# R기반의 data분석을 통한 사용자 편의성 증진을 위한 웹사이트 설계 및 구현

윤경섭*, 김연홍O
O*인하공업전문대학 컴퓨터정보과
e-mail: ksyoon@inhatc.ac.kr*, yeounhong@nate.comO

# Design and implement Web sites for greater user convenience through R based data analysis

Kyung Seob Yoon*, Yeon Hong KimO
O*Dept. of Computer Science, Inha Technical College

● 요 약 ●

우리 사회는 데이터를 기반으로 진화 하고 있어 데이터분석을 할 수 있는 통계패키지가 오늘날 상용화 되고 있다. 상용화되고 있는 통계패키지를 이용해 본 논문에서는 통계패키지 R을 Model1구조가 아닌 Model2 MVC구조로 적용하여, 웹사이트의 유지보수와 코드 효율성을 증진시키고자 한다. 이를 이용하여 웹 스크래핑을 통한 데이터를 수집 후 데이터 분석을 토대로 사용자가 분석내용을 쉽게 이해할 수 있도록, 편의성을 증진시키고 검색 할 수 있는 웹사이트를 설계 및 구현 하고자 한다.

**키워드:** 웹 스크래핑(Web Scraping),R, Model2, 웹사이트(Web Site), MVC, 데이터분석

## I. 서론

최근 빅데이터에 대한 관심이 높아지면서 오픈소스 기반의 데이터 수집 도구들이 많이 등장하고 있다.[1,2] 그로 인해 통계패키지들은 여러 분야에서 데이터를 분석할 때 사용되고, 오늘날 우리사회는 점차 데이터를 기반으로 하면서 데이터 분석과 활용이 기업 경쟁력에 많은 비중을 차지하고 있다. 그중 통계패키지 R은 공개소프트웨어지만 상용소프트웨어 SAS ,SPSS, Minitab [3] 못지 않게 다양한 확장성과 범용성을 가지고 있다.

예를 들어 R기반 웹사이트가 구현되면서 웹사이트에서 수집된 데이터를 처리하여 분석할 수 있도록 구현되고 있다.[3] 하지만 Model1방식으로 구현되어 JSP를 통해서만 처리되고 있기 때문에 유지보수의 문제점과 코드를 수정할 경우 많은 어려움이 존재한다.

본 논문에서는 통계패키지 R을 MVC구조와 연동 하여, 웹에서도 수집된 데이터를 처리, 분석하는 것 뿐 아니라 유지보수와 코드의 효율성을 높이고자 한다. 또한 사용자에게 데이터 분석내용에 대한 근원지를 알 수 있게 하기위해, 내용의 출처를 밝혀서 이해도를 높여 사용자 편의성을 증진시키는 웹사이트를 설계 및 구현한다.

## II. 관련 연구

R과 JSP를 활용해 웹과 연동되는 데이터 분석시스템이 존재한다. R 프로그램을 서버에 설치하고 클라이언트가 서버에 처리하고자 하는 데이터를 요청하면 서버의 R프로그램이 연산을 하여 결과를 응답해주는 구조이다. 일반적으로 R프로그램을 PC에 설치하여 데이터를 분석하는 것과는 다른 방법으로, 웹을 연동해서 사용한다. 이 방식은 웹페이지와의 연동을 위해서 Rserve 라이브러리를 이용해 구현 된다. Rserve서버의 631포트를 사용하여 데이터 통신후, Java프로그램을 구현 후 JSP가 최종 데이터를 Rserve 객체에 접근하여 로직을 수행하는 방식이다.[4]

그러나 해당 방식은 MODEL1 방식으로 유지보수와 코드의 효율성이 떨어진다.

본 논문에서는 이를 보완할 수 있도록, MODEL2 방식인 MVC 구조를 이용한다. MVC구조는 Model, View, Controller의 세 개의 컴포넌트로 구분하는 아키텍처로, 유저 인터페이스와 비즈니스 로직들을 서로 분리하여 개발하는 방법이다. 이를 이용해 웹 스크래핑을 통해 가져온 데이터를 분석하는 사이트를 설계 및 구현을 제안한다.

