# 공공이슈 추출을 위한 뉴스 빅데이터 분석 시스템*

김승주, 윤창근, 이차헌, 박동환, 이해준, 박혁주, 이용규
동국대학교 컴퓨터공학과 - 서울
e-mail : pasion2585@naver.com

# News Big Data Analysis System for Public Issue Extraction

Seung Ju Kim, Chang Geun Yoon, Cha Hun Lee, Dong Hwan Park,
Hae Jun Lee, Hyeok Ju Park, Yong Kyu Lee
Department of Computer Science and Engineering, Dongguk University-Seoul

요 약

대중의 관심인 공공이슈를 파악하기 위하여 다양한 종류의 빅데이터를 분석하는 연구가 진행되고 있다. 그러나 기존의 연구에서는 키워드의 노출 횟수만 파악하여 결과로 반영한다. 본 논문은 포털 사이트로부터 얻은 언론사별 뉴스 빅데이터를 이용하여 키워드별 노출 빈도수, 댓글 수 및 추천 수를 반영한 분석 방법을 제안하였다. 공공이슈를 추출하여 얻어낸 키워드들을 워드클라우드, Sankey다이어그램과 같은 형태로 시각화하여 사용자에게 제공한다. 제안된 방법을 사용하면 대중의 반응을 반영한 분석 결과를 확인 할 수 있다.

## 1. 서 론

빅데이터 분석은 전 세계에서 다양한 분야에 활용되고 있으며, 방대한 양의 데이터를 분석하기 위하여 다양한 기술들이 개발되고 있다. 빅데이터 분석은 데이터를 수집 및 분석하고 정보를 추출하여 새로운 사실을 알아내거나 향후 일어날 일들을 예측할 수 있다[1]. 그 중에서 뉴스 기사 및 소셜 네트워크 서비스(SNS)의 분석은 공공이슈를 확인하거나 대중의 관심을 확인하기 위하여 활용되고 있다[2].

기존 시스템은 데이터를 분석하는데 있어서 크롤링하여 수집한 단어의 노출 횟수를 카운팅하는 등 단순한 방법으로 분석하고 시각화하는데 그친다. 이를테면 SNS에 나타난 해당 페이지에 대한 긍정/부정적인 정보들을 토대로 이를 시각화하여 나타내줄 뿐이다[2]. 하지만 본 시스템은 단순히 웹을 통하여 얻을 수 있는 단편적인 정보들을 분석하여 시각화하는데 그치지 않고 여러 언론사로부터 기사를 크롤링하여 정보를 수집하고 분석하는 과정에서 여러 가지 요소들을 결합하여 공공이슈의 대한 중요도를 계산하고, 그 결과를 시각화 하고자 한다.

본 논문에서 제안하는 시스템은 단순한 형태의 분석 시스템과 달리 해당 뉴스 기사의 대중의 관심을 통합적으로 고려하여 키워드에 대한 신뢰성을 높이는데 차이점이 있다. 이를 위하여 해당 키워드가 포함 된 뉴스 기사의 댓글 수, 추천 수를 통해 뉴스 기사를 읽은 사람들의 반응 정도를 반영하고, 뉴스 기사가 등록된 시간에 따라 가중치 값을 계산하여 공공이슈에 대한 중요도를 파악하고자 하였다.

본 논문의 구성은 2장에서는 기존에 사용되는 분석 플랫폼과 형태소 분석에 대해 설명한다. 3장에서는 공공이슈 추출 시스템에 대한 구성도 및 사용된 기술에 대한 구체적인 내용을 설명한다. 4장에서는 시스템의 전체적인 흐름도 및 결과 화면을 제시한다. 마지막으로 5장에서는 앞서 언급한 내용을 정리하며 결론을 맺는다.

## 2. 관련 연구

### 2.1 분석 플랫폼

현재 사용되고 있는 분석 플랫폼들은 주로 SNS에 게시된 글이나 뉴스 기사에서 키워드를 추출한다. 키워드별 노출 빈도수를 얻어낸 다음 이를 이용해 시각화하고, 사용자에게 의사 결정에 도움이 될 만한 정보를 제공한다. 또는 개인이나 기업의 운영 지표로 사용한다[3][4]. 본 논문에서는 뉴스 기사에서 노출 빈도를 통해 얻어낸 키워드에 추가로 기사의 추천 수 등 사용자들의 의견이 반영된 요소들을 포함하여 데이터를 분석하는 방법에 대해 제안한다.

