드(Word Cloud)는 단어분류 또는 문법적 구조분석 등의 자연언어 기술에 기반하여 문서의 단어들을 분류하여 그 빈도를 한눈에 보기 쉽게 나타낸 시각화된 결과물이다.

<그림 5> webdriver를 이용한 PhantomJS환경을 선언하여 구현한 인터넷 뉴스 기사 알고리즘이다. Webdriver는 사람이 실제 웹 브라우저를 클릭하고 드래그하듯이 컴퓨터 환경에서 자동으로 시뮬레이션을 지원

<그림 6> 본 연구의 분석 결과1

하는 모듈이며 PhantomJS는 화면으로 표현하는 것이 아니라 백그라운드 작업이 가능하게 하여 크롤링 속도를 높이는 역할로 사용하였다. ㉮의 표시에서 출발하여 클래스 프로그램으로 키워드인 뉴스 검색어, 검색할 범위인 시작날짜와 끝날짜 를 입력하여 CrawlNews 함수를 호출하는 구문이다. ㉯는 검색한 관련 키워드가 있는 뉴스 링크를 가져오는 함수이며 ㉰ 함수는 ㉯ 함수에서 전달 받은 기사 링크에 접속하여 관련 기사의 제목과 내용을 크롤링하는 함수로 구성되어 있다.

4. 연구 분석 및 결과

수많은 단어들에서 패턴을 찾고자 하는 연구가 활발히 진행되고 있다. 데이터의 양에 비례하여 수집 시간, 분석 시간 또한 이원화 되어 급변해가는 여론을 예측하기엔 어려움도 많다(임좌상•김진만, 2014; 이철성 외, 2013).

본 연구는 이미 설명 하였듯이 웹 환경에서 자료 수집과 분석이 실시간으로 이루어져 시각화를 확인할 수 있는 웹 페이지로 구현하였다. 서버의 개발자 환경을 이용하여 빅데이터 분석 킷(kit)을 구성함으로 관련 연구자들이 직접 시스템을 구성할 수 있는 기회를 마련한 연구라 본다. 본 실험은 2017년 12월 1일에서 10일까지 “산업혁명”의 키워드로 실시간 뉴스 기사를 분석하고자 하였다. <그림 6>은 크롤링 된 뉴스 기사를 활용한 워드 클라우드 분석의 결과이다.

모든 산업에서 빅데이터 활용은 고객의 니즈 분석에 큰 역할을 하고 있다. 일하는 방식과 새로운 일자리, 비즈니스 전략 등 인류의 대변혁을 예고하고 있다. <그림 6>를 살펴보면, “산업”, “기술”, “기업”, “로봇”, “교육”, “일자리”, “인공지능”, “개발”, “전략”, “서비스산업”, “창업자”, “혁신” “빅데이터” 등의 키워드가 상위에 랭킹되어 있으며 “산업혁명”과 인접하게 나열된 것을 확인할 수 있다.

<그림 7>은 10월 1일에서 10일간 “기업”이란 키워드를 검색어로 하여 나타난 워드 클라우드 결과이다. “특허”, “사업”, “산업”, “경제”, “향균”, “기술”, “제품”, “중소기업”, “금융”, “생활용품”, “출원”, “국정감사“ 등이 기업 중소기업, 제품, 개발, 특허, 기술, 생활용품, 출원, 품질 등 기업 관련 뉴스 기사에서의 빈도 결과이다. 기업들의 “특허” 관련 이슈와 국정감사로 인한 “의원”, “향균”과 관련된 유해 세균을 제거하기 위한 “수세미”, “바이오”, “지퍼백”, “비누” 등 “상품” 관련 키워드로 노출된 것으로 나타났다.

5. 결론

본 연구는 신용카드 크기의 라즈베리 파이를 활용하여 웹 서버 구축과 인터넷 신문을 활용한 실시간 분석 시스템을 설계 및 구현, 그리고 실험까지 진행하였다. 그 결과를 종합적으로 정리해보면 다음과 같다.

초소형 PC이며 교육용 컴퓨터로 활용되고 있는 라즈베리 파이에 운영체제인 라즈비안을탑재하여 GUI 환경의 사용자 인터페이스로 구동환경을 이용하였다. 오픈 소스 계열의 웹 서버인 Apache를 활용하여 연구 PC 이외의 디바이스에서 접속이 가능한 웹 서버환경을 구현하였다. 웹 마이닝을 통한 웹 페이지 크롤링은 라즈비안에서 기본으로 제공하는 Python3.5를 활용하여 selenium 웹 드라이버를 설치하였다. Webdriver의 PhantomJS를 이용하여 웹 페이지 크롤링을 자동화하기 위해 연구자의 수작업 과정을 마치 컴퓨터가 수행하도록 처리하는 기능을 이용하였다. 수집된 데이터의 실시간 분석과 시각화 과정은 리눅스 R를 활용하여 진행하였으며 웹 페이지 구현은 동적인 웹 문서를 빠르고 쉽게 작성할 수 있는 PHP로 개발하여 빠른 웹 환경을 제공하였다. 빅데이터의 연구와 교육을 위한 플랫폼의