## III. 설계

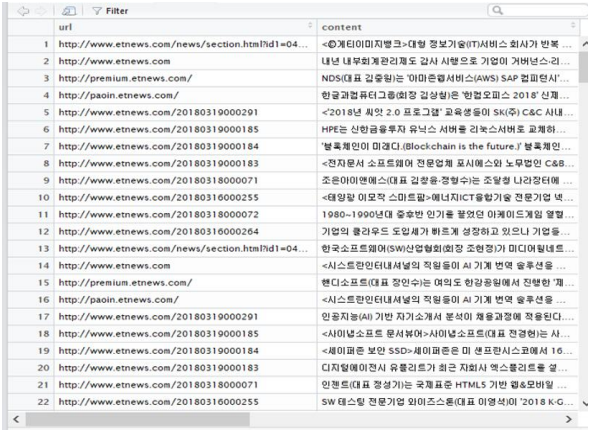
본 논문에서 개발하고자 하는 R기반 데이터 분석 웹사이트는 R을 활용해 데이터를 수집 후, 분석해 사용자가 편리하게 보고, 검색할

수 있도록 MVC구조로 웹사이트 구현하는 것을 목적으로 한다.웹페이지와의 연동을 위해서 R프로그램에서 제공하는 Rserve라이브러리를 사용한다.

Rserve[3]라이브러리는 C, PHP, Java와 같은 개발 언어를 통해 외부 프로그램과 R을 연동하는 TCP/IP 통신 소켓 서버로써, GPL 라이선스 정책과 함께 공개된 라이브러리다. 초기화등의 별도의 설정 없이 통신이 가능하고 통계 연산을 위한 R프로세스의 Back-End 연동을 위해 활용되고 있다.

해당 Rserve를 사용하기 위해 rforge.net에서 공개한 Rserve 라이브러리를 설치 후, Rserve와 데이터 통신을 할 수 있도록 Java 프로그램을 구현한다. 구현한 Java를 Rserve 객체에 접근하기 위해 JRClient API를 사용 후 MVC 구조로 구현한다. MVC 구조가 Rserve 라이브러리를 인식할 수 있도록 pom.Xml에 등록해 웹페이지 연동 설정을 한다.

연동 후 R에서 제공하는 rvest함수를 통해 html웹페이지 구조를 파악해 필요한 데이터의 노드를 확인한다. 확인 후 read_html함수로 해당 주소 html 소스 코드를 웹 스크래핑 해 온다. 필요한 데이터만 추출하기 위해 html_node함수를 이용해 data를 분석해 추출 노드 부분을 찾아준다. 해당 노드에서 html_text 함수를 활용해 text형태로 데이터를 수집한다. 그림1은 해당 방식으로 데이터를 수집해온 모습을 구현한 화면이다.



| | url | content |
|---|---|---|
| 1 | http://www.etnews.com/news/section.html?id1=04... | <㈜케티아미지빅크>대형 정보기술(IT)서비스 회사가 반복 ... |
| 2 | http://www.etnews.com | 내년 내부회계관리제도 감사 시행으로 기업이 거버넌스·리... |
| 3 | http://premium.etnews.com/ | NDS(대표 김중원)는 '아마존웹서비스(AWS) SAP 컴피턴시'... |
| 4 | http://paoin.etnews.com/ | 한글과컴퓨터그룹(회장 김상철)은 '한컴오피스 2018' 신제... |
| 5 | http://www.etnews.com/20180319000291 | <'2018년 최앗 2.0 프로그램' 교육생들이 SK(주) C&C 사내... |
| 6 | http://www.etnews.com/20180319000185 | HPE는 산학금융투자 유닉스 서버를 리눅스서버로 교체하... |
| 7 | http://www.etnews.com/20180319000184 | '블록체인이 미래다.(Blockchain is the future.)' 블록체인... |
| 8 | http://www.etnews.com/20180319000183 | <전자문서 소프트웨어 전문업체 포시에스와 노무법인 C&B... |
| 9 | http://www.etnews.com/20180318000071 | 조은아이앤에스(대표 김창윤·정형수)는 조달청 나라장터에 ... |
| 10 | http://www.etnews.com/20180316000255 | <태양광 이모작 스마트팜>에너지ICT융합기술 전문기업 넥... |
| 11 | http://www.etnews.com/20180318000072 | 1980~1990년대 중후반 인기를 끌었던 아케이드게임 열협... |
| 12 | http://www.etnews.com/20180316000264 | 기업의 클라우드 도입세가 빠르게 성장하고 있으나 기업들... |
| 13 | http://www.etnews.com/news/section.html?id1=04... | 한국소프트웨어(SW)산업협회(회장 조현정)가 미디어링네트... |
| 14 | http://www.etnews.com | <시스트란인터내셔널의 직원들이 AI 기계 번역 숭후션을 ... |
| 15 | http://premium.etnews.com/ | 핸디소프트(대표 장인수)는 여의도 한강공원에서 진행한 '채... |
| 16 | http://paoin.etnews.com/ | <시스트란인터내셔널의 직원들이 AI 기계 번역 숭후션을 ... |
| 17 | http://www.etnews.com/20180319000291 | 인공지능(AI) 기반 자기소개서 분석이 채용과정에 적용된다.... |
| 18 | http://www.etnews.com/20180319000185 | <사이냅소프트 문서뷰어>사이냅소프트(대표 전경헌)는 사... |
| 19 | http://www.etnews.com/20180319000184 | <세이퍼존 보안 SSD>세이퍼존은 미 샌프란시스코에서 16... |
| 20 | http://www.etnews.com/20180319000183 | 디자털에이전시 유플리트가 최근 자회사 엑스플리트를 설... |
| 21 | http://www.etnews.com/20180318000071 | 인젠트(대표 정성기)는 국제표준 HTML5 기반 웹&모바일 ... |
| 22 | http://www.etnews.com/20180316000255 | SW 테스팅 전문기업 와이즈스톤(대표 이영석)이 '2018 K-G... |