### 2.2 형태소 분석

자연어 처리에서 형태소 분석이란 어떤 문장을 최소한의 의미를 갖는 단위인 형태소 단위로 나누어 구별하는 것이다. 한국어 형태소 분석기는 일반적으로 세종 품사 태그[5]를 기준으로 품사를 구분하였으며, 명사의 경우 시스

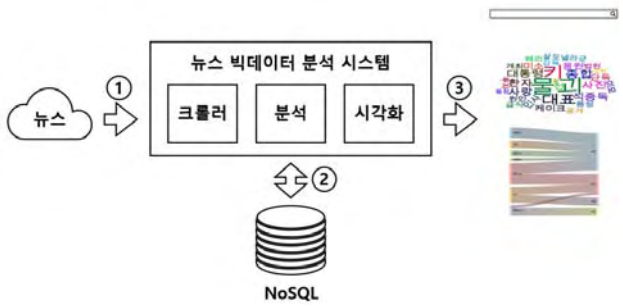
* 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음(2016-0-00017)

템사전을 바탕으로 확인하여 합성명사일 경우에는 각각의 명사로 나누어 분류한다.

본 논문에서 사용하는 형태소 분석기는 Seunjeon[6]으로 이는 mecab-ko-lib라는 MeCab용 한국어 형태소 사전을 이용한 것이다.

## 3. 뉴스 빅데이터 분석 시스템 구성

본 논문에서 제안하는 시스템의 전체 구성도는 (그림 1)과 같다. ①에서 크롤러를 이용하여 포털 사이트에 있는 뉴스 빅데이터를 수집하여 NoSQL에 저장한다. ②에서는 크롤링 한 데이터를 NoSQL에서 조회하여 분석하고, 분석된 결과를 NoSQL에 저장한다. ③에서는 클라이언트가 서버에 공공이슈 키워드를 요청하면 서버는 분석 결과를 전송하고, 클라이언트는 전송받은 정보를 시각화하여 출력한다.




(그림 1) 뉴스 빅데이터 분석 시스템 전체 구성도

### 3.1 뉴스 빅데이터 크롤링

본 논문에서 제안하는 시스템에서는 웹의 뉴스 기사 빅데이터를 수집하는 기술인 크롤링(Crawling)을 사용한다. 이를 위해 웹 페이지의 데이터를 가져오기 위한 Jsoup과 JSON 데이터를 추출하기 위한 JSONparser 라이브러리를 사용한다. 이러한 크롤러(Crawler)는 언론사마다 웹페이지의 구성이 다르므로 데이터를 수집하기 위해서는 각각의 언론사마다 서로 다른 크롤러가 필요하다. 따라서 다른 언론사들의 뉴스를 종합하여 보여주는 포털의 크롤러를 만들어 여러 언론사들의 뉴스 기사들의 데이터를 수집하도록 하였다.

크롤러는 30분의 주기로 실행이 되며, 사용자들의 반응 정보를 수집하기 위해 기사가 쓰인 시간으로부터 한 시간의 간격으로 기사를 수집한다. 검색 시간 범위는 주기 사이에 일어날 수 있는 데이터 누락을 막기 위하여 주기보다 더 큰 시간 범위인 3시간 이내로 설정한다. 수집한 데이터는 제목, 기사가 작성된 시간, 기사 내용, 언론사, URL, 카테고리, 댓글 수, 추천 수로 분류하여 NoSQL에 저장한다. 이 때, 동일한 정보를 가진 데이터가 저장되는 것을 방지하기 위하여 제목과 URL로 중복 여부를 검사한다.

