Journal of Knowledge Information Technology and Systems(JKITS), Vol. 15, No. 1, pp. 103-111, February 2020
https://doi.org/10.34163/jkits.2020.15.1.010

**Journal of Knowledge Information Technology and Systems**

ISSN 1975-7700

http://www.kkits.or.kr

# A Hybrid Approach Using Python and Excel for Extracting Valid Cases in Content Analysis

**Dong-Sung Kim**[1], **Hyung-Suk Kim**[*2]

[1]*Department of Advertising and PR, Hanyang University*
[2]*Department of Advertising and PR, Chungwoon University*

## ABSTRACT

Recent big data gives scientists the opportunity to collect and analyze more quantity of data and more variety of data more quickly than they could ever do. 3Vs(Volume, Variety, Velocity), the basic characteristics of big data, can solve social science research's key problems of generalization with research results, accessibility to necessary data, and excessive manpower or time. So, big data is attracting attention as a research method for social scientists. However, many researchers are blind and improperly applying big data without worrying and solving the representativeness and validity of the sample. Therefore, this study proposes a hybrid research method that can maximize the merits of both quantitative and qualitative research methods: generalization and validity. To do this, we researched the trend of fairness-oriented issues related to domestic large companies over a year and crawled 21,900 articles of Naver ranking news with Python using the BeautifulSoup library. And 1,876 final valid data for content analysis were extracted by using the FIND function of Excel. Consequently, we propose an alternative by modeling the process and method for content analysis: crawling news articles by python and filtering valid cases by Excel FIND function. After this research, it is expected that various collaborations will be carried out to create synergy between big data technologies and social science fields.



**KEYWORDS:** Hybrid method, Valid cases extraction, Python, Excel FIND function, Web crawling, Content analysis, Big data

**ARTICLE INFO:** Received 17 January 2020, Revised 6 February 2020, Accepted 7 February 2020.

[*]Corresponding author is with the Department of Advertising & PR, Chungwoon University, 113 Sukgol-ro Michuhol-gu Incheon, 22100, KOREA.
*E-mail address*: kennyg@chungwoon.ac.kr

## 1. 서 론

2020년 1월, 모든 산업 영역에서 개인을 알아볼 수 없게 안전한 기술적 처리(비식별화)를 거친 가명·익명 정보의 경우에는 산업적 연구, 상업적 통계 목적이라면 이를 개인동의 없이 활용할 수 있도록 하는 데이터 3법(개인정보보호법·정보통신망법·신용정보보호법) 개정안이 국회를 통과하여[1], 본격적으로 우리 사회의 모든 분야가 빅데이터(big data)를 활용한 획기적인 변화가 예상된다. 다양한 종류(variety)의 대규모 데이터(volume)를 신속하게 처리(velocity)할 수 있는 빅데이터의 시대는 과학자들의 연구 패러다임도 크게 변화시키고 있다. 특히 소수의 사람들에 대한 관찰 및 질문을 중심으로 진행해 온 대부분의 사회과학자들에게 기존 연구의 수준과는 비교할 수 없을 정도로 더 다양하고 더 많은 데이터를 더 신속하게 수집하고 분석할 수 있는 기회를 제공하고 있다. 이는 사회과학 연구에서 가장 큰 연구 한계로 제기되는 연구 결과의 일반화, 필요 데이터에 대한 접근성, 그리고 연구에 소요되는 인력 및 시간 등의 노력 문제를 해결해 줄 수 있다는 의미이다. 하지만 수집된 모든 빅데이터가 대표성과 타당성이 있다고 가정하고 연구 결과를 제시한다면 오히려 심각한 오류가 발생할 수도 있다. 특히 내용분석을 위한 분석용 데이터는 타당하고 엄격한 필터링을 통해서 준비되어야 이를 극복할 수 있다.