Fig. 1. Data Collection using R

수집된 데이터는 비정형화된 데이터이므로, 정형화된 데이터로 가공을 해야 된다. 그림 2와 같이 KAIST에서 제공하는 KAIST 품사 태그셋 SimplePos09을 이용하여 정형화된 데이터로 가공을 시도한다.

그림3에서 제시한 데이터 분석 알고리즘을 이용해 형태소 태그별 인덱싱을 설정한다. 설정 후 str_match함수를 통해 필요한 명사(체언)만 추출하여, 태그별 인덱싱 부분을 제외한 단어만 저장해 정형화된 데이터로 가공한다. 가공 후 데이터를 빈도수와 관련하여 내림차순으로 저장한다.

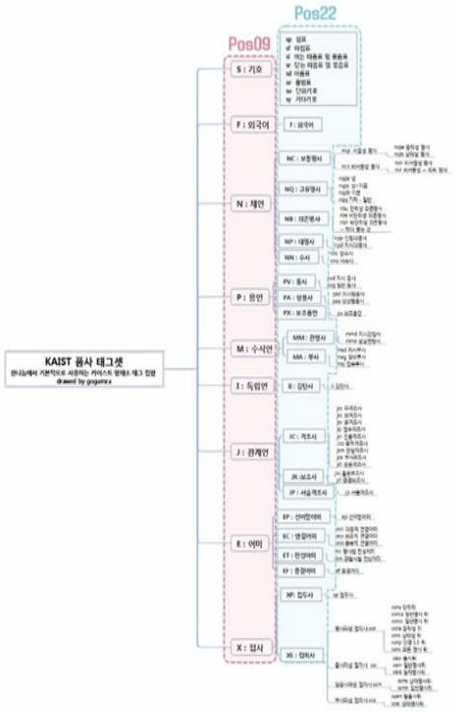



Fig. 2. KAIST Part of speech set form

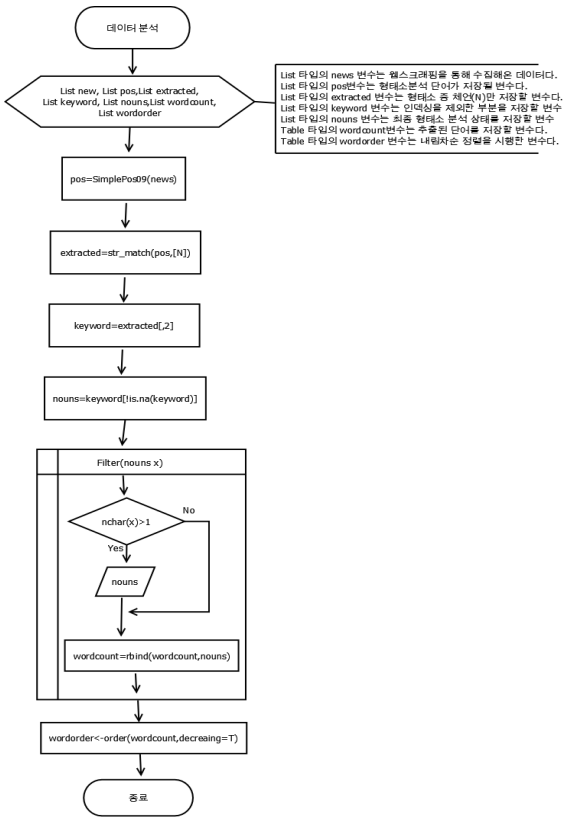



Fig. 3. Data Analysis Algorithm

또한 R에서 제공하는 Apriori 알고리즘[6]을 이용하여 지지도,신뢰도,향상도를 설정을 통해 연관규칙 분석을 시도하여 상관성을 분석한

다. 상관성 분석결과를 참고하여 연관어 MAP을 표시 한다.

데이터의 근원지를 파악하고 분석 내용에 대한 이해도를 높이기 위해 임의로 수집한 데이터 내용과 출처를 표시했지만, 현재 수집된 데이터에서만 국한 될 뿐만 아니라 여러 분야 데이터 수집에서 활용해 사용자가 쉽게 볼 수 있도록 편의성을 증진시키고자 설계 및 구현한다.