(그림 2)는 크롤러의 동작 과정을 설명하기 위한 웹 페이지의 HTML(HyperText Markup Language)을 나타낸 것이다. 여기서 필요한 데이터가 ③의 “support”라는 문자일 경우에 추출과정은 다음과 같다. 단어 “support”는 순서대로 ①의 <body>, ②의 , ③의 <p> 태그로 둘러싸여 있으므로 필요로 하는 데이터의 범위를 좁혀나가면서 태그를 선택한다. ②의 경우 가 여러 개 존재하므로 class의 값이 “content”를 태그를 선택한다. 마지막으로 ③의 <p>안에 있는 “support” 데이터 만 추출하여 수집한다.

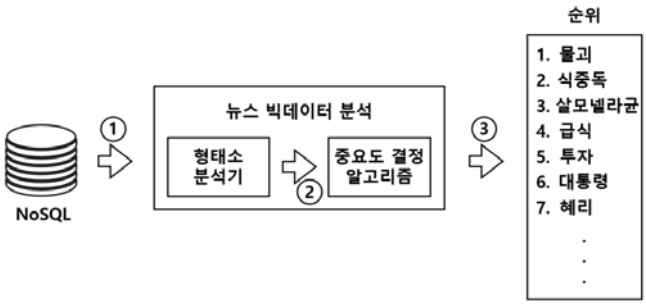
```
<!DOCTYPE html>
<html>
  <head>
    <title> news </title>
  </head>
①<body>
  ②<div class="content">
     ③<p> support </p>
    </div>
    <div class="sub">
       Event
    </div>
  </body>
</html>
```

(그림 2) 웹 페이지의 HTML 예시

언론사의 스포츠면 뉴스 기사들은 JavaScript를 사용하므로 Jsoup을 사용하여 데이터를 추출하는 것이 불가능하다. 따라서 HTML이 아닌 JSON의 데이터를 JSONparser 라이브러리를 사용하여 뉴스 기사들을 수집하였다. JSONparser 또한 Jsoup과 같이 유사한 방식으로 데이터를 수집한다.

### 3.2 공공이슈 추출

본 논문에서 제안하는 시스템에서는 키워드를 명사로 추출하기 위해 형태소 분석기를 사용한다. 그리고 수식을 이용하여 점수를 계산하고, 이 점수를 사용하여 단어들의 순위를 정한다.




(그림 3) 뉴스 빅데이터 분석 방법

(그림 3)은 노출 가중치를 적용하여 뉴스 빅데이터를 분석하는 방법이다. ①은 NoSQL에 저장되어 있는 뉴스 빅데이터를 조회한다. 그리고, Seunjeon 형태소 분석 라이브

러리를 이용하여 1~3시간 사이에 기사들의 제목에서 명사를 키워드로 추출하고 키워드별로 출현 빈도를 체크한다. ②에서 식 (1)을 이용하여 기사별로 구독자 반응 가중치를 산출한다. 그리고 식 (2)를 이용하여 키워드가 등장하는 기사의 구독자 반응 가중치를 합하여, 키워드 중요도를 계산한다. ③은 ②에서 계산된 키워드 중요도 결과를 내림차순 정렬하여 리스트를 작성을 한다. 이 리스트에 있는 1~30위까지의 키워드들은 Wordcloud에 사용되고, 시각화되는 글자 크기는 키워드 중요도에 의한 순위를 기준으로 결정한다.

식 (1)은 기사별 구독자 반응 가중치를 계산하는데 사용된다. C는 기사에 달린 댓글의 수이고, L은 기사를 읽고 사용자들이 클릭한 "좋아요"의 수이고, R는 기사의 추천 수이다. T는 기사가 등록되고 경과된 시간으로 루트를 취하여 나누었다. 시간에 따라 값의 차이가 많이 나지 않게 하기 위해 루트를 사용하였다. 기사별로 구독자 반응 가중치 값을 계산하고, 기사에 포함된 키워드는 동일한 구독자 반응 가중치 값(W)을 갖는다.

$$W = \frac{C + L + R}{\sqrt{T}} \quad \text{식 (1)}$$

W = 구독자 반응 가중치
C = 댓글 수
L = 좋아요 수
R = 추천 수
T = (현재 시간 - 기사가 등록된 시간)(분)

식 (2)는 키워드 중요도 계산하는데 사용된다. W는 식 (1)에서 계산한 각 기사의 구독자 반응 가중치 값이다. m은 키워드가 등장하는 뉴스의 수이다. 키워드가 등장하는 기사의 구독자 반응 가중치 값을 더하여 키워드 중요도(I)를 계산한다.