본 연구는 빅데이터를 만능이라고 맹신하고 너무 쉽게 생각하여 부적절하게 적용하고 있는 사회과학자들이 빅데이터를 제대로 활용하여 양적 연구방법과 질적 연구방법 모두의 장점, 즉 일반화와 타당성을 지닌 하이브리드형 연구방법에 대한 가이드를 제시하고자 한다. 이를 위해서 빅데이터, 파이썬 크롤링, 엑셀의 FIND 기능 등에 대한 기본적 이론 배경을 살펴 보고, 1년 동안의 국내 대기업 관련 공정/배려 이슈 트렌드에 대한 내용분석 연구를 하나의 사례로써 수행하면서 데이터 수집과 유효 케이스 추출의 과정 및 방법을 모델링하여 제안한다.

## 2. 이론적 배경

### 2.1 빅데이터

빅데이터란 기존 데이터베이스 관리도구의 능력을 넘어서는 대량의 정형 또는 비정형 데이터로부터 가치를 추출하고 결과를 분석하는 기술을 의미한다[2]. 이 정의에서 '기존 DB 관리도구'라는 개념은 상대적인 기준이기에 이전과 이후를 분리하는 개혁성과 정확한 기준 시점에 대한 모호성이라는 두 가지 의미를 모두 지닌다고 볼 수 있다. 따라서 빅데이터에 대한 보다 명확한 개념을 알기 위해서는 그 특징들을 확인해야 한다. 일명 3V라고 하는 공통적인 기본 특징은 매우 큰 용량(terabyte나 petabyte 정도)의 데이터를 의미하는 Volume(크기), 다양한 출처(source) 및 형식(정형, 반정형, 비정형)의 Variety(다양성), 그리고 데이터의 생성 및 수집의 신속성을 뜻하는 Velocity(속도)이다. 이외에도 Veracity(정확성), Value(가치), Variability(가변성), Visualization(시각화) 등의 'V' 자로 시작하는 특징들을 추가하기도 한다. 이와 같은 특장점으로 인하여 21세기의 원유라고 하는 빅데이터는 기업뿐만 아니라 의료, 정치/정책, 교육 등의 광범위한 분야에서 활용되고 있다[3]. 박대민(2013)은 저널리즘연구에서 기사는 빅데이터화되고 그로 인해 기존의 방법론을 빅데이터 분석과 접목시키는 작업이 필요하며, 빅데이터 방법에 기초하여 이론을 도출하는 작업이 필요하다고 주장한 바가 있다[4]. 이처럼 '기존'의 데이터와는 다른 빅데이터의 개혁성은 기존의 연구자들, 특히 사회과학자들에게도

새로운 연구 방법과 성과의 기회를 제공한다고 볼 수 있다.

## 2.2 빅데이터 처리과정과 기술

빅데이터 처리 과정과 각 단계별 기술은 다음과 같이 요약정리할 수 있다[5-8]. 우선 내부데이터(자체 DBMS), 외부데이터(소셜미디어, 공공데이터와 같은 데이터웨어하우스 등), 그리고 미디어(이미지, 영상 등) 중 원하는 데이터 소스를 결정하고, 수동 또는 자동(로그수집기, 크롤링, 센싱)으로 데이터를 수집・통합한 후, 정형/반정형/비정형 그리고 등급화한 데이터를 저장한다. 그리고 데이터셋에서 분석을 위해서 필요한 데이터 정제처리과정을 거친 후, 자연어 처리에 기반을 둔 텍스트 마이닝, 모든 웹문서와 댓글 등에서 의견을 수집・분석해 평판을 추출해 내는 평판 분석, 소셜 네트워크 내 영향력, 관심사, 성향 및 행동 패턴을 추출하는 소셜 네트워크 분석, 데이터 간의 유사도 및 거리에 기반을 둔 클러스터 분석 등 다양한 분석 방법들을 적용한다. 마지막으로 분석 결과를 그래프, 스프레드 시트, DB, 인포그래픽 등 여러 형태의 시각적 표현으로 제시한다.

이러한 빅데이터 처리 과정 중에서 본 연구는 연구자들이 분석을 위해서 대용량 데이터를 수집하고 필요한 정제된 유효 데이터를 보다 합리적이고 타당하게 준비하는 과정까지를 사회과학적 관점에서 제시하고자 한다.