<그림 7> 번 연구의 분석 결과2

개발은 많은 비용과 복잡한 개발 환경의 지식이 필요하였다. 본 연구는 오픈 소스 하드웨어 시스템 구축과 오픈 소스 소프트웨어를 활용하여 저비용의 교육용 빅데이터 분석 플랫폼을 설계하고 실제 개발함으로써 빅데이터 분석을 위한 시스템 설계에 그 목적을 두고 있다. 오픈 소스 하드웨어, 소프트웨어 기술을 활용하여 웹 환경의 데이터를 크롤링하는 수집 과정과 자연어 처리를 통한 분석과 시각화 과정을 실시간으로 처리하기 위한 문제를 해결하고자 노력하였다. 라즈베리 파이의 개발 환경과 빅데이터 분석이 가능한 콘텐츠 기술의 조합으로 빅데이터 분석 킷(kit)를 제공하므로 마이닝 처리가 가능한 서비스 제품화까지 고려할 수 있다.

본 연구는 저비용 연구환경과 실용성을 앞세워 오픈 소스 하드웨어, 소프트웨어를 활용한 빅데이터 분석 플랫폼 개발 환경을 설계하고 있다. 고성능 컴퓨터를 앞세워 빠른 처리 속도와 처리량을 측정하기 위한 시스템 설계가 아니라 연구 환경을 보다 교육적 관점에서 접근하여 빅데이터 산업의 자유로운 연구와 연구 기반을 조성하기 위한 시스템을 설계하였다. 인문사회계열의 연구자들이 웹 서버 시스템을 자유롭게 운영할 수 있는 연구 환경을 직접 조성할 수 있으며, 휴대가 용이하여 이동이나 간편 설치가 가능하다. 또한, 분산 처리 시스

템 구현이 이공계 연구자들의 영역을 벗어나 ICT 비전문 연구자들에게도 쉽게 실습과 연구를 병행할 수 있는 환경으로 발전할 수 있다.

저 비용의 웹 서버 환경이지만 웹 마이닝 연구 환경의 다양한 가능성을 보여주었다. 인터넷 뉴스뿐만 아니라 SNS 분석 과정도 함께 진행한다면 빈도수 분석을 통한 향후 추세, 트렌드 변화, 선호순위와 특정 이슈의 여론이나 의견을 가공하여 감정적 통계치를 이용한 오피니언 마이닝, 상권 분석, 여론조사 등 빅데이터 분석 시스템으로 활용가치를 확인하였다. 또한, 데이터들의 분류와 검색이 용이한 SNS 해시태그를 활용하여 태그의 감성 분류에 대한 연구로 이어져 감성 분석 시스템으로 확장이 가능하다.

하드웨어와 소프트웨어 환경을 쉽게 이해하고 구현하므로 빅데이터 연구 환경의 접근성에 중점을 둔 시스템 설계이다. 운영체제인 라즈비안, 개발 프로그램인 Python, R, PHP 등 모든 SW는 오픈 소스로 구성되어 있으며 누구나 개량하고 재배포 가능하며 무상으로 공개 및 공유할 수 있는 4차 산업혁명의 협업을 기본으로 하고 있다. 오픈 소스 하드웨어인 라즈베리 파이 보드의 부품이 자유롭게 교체되는 호환성과 다양한 사용자 인터페이스로 사물을 네트워크로 연결하는 있는 확장성을 확인하였다.

새로운 산업혁명에서 이끌고 있는 강력한 무기는 데이터이다. 기존 빅데이터 연구자들의 데이터 수집과 분석의 이원화된 연구환경에서 데이터 수집의 다양한 시도가 자유롭게 됨으로 현장 연구가 가능해진다. 웹 미디어의 실시간 자료 수집과 분석을 웹 서버에서 제어할 수 있는 웹 환경을 개발하여 IT 비전공자들의 커리큘럼으로 제시되어 연구자들의 현장 연구의 확산이 기대된다.

참고문헌

[국내 문헌]