$$I = \sum_{n=1}^{m} W_n \quad \text{식 (2)}$$

I = 키워드 중요도
W = 기사별 구독자 반응 가중치
m = 키워드가 등장하는 뉴스의 수

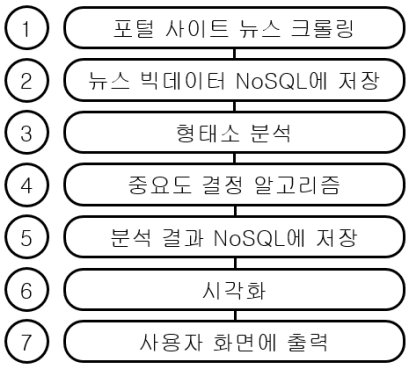
공공이슈가 될 수 있는 키워드들의 신뢰성을 높이기 위해 뉴스 기사를 이용한 사람들의 반응인 댓글, 좋아요, 추천 수를 반영하여 가중치를 계산한다. 시간은 단시간 안에 이용자들의 반응이 많을수록 더 주목이 된 것이기 때문에 가중치가 높게 나오게 하였다.

## 4. 구현 결과

### 4.1 시스템 흐름도

본 논문에서 제안한 시스템 동작의 전체 흐름도는 (그림 4)와 같다.

①은 Jsoup, JSONparser 라이브러리를 사용하여 포털 사이트의 뉴스들을 크롤링한다. ②는 크롤링한 데이터들을 NoSQL에 저장한다. ③은 NoSQL에서 뉴스 빅데이터를 가져와 Seunjeon 라이브러리를 이용하여 형태소 분석을 하고 명사를 뽑아낸 후 명사의 출현 빈도를 체크한다. ④에서 식 (1)을 이용하여 기사별로 구독자 반응 가중치를 계산한다. 그리고 키워드 중요도를 산출하여 내림차순으로 정렬한다. ⑤는 분석이 끝나면 생성된 데이터들을 NoSQL에 저장한다. ⑥는 사용자의 요청이 들어오면 NoSQL에 저장되어 있는 정보를 조회하여 wordcloud나 Sankey 다이어그램으로 시각화 한다. ⑦은 사용자의 웹 브라우저에 출력한다.

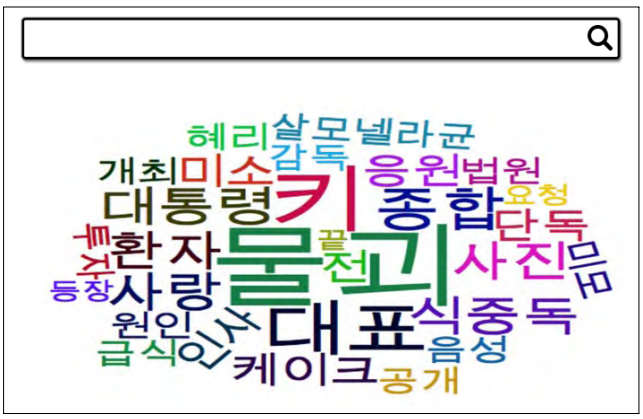



(그림 4) 시스템 전체 흐름도

### 4.2 결과 화면

4.2.1 공공이슈 Wordcloud

NoSQL에서 분석되어있는 1~3시간 이내의 뉴스 기사들 중에서 키워드 중요도가 가장 높은 상위 30개의 키워드를 선별한 후 Wordcloud 형태로 시각화하여 보여준다. (그림 5)는 공공이슈 Wordcloud의 출력 결과이다. 식 (2)에서 계산한 중요도 점수가 높을수록 키워드의 글자 크기가 더 크다. 그리고 글자를 누르면 그 키워드의 노출 횟수 등 세부 내용을 확인할 수 있고, (그림 6)과 같이 키워드가 포함된 뉴스기사들의 목록을 볼 수 있다. 또한 제목을 누르면 기사 본문으로 이동할 수 있다. 또한 검색창에 키