## VI. 구현

본 논문에서 제안한 웹사이트를 구현하기 위해 R과 MVC 구조로 연동하여 구현했다. 그림4는 웹사이트에 접속해 단어를 검색하는 모습을 구현한 화면이다.

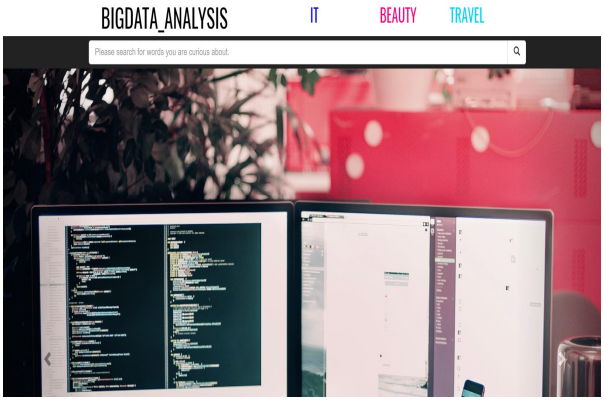



Fig. 4. Implementing a website default search page

그림5는 임의로 수집한 뉴스기사 데이터를 IT, BEAUTY, TRAVEL 분야 별로 분류해 인기 단어 12개를 보여주는 모습을 구현한 것으로, BEAUTY 분야를 보여주고 있다.

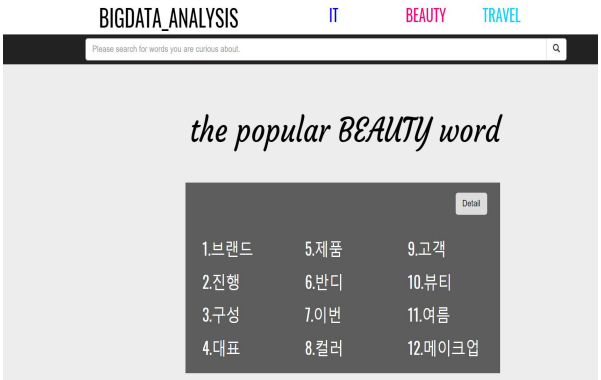



Fig. 5. Implement BEAUTY Partial Popular Topics Analysis Page

그림6과 그림7은 웹사이트 검색을 통해 검색단어를 분석하고 있는 모습을 구현한 화면이다. 그림6은 해당 단어와 연관성 있는 단어의 목록을 5개까지 보여준 화면으로, 연관어 단어의 수를 5개,10개,15개,20개까지로 설정이 가능하다.

연관 단어를 클릭하게 되면 해당 연관 단어로 다시 검색이 가능하도록 구현했다.


Fig. 6. Implement search word ' product ' analysis page 1

그림7은 사용자의 이해도를 높이기 위해 분석단어 데이터 부분을 추출한 것으로, 웹스크래핑을 통해 수집한 뉴스 기사의 일부분을 보여주고 있다.


Fig. 7. Implement search word ' product ' analysis page2

## VII. 결론

본 논문은 R언어 라이브러리 Rserve를 기반으로 MVC 구조로 구현 해 실시간 처리가 가능하게 했다. 이를 활용해 R 언어로 웹 스크래핑을 실행하여 데이터를 수집 후, KAIST에서 제공하는 품사 태그셋을 이용하여 정형화된 데이터로 가공한다. 가공된 데이터를 Apriori 알고리즘 상관성 분석을 통해 연관어 맵을 생성한다. 뿐만 아니라 사용자의 분석 이해도를 높이기 위해, 분석내용과 관련된 데이터를 하단에 보여준다.

설계된 웹사이트는 데이터를 분석해 결과를 보여줄 뿐만 아니라, 여러 분야 데이터 수집에서 활용될 경우 데이터를 하단에 보여줌으로써, 분석 내용에 대한 이해도를 높여 사용자 편리성을 증진시킬 것

이라 기대된다. 향후 연구는 실시간으로 데이터를 수집하여 웹사이트에 적용 할 수있도록 하는 방안을 모색할 것이다.

## REFERENCES

[1] YooHyunGeun, “A Study on the Application of Competition Law to Big Data”
[2] KiHwanNam, “havior-based recommender system using big data analysis”
[3] LeeYongPil, “Implementing Data Analytics Systems Using Statistical Package R and JSP”
[4] BaeJaeDong, “Comparison and Implementation of Interlocking Method between Statistical Applications and R”
[5] JungJinWoo, “Identifying sentence types in korean with morpho-syntactic analysis”
[6] ParkJungHo, “Searching for Information Using Apriori Algorithm Association Rule Mining Techniques”