## 2.3 파이썬 웹크롤링

보통 빅데이터로부터 유용한 정보를 추출하여 가치를 창출하기 위해서는 데이터의 추출과 변형 그리고 가치 창출을 위한 새로운 데이터베이스로의 로딩(Extraction, Transformation, and Loading) 단계가 필요하다. ETL 구현 시 컴퓨터 내의 수많은 문서를 수집하는 기술을 크롤링(crawling)이라고 하고 웹환경에서 각종 정보를 자동 수집하는 기술을 웹크롤링이라고 하는데, 최근에 큰 주목을 받고 있다[9,10]. 이는 사람이 직접 하기엔 비효율적인 반복적 작업을 컴퓨터가 신속하게 대신해주는 큰 이익을 주기 때문이다. 웹크롤링을 위한 다양한 개발 도구들이 있는데, 최근 가장 각광받고 성장하고 있는 언어는 오픈소스 소프트웨어인 파이썬(python)이다[11]. 파이썬은 컴퓨터 프로그래밍이 익숙하지 않은 비전공자들인 인문학이나 통계 분야 종사자들이 쓰기 쉽도록 여러 라이브러리들이 발달하면서 급격히 발전하고 있다[12].

이번 연구는 파이썬으로 웹에서 연구자가 원하는 대량 데이터를 자동 크롤링하는 과정 및 방법을 프로그래밍 비전문가들도 어렵지 않게 수행할 수 있도록 안내하고자 한다.

## 2.4 엑셀 FIND 기능

엑셀의 가장 대표적인 기능인 엑셀 함수는 수식(예약어)에 의해서 반복적이고 복잡한 일련의 계산, 변환, 정렬, 검색 등의 기능을 쉽고 빠르게 수행하는 하나의 약속이라고 할 수 있다. 웹크롤링을 통해서 원자료(raw data)를 수집하였지만, 각 케이스들이 연구의 적정 유효 데이터인지를 확인하기 위해서는 필터링 및 검증 과정이 필요하고 이때 유용하게 사용할 수 있는 엑셀 기능이 바로 FIND 함수이다. 이를 통해서 특정 데이터 영역에서 원하는 검색어나 숫자를 쉽고 빠르게 찾고 그 데이터의 유효성을 확인할 수 있다. 물론 검색 키워드, 순서, 그리고 조건 등에 대한 합리적이고 효율적인 노하우도 필요하다.

# 3. 크롤링과 필터링 과정 및 방법

### 3.1 내용분석 연구의 개요 및 사전조사

본 연구는 파이썬을 이용한 웹크롤링과 엑셀 FIND 함수를 이용한 유효데이터 필터링 과정 및 방법을 기업의 공정성 관련 이슈 트렌드 분석이라는 PR 분야의 연구 과정을 통해서 제시하고자 한다. 이는 긍정/부정 이슈 유형이나 상황적 위기 커뮤니케이션 이론(Situational Crisis Communication Theory)에서 제시한 책임성에 따른 위기 이슈 유형[13]과 같은 기존의 이슈 유형 분류에서 벗어나, 배려보다는 공정에 더 큰 가치 무게를 두고 있는 최근의 사회적 현상에 주목하여 국내 기업들에 대한 사회적 이슈 트렌드를 공정과 배려라는 새로운 이슈 유형 분류를 적용하여 살펴보고자 하는 연구이다. 2018년 10월 사전조사에서는 네이버의 '많이 본 뉴스' 중에서 화제성이 보다 크다고 할 수 있는 TV섹션 Top 31 뉴스를 대상으로 2018년 4월 3일부터 15일 간 총 558개 뉴스 중 유효 63개(2일 중복게재 23개)의 데이터를 수동으로 기사 하나하나 직접 수집하여 내용분석을 하였다[14,15]. 하지만 당시 적은 표본이 갖는 일반화 문제와 자료수집 과정에서의 너무 많은 시간의 소요 문제가 있었기에 본조사에서는 이를 극복할 수 있는 대안으로서 빅데이터 웹크롤링 기술 적용의 필요성을 절실하게 깨닫게 되었다.