(그림 5) 공공이슈 Wordcloud 출력 결과

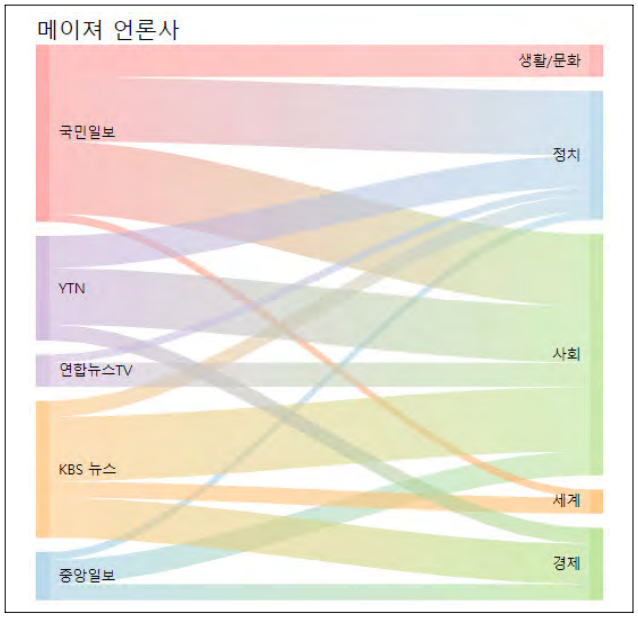
워드로 검색을 하면, 그 키워드를 포함한 뉴스 기사의 내용을 분석하여 공공이슈 Wordcloud를 출력한다.

| 날짜 | 제목 | 언론사 |
|---|---|---|
| 2018/10/01 15:35 | 유니온페이 '가을에 해외여행' 이벤트 진행 | 스포츠동아 |
| 2018/10/01 15:58 | 올 가을 풍요 알리는 황금 들녘 | 국민일보 |
| 2018/10/01 15:52 | 광주시, 전국 최초 가을 수학여행 안전대진단 | 노컷뉴스 |
| 2018/10/01 15:45 | 코스모스 구경하며 가을 만끽하는 어린이들 | 뉴스1 |
| 2018/10/01 15:43 | "이벤트 즐기며 제주로 가을여행 떠나볼까?" | 뉴스1 |
| 2018/10/01 15:39 | 물들어두고 싶은 가을 | 뉴스1 |
| 2018/10/01 15:38 | '가을이 가기 전에' | 뉴스1 |
| 2018/10/01 15:38 | '가을아 오래 머물러라' | 뉴스1 |
| 2018/10/01 15:37 | 가을 정취 느끼는 시민들 | 뉴스1 |
| 2018/10/01 15:24 | 가을 햇살 쬐는 황금물떼새 | 뉴스1 |
| 2018/10/01 15:24 | 황금물떼새의 가을 | 뉴스1 |
| 2018/10/01 15:13 | 가을에 돌아온 아이콘 | 연합뉴스 |
| 2018/10/01 15:13 | 경기도의 가을 속으로 걸어 들어가 보자-경기관광공사 10월의 명소 | 조선일보 |
| 2018/10/01 15:13 | 가을 감성 저격할 아이콘 | 연합뉴스 |
| 2018/10/01 15:11 | 가을 손짓하는 갈대 | 뉴스1 |
| 2018/10/01 15:11 | 가을 노래하는 갈대 | 뉴스1 |
| 2018/10/01 15:11 | [포토뉴스] 청년군 "가을여행 코스로 최적" | 프레시안 |
| 2018/10/01 15:09 | "가을 추자도 이렇게 즐겨요"…주민이 뽑은 관광 10선 | 뉴스1 |
| 2018/10/01 16:05 | 가을 눈총 풍경 | 연합뉴스 |
| 2018/10/01 15:10 | [날씨] 찬 바람 불며 서늘…내일 쾌청한 가을 | YTN |
| 2018/10/01 15:49 | 개그맨 유세석 · 플레빗크 제니, '아름다운 가을마을, 미추리' 홍연화정 | 스포츠경향 |
| 2018/10/01 15:48 | [HI★현장] '뷰티인사이드' 막오른 서현진X이민기 표 '가을 감성 멜로' | 한국일보 |
| 2018/10/01 15:46 | "인천공항 품은 멜로"…이제훈X배수빈 '여우각시별', 가을감성 겨울까[종합] | SBS |
| 2018/10/01 15:43 | 박지빈, 가을 남자 변신…"연기 점점 두려워" | 매일경제 |
| 2018/10/01 15:24 | "선다별" 가을 겨울 편, 오늘(1일) 첫 방송…관전포인트는? | MBN |
| 2018/10/01 15:23 | "인천공항의 모든 것"…'여우각시별' 이제훈♥배수빈, 시린 가을 녹일까 [종합] | 마이데일리 |
| 2018/10/01 15:22 | '공듀이루' 개천절 연휴 4강전..가을 감성 멜화 | 스타뉴스 |
| 2018/10/01 15:21 | "가을느낌 물씬"…박지빈, 한층 깊어진 남성美 [화보] | 엑스포츠뉴스 |
| 2018/10/01 15:20 | 박지빈, 한층 깊어진 남성미… 내가 바로 '가을 남자' | enews24 |
| 2018/10/01 15:20 | [T포토-Lab] 이민기-안재현 '격이 다른 모델핏, 남자의 가을패션' | TV리포트 |