1. 김세민, 최숙영, 2017, “공업계 특성화 고등학생을 위한 라즈베리파이를 활용한 프로그래밍 수업 방안,” *한국정보통신학회논문지,* (21:1), pp. 165-172.
2. 김영근, 김승현, 조민희, 김원중, 2014, “학습 시스템을 위한 빅데이터 처리 환경 구축,” *한국전자통신학회 논문지*, (9:7), pp. 791-797.
3. 김영근, 조민희, 김원중, 2016, “라즈베리파이 보드 기반의 빅데이터 분석을 위한 학습 시스템,” 한국전자통신학회 논문지, (11:4), pp. 433-439.
4. 김은우, 금득규, 2014, “빅데이터 분석: Social BigDate 서비스 분석플랫폼 구축기술 소개,” *정보처리학회지,* (21:3), pp. 35-42.
5. 김윤경, 2017, “제 4 차 산업혁명 시대의 국내환경 점검과 정책 방향,” *한국경제연구원*, pp. 1-16.
6. 김정선, 권은주, 송태민, 2014, “분석지의 확장을 위한 소셜 빅데이터 활용연구-국내” 빅데이터” 수요공급 예측,” *지식경영연구*, (15:3),pp. 169-188.
7. 박성원, 2017, “새로운 과학기술이 일하는 방식에 미치는 영향,” *과학기술정책*, (27:7), pp. 26-31.
8. 서병조, 나성욱. 2017, “지능정보화 시대에 대비한 네트워크 발전전략 연구,” *한국통신학회지*(정보와 통신), (34:8), pp. 30-37.
9. 서새남, 2017, “4 차 산업혁명 주요기술에 대한 법적 고찰-한국 및 중국을 중심으로,” *문화· 미디어· 엔터테인먼트 법*, (11:1), pp. 141-172.
10. 조성룡, 2012, “데이터 환경의 변화와 분산 데이터베이스 시스템,” *정보과학회지,* (30:5), pp. 21-28.
11. 안정국, 김희웅, 2015, “Building a Korean Sentiment Lexicon Using Collective Intelligence,” *지능정보연구*, (21:2), pp. 49-67.

12. 이종화, 이현규, 2017, “해시태그를 이용한 실시간 연관 규칙 분석 연구,” *인터넷전자상거래연구*, (17:4), pp. 105-117.
13. 이철성, 최동희, 김성순, 강재우, 2013, “한글 마이크로블로그 텍스트의 감정 분류 및 분석,” *정보과학회논문지: 데이터베이스*, (40:3), pp. 159-167.
14. 임좌상, 김진만, 2014, “한국어 트위터의 감정 분류를 위한 기계학습의 실증적 비교,” *멀티미디어학회논문지*, (17:2), pp. 232-239.
15. 정민영, 2016, “포털사이트 실시간 검색키워드의 주간 핵심 이슈 선정 및 차이 분석,” *디지털융복합연구*, (14:12), pp. 237-243.
16. 조재희, 2004. “OLAP 과 데이터마이닝을 이용한 조직내 분석지 생성에 관한 사례연구,” *지식경영연구*, (5:1), pp. 69-82.
17. 황보람, 김성규, 2016, “라즈베리파이를 이용한 빅데이터 처리 학습 환경 구축,” *디지털융복합연구*, (14:4), pp. 251-258.
18. Lee, J. H. and Lee, H. K., 2015, “A Study on Unstructured Text Mining Algorithm through R Programming based on Data Dictionary,” *Journal of the Korea Society Industrial Information System*, (20:2), pp. 113-124.
19. Lee, J. H., Le, H. S., and Lee, H. K., 2016, “Research on Methods for Processing Nonstandard Korean Words on Social Network Services,” *Journal of the Korea Industrial Information Systems Research*, (21:3), pp. 35-46.

[국외 문헌]

1. Chen, H., Chiang, R.H. and Storey, V.C., 2012. “Business intelligence and analytics: from big data to big impact.,” *MIS quarterly*, pp. 1165-1188.
2. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996, “From data mining to knowledge discovery in databases,” *AI magazine,* (17:3), pp. 37-54.
3. Liu, B., 2013, *Web Data Mining, 2nd edition, Springer.*
4. Sivakumar, P., 2015, “Effectual Web Content Mining using Noise Removal from Web Pages,” *Wireless Personal Communications*, (84:1), pp. 99-121.
5. Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J., 2016, *Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann.*
6. Wu, X., Zhu, X., Wu, G.Q. and Ding, W., 2014. Data mining with big data. *IEEE transactions on knowledge and data engineering*, (26:1), pp. 97-107.
7. Zhang, Q. and Segall, R. S., 2008, “Web mining: a survey of current research, techniques, and software,” *International Journal of Information Technology & Decision Making,* (7:4), pp. 683-720.

저 자 소 개

이 종 화 (Jong-Hwa Lee)

현재 부경대학교 정보전산원 강사로 재직 중이다. 부경대학교에서 경영학 박사 학위를 취득하였다. 주요 관심분야는 빅데이터, 웹마이닝, 감성분석 등이다. 지금까지 Journal of the Korea industrial Information Systems Research, The Journal of Information Systems, The Journal of Internet Electronic Commerce Research 등 주요 학술지에 빅데이터 및 마이닝 관련 논문을 발표하였다.

이 현 규 (Hyun-Kyu Lee)

현재 부경대학교 경영대학 경영학부 교수로 재직 중이다. 연세대학교에서 경영학 박사 학위를 취득하였고, 학위취득 후 아더앤더슨 비즈니스컨설팅에서 다년간 컨설턴트로서 활동하였다. 주요 관심분야는 정보시스템전략, 빅데이터 분석 등이다. 지금까지 The Journal of Internet Electronic Commerce Research, The Journal of Information Systems, Journal of the Korea industrial Information Systems Research, Korean Management Review, Information Systems Review 등 주요 학술지에 다수의 논문을 발표하였다.