### 3.2 데이터 소스의 결정

본조사에서의 이슈 내용분석을 위한 자료는 국내 대표 포털사이트인 네이버의 랭킹뉴스 중 '많이 본 뉴스'의 경제와 사회 두 섹션에서 각각 매일 30개씩 1년 치(각 기사에 대한 view 수가 확인가능한 2018년 4월 3일부터 2019년 4월 2일까지)의 기사들이다. 국내의 압도적인 1위 포털사이트에서 경제와 사회 분야에서 일간 조회수 Top 30의 기사들만을 선정하였기에 한국 사회의 최대 이슈로서의 대표성을 갖는 표본이라고 할 수 있다.


그림 1. 네이버의 랭킹뉴스 중 많이 본 뉴스(경제/사회 섹션)
Figure 1. Ranking news in Naver(economy/society section)

또한 신문기사는 비정형 데이터로써 다양한 데이터 포맷에 대한 메타데이터 검색, 중복 정보 및 노이즈를 제거하기 위한 기술, 개체 추출 정확도 제고를 위한 기술 등이 요구되는데[16], 네이버의 뉴스는 대부분의 신문기사들을 통합·표준화하여 데이터를 제공하기에 크롤링과 분석에 더욱 유용하다. 따라서 공공의 이슈를 다루는 PR학 분야에서는 네이버의 랭킹뉴스(많이 본 뉴스)는 대표성을 갖춘 최적의 이슈 빅데이터 소스라고 할 수 있다.

### 3.3 데이터 웹크롤링

연구에 필요한 기사들은 네이버 공개 API(Open Application Programming Interface)를 활용하여 파이썬으로 크롤링했는데, JS/CSS, DOM 등 형식에 구애받지 않고 편리하게 크롤링할 수 있는 셀레니움 웹드라이버(Selenium-WebDriver)를 이용하여 크롬 브라우저를 가상 웹드라이버로 설정하고 네이

버 랭킹뉴스 기사의 URL을 추출하여 파이썬의 대표적인 라이브러리인 뷰티풀수프(BeautifulSoup)를 이용해서 데이터를 수집하였다.

```
from selenium import webdriver
from bs4 import BeautifulSoup
import time
import datetime
import numpy as np
import csv
분야 = '101' ###경제섹션, 사회섹션은 102###
크롬경로 = 'C:/python/chromedrive.exe"
start_date = '2018-04-03'
end_date = '2019-04-02'
저장할경로 = 'C:/python/navernews/'

options = webdriver.ChromeOptions()
options.add_argument('window-size=1920x1080')
options.add_argument("user-agent=Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_6) ₩
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.100 Safari/537.36")

navernews = webdriver.Chrome(r"C:/python/chromedriver.exe",options=options)
기본경로 = r'https://news.naver.com/main/ranking/popularDay.nhn?rankingType=popular_day&sectionId={}&date={}'.format
총뉴스 = list()

now_date = datetime.date(int(start_date[:4]),int(start_date[5:7]),int(start_date[8:]))
stop_date = datetime.date(int(end_date[:4]),int(end_date[5:7]),int(end_date[8:]))
while now_date != stop_date:
 try:
 날짜 = now_date.isoformat()
 navernews.get(기본경로(분야,날짜.replace('-','')))
 time.sleep(4)
 많이본뉴스 = BeautifulSoup(navernews.page_source,'html.parser')
 for count, one in enumerate(많이본뉴스.find_all('div',class_='ranking_text')):
뉴스날짜 = 날짜
뉴스순서 = count
뉴스링크 = one.find('a')['href']
뉴스제목 = one.find('a')['title'],
뉴스매체 = one.find('div',class_='ranking_office').text
                    .....<후략> ....
```

그림 2. 네이버 랭킹뉴스 크롤링 알고리즘(일부)
Figure 2. Crawling algorithm for Naver ranking news articles (the front part)

참고로 네이버는 지속적으로 크롤링 방지 기술을 적용하기에 연구를 위한 파이썬의 크롤링 알고리즘도 계속 업데이트해야만 실제 이용이 가능하다. 따라서 <그림 2>에서 일부 제시하고 있는 파이썬 알고리즘은 이번 크롤링 작업에서만 유의하기에 이를 참고하여 이후 연구에서는 새로운 알고리즘을 작성하여 적용해야 한다.