(그림 6) 키워드가 들어간 뉴스 목록

4.2.2 Sankey 다이어그램

(그림 7)은 메이저언론과 마이너언론으로 나누어서 기사들이 어떤 분야의 기사가 수집되었는지 기사분야 별 비율을 한눈에 보기 쉽게 Sankey 다이어그램으로 표현한 것 중 메이져 언론사를 나타낸 것이다. 포털에서 지정된 카테고리로 항목을 나누어 크롤링으로 수집된 뉴스기사의 수를 분석해 기사가 많이 나오는 언론사를 메이저언론, 기사의 수가 적은 언론사를 마이너언론로 등록하여 화면에 나타낸다.




(그림 7) Sankey 다이어그램 출력 결과

## 5. 결 론

본 논문에서 제안된 시스템은 댓글 수, 추천 수, 작성 시간 등을 반영하여 뉴스 빅데이터를 분석하였다. 그 결과 기사에 대한 대중의 관심도를 적용한 분석 결과를 얻을 수 있었다. 이를 통하여 기존의 단순히 키워드의 노출 횟수만을 적용한 시스템에서는 고려되지 못한 기사의 반응 정도를 반영 할 수 있었다.

대중의 관심도가 적용되어 기사에 대한 대중의 관심도가 크면 키워드의 출현 횟수가 적어도 결과화면에 노출되게 하였고, 기사가 중복되어 게재되어도 이용자의 관심도가 떨어지는 이슈들을 적절한 비율로 배제시킬 수 있게 되었다. 이로 인해 뉴스분석 시스템의 목적인 공공이슈 분석에 대한 신뢰도를 더 향상시켰다고 볼 수 있다.

반면, 정보를 시각화해서 보여주는 다이어그램 종류가 한정적인데, 뉴스데이터와 호환되는 형태의 다이어그램을 찾고, 응용 및 변형하여 더 많은 종류의 다이어그램, 그래프를 사용자에게 제공할 수 있게 할 예정이다. 또한, UI부분을 사용자에게 친화적일 수 있도록, 형태를 발전시켜 나갈 것 이다.

## 참고문헌

[1] 김상락, 강만모, “빅데이터 분석 기술의 오늘과 미래”, 정보과학회지, Vol.32, No.1. pp.8-17, 2014.

[2] 임광혁, “SNS 빅데이터 분석 기술 동향 및 발전방향”, 한국콘텐츠학회지, Vol.15, No.2, pp.38-43, 2017.

[3] 김재생, “빅데이터 분석 기술과 활용사례”, 한국콘텐츠학회지, Vol.12, No.1, pp.14-20, 2014.

[4] 조영임, “빅데이터의 이해와 주요 이슈들”, 한국지역정보화학회지, Vol.16, No.3, pp.43-65, 2013.

[5] 국립국어원 언어정보나눔터, https://ithub.korean.go.kr/user/main.do

[6] 은전한닢, https://bitbucket.org/eunjeon/seunjeon