# 사회적 핵심 이슈 키워드 정보를 활용한 채팅 프로그램 개발

윤경섭*, 정원혁O
OO인하공업전문대학 컴퓨터정보공학과,
*인하공업전문대학 컴퓨터정보공학과
e-mail: ksyoon@inhatc.ac.kr*, cjswo7119@gmail.comO

# Development of chatting program using social issue keyword information

Kyung-Suob Yoon*, Won-Hyeok JeongO
*Dept. of Computer Science, Inha Technical College,
ODept. of Computer Science, Inha Technical College

● 요 약 ●

본 논문에서 이슈 키워드 추출을 위해 텍스트 마이닝(Text Mining) 기술을 요구한다. 사회적 이슈 키워드를 추출하기 위해 키워드 수집 모델이 되는 사이트에서 크롤링(crawling)을 수행한 뒤, 형태소 단위 의미있는 단어를 수집하기 위해 형태소 분석(morphological analysis)을 수행한다. 한국어 형태소 분석을 위해 파이썬의 코엔엘파이(KoNLPy) 패키지를 활용한다. 형태소 분석을 통해 나뉘어진 단어에서 통계를 내어 이슈 키워드 추출한다. 이슈 키워드를 뒷받침할 연관 단어를 분석하기 위해 단어 임베딩(Word Embedding)을 수행한다. 단어 임베딩 수행을 위해 Word2Vec 모델 중 Skip-Gram 방법론을 적용하여 연관 단어를 분석하도록 개발하였다. 웹 소켓(Web Socket) 통신을 통한 채팅 프로그램의 상단에 분석한 이슈 키워드와 연관 단어를 출력하도록 개발하였다.

**키워드:** 텍스트 마이닝(Text Mining), 크롤링(crawling), 형태소 분석(morphological analysis), 코엔엘파이(KoNLPy), 단어 임베딩(Word Embedding), Word2Vec, Skip-Gram

## I. Introduction

현대 사회에서 이슈화되는 사건이 늘어나면서 국민청원을 통해 소수의 사람들도 인권을 보장 받을 수 있는 긍정적인 국민의식을 기를 수 있었다. 그러한 과정 속에 보탬이 되고자 우리는 함께 생각을 공유하고, 동참의지를 기를 수 있도록 실시간 소통의 장을 형성한 후, 사회적 운동에 동참할 수 있도록 도모함이 필요했다.

사회적 이슈가 되는 기사들의 핵심 키워드를 수집하고, 연관단어를 탐색하여 채팅 주제 선정을 마친다. 주제와 관련하여 사용자들끼리 소통을 진행할 수 있게 소통의 장을 형성하고, 사회적 관심거리에 대한 폭넓은 시야와 견해, 의견으로 목적의식을 기를 수 있도록 방향을 제시한다. 생성되는 채팅방은 '방장'의 개념을 제거하여 한 사람의 이념에 따르지 않고 모두의 의견을 공유 가능하도록 설계하는 것이 목표이다. 작은 힘이 세상을 움직인다. 소수의 힘만으로 해결할 수 없는 큰 사건이 도태되지 않도록 계속 언급하고 생각을 공유한다. 모두의 동참의지를 일어서게 만들고 전파될 수 있게 하여, 긍정적인 국민의식을 기르는 사회 운동이 될 수 있다.

## II. Preliminaries

소통의 매개체로 활용될 정보 및 키워드를 추출하기 위해서 수집이 필요한 분석모델이 필요하고, 포함된 데이터를 가공하여 의미있는 정보를 추출하는 과정인 텍스트 마이닝 기술이 필요하다.

분석모델에서 추출하는 데이터는 뉴스기사 중 '문장' 형태로 데이터를 제공한다. 텍스트 마이닝 기술로써 소통 매개체로 활용될 키워드는 첫째, '의미적 분석이 가능'하며 둘째, '분해가 되지 않는 가장 작은 말의 단위의 단어'여야 한다. '의미적 분석이 가능'한 경우 단어를 통해 사용자는 해당 단어를 인지하고 의사소통이 가능하도록 수행되어야 한다. '가장 작은 말의 단위의 단어'의 경우, 둘 이상의 단어가 출력될 경우에 단어 임베딩을 통한 연관 단어 탐색 시, 두 단어로 형성된 구절의 경우 연관 단어의 탐색이 불가하다. 두 가지 조건을 만족하는 한국어 단위 구조는 형태소가 적합하여 시스템 구현 시, 분석모델에서 수집한 문장 형태의 뉴스기사에서 형태소 분석 기능이 필요했고, 분석된 형태소를 통해 집계를 내어 가장 많이 언급된 키워드를 색출하는 작업을 통해 이슈 키워드를 탐색 하는 과정이 필요하다.