자동 수집된 초기 데이터는 총 21,900개 뉴스(=30개뉴스×365일×2섹션)의 케이스번호, 섹션명, 게시일, 日랭킹, 기사제목, 기사내용 앞부분, 언론사명, View 수, 공감지표들(좋아요, 훈훈해요, 슬퍼요, 화나요, 후속기사 원해요 등), 그리고 댓글지표들(성별, 연령별) 등으로 구성된다.


그림 3. 수집된 원자료와 IF함수를 활용한 필터링
Figure 3. Raw data and filtering by excel IF function

## 3.4 유효케이스 필터링

약 17년간 80,428,892건의 네이버 뉴스 빅데이터 분석 연구처럼[17] 수집된 빅데이터에 대한 단순 자동 빈도분석이나 연관어 분석의 경우는 큰 문제가 아니겠지만, 연구자들이 수동으로 21,900개나 되는 뉴스들을 내용분석한다는 것은 감당할 수 없을 정도의 엄청난 인내와 시간을 요구한다. 따라서 그 중에서 연구의 대상이 되는 유효케이스를 선정하는 합리적인 다음 단계를 거쳐야 한다. 이는 연구의 모집단에 따라서 달라지는데, 본 연구에서는 한국 경제에서의 중요성과 조직 규모의 동질성을 고려하여 자산 총액 16조 원 이상인 재계 순위 20위까지의 대기업(공정거래위원회, 2019년 5월 15일 발표)을 내용분석 대상으로 설정했다. 그리고 기업보다는 이슈 속성이 더 중요하기에 산업군 전체에 대한 기사는 제외하고 해당 기업이 단독 또는 핵심 조직으로서 게시된 기사들만을 추출하였다.

유효 케이스는 엑셀 프로그램에서 IF 함수를 이용하여 기업집단명(그룹명), 기업집단 유사명, 동일인명(대표명)과 주요 인물명, 그리고 사건/이슈명 등으로 검색하였다. 이를테면, 삼성과 관련한 기사를 추출하기 위해서 '기사제목' 데이터값에서 기업집단명인 '삼성', 동일인명인 '이재용', 그리고 주요 인물인 '이건희' 등을 검색하는 FIND 함수로 해당 셀에 "=IF(COUNT(FIND({"삼성", "이재용", "이건희"},F2)),"삼성",0)" 이라고 입력하고 이를 모든 열(F2에서 F21901까지)에 적용시켜서 삼성과 관련한 뉴스들만 필터로 선택하여 유효기사 해당여부를 수동으로 직접 판단하였다. 기사제목에서 추출이 안되는 경우도 있을 수 있어서 '기사내용 앞부분' 데이터값에서도 동일한 방법을 적용하였다. 참고로 검색 시 2가지 유용한 팁을 제시하자면, 삼성과 같이 상대적으로 큰 규모의 조직이거나 한진(대한항공)과 같이 사회적 큰 이슈가 된 조직처럼 관련 기사가 많을 것 같은 조직보다는 적은 조직을 먼저 검색해서 소수의 케이스를 쉽게 걸러내고, "물컵(한진)", "분식회계(삼성)" 처럼 특정 사건과 관련한 키워드를 통해서 중복되지 않는 케이스를 우선 추출하면 혼동없이 더 빠르게 분류할 수 있다.

표 1. 유효케이스 필터링을 위한 주요 검색어와 결과
Table 1. Filtering keywords and results for valid cases

| 기업집단명 | 주요유사명 | 동일인/주요인물 | 기사수 |
|---|---|---|---|
| 삼성 | | 이재용<br>이건희 | 540 |
| 현대자동차 | 현대차<br>현대車<br>기아차 | 정몽구<br>정의선 | 164 |
| 에스케이 | SK | 최태원 | 57 |
| 엘지 | LG | 구광모<br>구본무 | 77 |
| 롯데 | | 신동빈<br>신격호, 신동주 | 46 |
| 포스코 | POSCO | ㈜포스코<br>권오준, 최정우 | 15 |
| 한화 | | 김승연 | 24 |
| 지에스 | GS | 허창수 | 0 |
| 농협 | | 농협협동조합중앙회<br>김병원 | 3 |
| 현대중공업 | 현대중<br>현대重 | 정몽준<br>강환구 | 45 |
| 신세계 | SSG<br>이마트 | 이명희<br>정용진 | 19 |
| 케이티 | KT | ㈜케이티<br>황창규 | 43 |
| 한진 | 대한항공<br>진에어 | 조원태, 서용원<br>조양호, 이명희<br>조현아, 조현민 | 608 |
| 씨제이 | CJ | 이재현<br>이재환 | 28 |
| 두산 | | 박정원<br>박서원 | 7 |
| 부영 | | 이중근 | 4 |
| 엘에스 | LS<br>E1 | 구자홍 | 0 |
| 대림 | e편한세상 | 이준용 | 7 |
| 미래에셋 | | 박현주 | 0 |
| 에쓰오일 | S오일 | 에쓰-오일㈜<br>오스만 알 감디<br>후세인 에이알-카타니 | 2 |
| 중복 | | | 187 |
| 검색어의 본질적 개념과 무관 | | | 441 |
| 해당사항 없음(검색어 불포함) | | | 19,583 |
| 전체 | | | 21,900 |

필터링 과정 중에 삼성과 LG의 디스플레이 경쟁 뉴스처럼 2개 이상의 조직을 주요하게 다룬 기사들(187개)도 있었고, '삼성' 키워드로 검색은 되었지만 실제 기사 내용은 '삼성동 아파트 시세'에 대한 것처럼 연구대상 조직과 관련 없는 개념의 케이스들(441개)도 존재했다. 결과적으로 21,900개 기사 중에서 20대 그룹과 관련한 유효 기사는 1,876개(약 8.57%)로 최종 추출되었다. 이러한 과정과 결과를 통해서 우리는 빅데이터 크롤링으로 수집한 데이터와 검색어로 단순 추출한 데이터들이 실제 분석에서는 유효하지 않은 데이터일 수도 있음을 확인할 수 있다.

## 4. 결 론

본 연구는 빅데이터 전문가가 아닌 사회과학 분야의 연구자들도 빅데이터 기술을 활용하여 다양한 데이터를 보다 많이 그리고 빠르고 편리하게 수집할 수 있도록 PR 분야의 이슈 트렌드 내용분석 연구 과정을 사례로 제시하였다. 그리고 수집된 대용량 데이터가 가질 수 있는 오류를 점검하기 위해서 유효 케이스를 필터링하는 과정까지 진행하였다. 이를 통해서 사회과학자들에게 양적 연구가 지닌 일반화의 장점과 질적 연구가 지닌 타당성의 장점을 모두 충족할 수 있는 하나의 방법 대안을 제시하고자 하였다.

전체적인 프로세스를 요약하면, 우선 내용분석의 대상이 되는 데이터 소스를 결정하고나서 원하는 데이터값들을 파이썬 코딩으로 크롤링하고 그 중에서 유효한 데이터를 엑셀 프로그램을 활용하여 필터링하는 과정으로 압축할 수 있다. 즉, 이슈 트렌드에 대한 연구이기에 사회적 이슈를 확인할 수 있는 포털사이트 네이버의 랭킹 뉴스를 데이터 소스로 선정한 후, 경제와 사회 섹션의 1년치 기사를 BeautifulSoup 라이브러리를 활용한 파이썬 코딩으로 제목, 기사내용 앞부분, 공감 및 댓글 지표들 등을 크롤링하고, 수집된 원자료 중에서 해당 기업과 관련된 유효한 기사들인지를 엑셀 FIND 함수를 이용하여 필터링 및 검증하였다. 그 결과 21,900개 기사 중 약 8.6%인 1,876개 기사가 내용분석의 유효 데이터로 최종 추출되었다. 결국 데이터 수집을 위한 빅데이터 크롤링과 유효 데이터 추출을 위한 엑셀 필터링 작업을 모두 거쳐야 양적과 질적 차원에서의 제대로 된 분석 데이터를 준비할 수 있다는 것을 확인할 수 있었다.

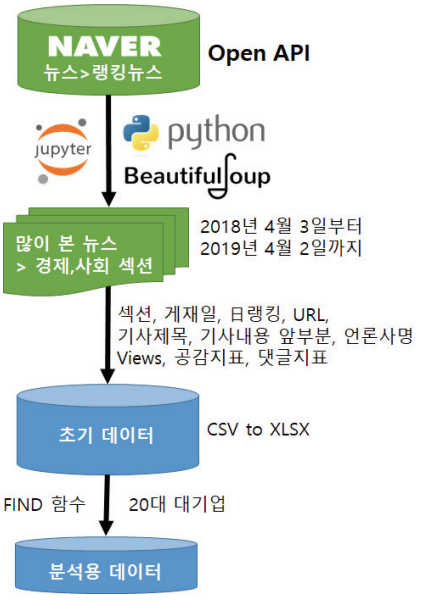


그림 4. 네이버 랭킹뉴스 웹크롤링과 엑셀 필터링 과정
Figure 4. Web crawling and filtering for Naver news articles

그렇지만 본 연구에서 제시하는 데이터 수집 및 유효 케이스 추출 방법만이 유일한 대안이 아니며 빅데이터 보유 주체들도 지속적인 방어 노력을 하기에, 더 다양한 방법의 개발과 업그레이드 및 업데이트가 필요하다. 또한 빅데이터나 프로그래밍 언어 전문가들의 수준에서 보자면, 본 연구에서 제시하는 코딩이나 방법은 그렇게 높은 수준이 아니

라고 할 수 있다. 하지만 광고, PR, 사회학 등의 사회과학 분야의 연구자들에게는 본 연구가 생소하거나 어렵게 느껴져 왔던 빅데이터 기술을 활용하여 연구의 영역을 확장하고 결과에서도 일반화와 타당성 모두를 제고할 수 있는 유용한 가이드가 될 수 있을 것이다. 또한 본 연구를 계기로 빅데이터 분야와 사회과학 분야의 연구가 시너지를 발휘할 수 있는 다양한 협업작업이 이루어지길 기대한다.

이후의 연구에서는 양적과 질적 방법론을 결합했다고 할 수 있는 본 하이브리드형 방법을 활용하여 정치, IT 등 다른 영역에서의 이슈 분석, 10년 단위의 장기적인 이슈 분석, 그리고 매년 정기적인 이슈 트렌드 분석까지 진행하고자 한다. 또한 가능하다면 파이썬이나 R 등의 프로그래밍 자체에서 엑셀의 FIND 기능과 유효케이스 판단 알고리즘을 적용할 수 있는 기술적 대안도 모색하고자 한다.

## References


[1] J-M. Kim, Our future that three data acts w will change, http://www.opinionnews.co.kr/news/articleView.html?idxno=28580, Jan. 2020.

[2] Big data definition, https://ko.wikipedia.org/wiki/%EB%B9%85_%EB%8D%B0%EC%9D%B4%ED%84%B0, Jan. 2020.

[3] K-H. Kim, and Y-J. Kim, *A study on the enterprise's big data utilization*, Journal of Knowledge Information Technology and System, Vol. 14, No. 5, pp. 445-453, 2019.

[4] D-M. Park, *Analysis of news information network as big data analysis method of news articles*, Korean Journal of Journalism & Communication Studies, Vol. 57, No. 6, pp. 234-262, 2013.

[5] Big data Processing and Technologies, https://ikkison.tistory.com/69, May. 2019.

[6] N-J. Kim, and E-I. Choi, *Security technology framework in big data environment*, Journal of Knowledge Information Technology and System, Vol. 11, No. 3, pp. 251-260, 2016.

[7] K-I. Jeong, H-N. Park, B-G. Jung, J-S. Jang, and M-A. Jung, *Big data and information security*, Journal of The Korea Knowledge Information Technology Society, Vol. 3, No. 10, pp. 17-22, 2012.

[8] Y-Y. Joe, *Understanding of big data and major issues*, The Korean Association for Regional Information Society, Vol. 3, No. 16, pp. 43-65, Oct. 2013.

[9] H-S. Kim, N. Han, and S-J. Lim, *Web crawler service implementation for information retrieval based on big data analysis*, Journal of Digital Contents Society, Vol. 18, No. 5, pp. 933-942, 2017.

[10] W-S. Cho, J-E. Lee, and C-H. Choi, *Refresh cycle optimization for web crawlers*, The Journal of the Korean Contents Association, Vol. 13, No. 6, pp. 30-39, 2013.

[11] J-H. Lee, *Building an SNS crawling system using python,* Journal of the Korean Industrial Information Systems Research, Vol. 23, No. 5, pp. 61-76, 2018.

[12] Crawling softwares, https://namu.wiki/w/%ED%81%AC%EB%A1%A4%EB%A7%81, Aug. 2019.

[13] T. Coombs, *Ongoing crisis communication: planning, managing and responding*, Thousands Oaks, CA: Sage, 1999.

[14] D-S. Kim, and H-S. Kim, *Analysis of issue type trends with Korean domestic companies: the crisis-era of fairness value*, Korean

Academic Society for Public Relations, Autumn Conference, pp. 35-35, 2018.

[15] D-S. Kim, and H-S. Kim, *Analysis of fairness issue trends with Korean domestic companies*, Korean Academic Society for Public Relations, Spring Conference, pp. 85-88, 2019.

[16] D-S. Hong, S-D. Jeon, and H-G. Kim, *Analysis of digital forensics technology trends based on big data*, Journal of Knowledge Information Technology and System, Vol. 9, No. 1, pp. 51-63, 2014.

[17] H-Y. Song, and J-H. Yang, *Online news portal service and changes in news distribution: big data analysis of Naver news in 2000-2017*, Korean Journal of Journalism & Communication Studies, Vol. 61, No. 4, pp. 74-109, 2017.

## 내용분석 유효 케이스 추출을 위한 파이썬과 엑셀 활용 하이브리드형 접근법

**김동성[1], 김형석[2]**

[1] *한양대학교 광고홍보학과 겸임교수*

[2] *청운대학교 광고홍보학과 교수*

### 요 약

빅데이터의 기본적인 특성인 3V(Volume, Variety, Velocity)는 연구 결과의 일반화, 필요 데이터에 대한 접근성, 그리고 소요되는 인력 및 시간 등의 노력 문제를 해결해줄 수 있기에 빅데이터는 최근에 사회과학자들에게도 유용한 연구방법으로 각광받고 있다. 하지만 대표성과 타당성에 대한 고민과 해결의 노력 없이 빅데이터를 맹신하고 부적절하게 적용하고 있는 연구자들이 적지 않다. 이에 본 연구에서는 내용분석에서 양적과 질적 연구방법 모두의 장점을 최대화할 수 있는 하이브리드형 연구방법을 제시하고자 한다. 이를 위해서 1년 동안 국내 대기업의 공정/배려 관련 이슈 트렌드를 조사하면서 BeautifulSoup 라이브러리를 이용하여 Python으로 네이버 랭킹뉴스 21,900개 기사를 크롤링하고 그 중에서 재계 20위권 대기업 관련 기사를 핵심 검색어로 엑셀의 FIND 함수를 활용하여 1,876개의 최종 분석용 유효 데이터를 추출하면서 그 과정과 방법을 모델링하여 제안한다.

## 감사의 글

본 연구는 2017년 청운대학교 학술연구조성비 지원에 의하여 연구되었음.

**Dong Sung Kim** received both the M.S.(2000) and the Ph.D.(2014) from the Department of Advertising and Public Relations at Hanyang University. He is currently a representative of a PR agency and a adjunct professor of the Department of Advertising and Public Relations at Hanyang University. His current research interests include communication technologies, new media, and good PRs.

*E-mail address*: prhows@naver.com

Hyung Suk Kim received both the M.S.(2000) and the Ph.D.(2007) from the Department of Advertising and Public Relations at Hanyang University. He is currently a professor of the Department of Advertising and Public Relations at Chungwoon University. His current research interests include big data analysis, public fairness, and authenticity of PR.

*E-mail address*: kennyg@chungwoon.ac.kr