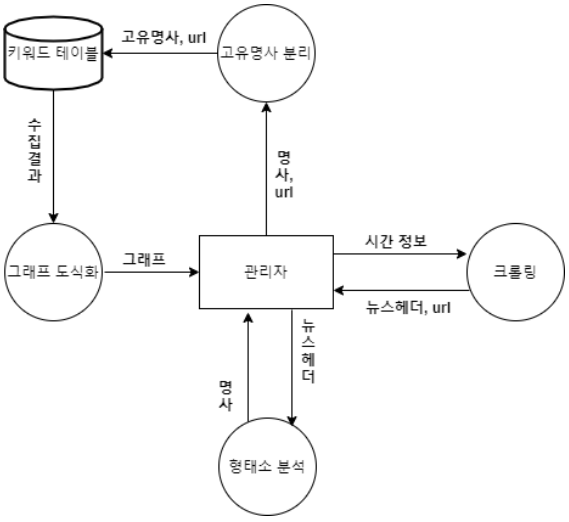


Fig. 1. Text Mining Process

사용자에게 이슈 키워드만 제공할 경우, 사회적 관심거리 중 해당 이슈 키워드가 차지하는 비중이 비교적 눈에 띄지 않다면 또는 이슈 키워드만으로는 인지가 불가능할 경우를 대비해야 한다. 이슈 키워드와 관련된 연관 단어를 함께 출력하여 사용자에게 제공 시, 사회적 관심거리와 연상된 주제를 상기시킬 계기를 제공할 수 있게 된다. 분석 모델 중 뉴스기사에서 보이는 특징 중 하나는 사용자들의 흥미를 유발할 수 있는 기사 제목으로 작성되어 있다. 기사 내용 중 주목받는 키워드를 포함시켜 키워드를 뒷받침할 연관단어를 주변에 포진시켜 놓은 것이 특징이다. 단어 임베딩 중 단어상 연관있는 단어는 주변에 밀집한다는 가정하에 각 연관단어를 예측하는 Skip-Gram 방법이 적합하다 판단하였다.

## III. Design

### 3.1 핵심 키워드 수집을 위한 크롤링, 형태소 분석, 단어 임베딩 수행

핵심 이슈 키워드 선정을 위해서는 단어 크롤링이 선행되어야 한다. 이를 위해 국내에서 이용자 수가 가장 많은 '네이버 뉴스'를 수집 모델로 선정했다. 네이버 뉴스 상단의 '가장 많이 본 뉴스' 메뉴의 기사들 중 사회 섹션의 기사 제목들을 크롤링하기로 결정했다.

크롤링한 기사 제목 데이터는 문장 형태로 되어 있으므로 핵심 키워드를 추출하기 위해서는 고유명사 단위로 리스트를 재구성하는 작업이 선행되어야 한다. 따라서 대표적인 Python 형태소 분석 라이브러리인 KoNLPy(코엔엘파이)를 활용하여 형태소 분석을 수행했다. 추출한 단어들 중 뜻이 없는 한 글자 데이터가 다수 존재하는 것을 확인할 수 있었고, 두 글자 이상의 단어들만 리스트에 추가하는 방식으로 문제를 해결했다. 그 후, 각 기사당 키워드 리스트를 갖는 2차원 리스트 형태로 재구성하는 작업을 수행했고, '단어', '수집 날짜', '추출 URL', '기사 제목' 순서대로 DB에 저장했다.

| 123 relt_no | ABC relt_kwrd | ABC cnt_kwrd | 123 daily | 123 kw_rank | kw_date |
|---|---|---|---|---|---|
| 29 | 왕기준 | 유도 | 1 | 0 | 2020-05-02 00:00:00.0 |
| 30 | 성폭행 | 유도 | 1 | 1 | 2020-05-02 00:00:00.0 |
| 31 | 은메달 | 유도 | 1 | 2 | 2020-05-02 00:00:00.0 |
| 32 | 미성년 | 유도 | 1 | 3 | 2020-05-02 00:00:00.0 |
| 33 | 혐의 | 유도 | 1 | 4 | 2020-05-02 00:00:00.0 |
| 34 | 올림픽 | 유도 | 1 | 5 | 2020-05-02 00:00:00.0 |
| 35 | 구속 | 유도 | 1 | 6 | 2020-05-02 00:00:00.0 |
| 36 | 리스트 | 유도 | 1 | 7 | 2020-05-02 00:00:00.0 |

Fig. 2. DB 저장 내용

재구성된 2차원 리스트에서 연관 단어를 추출하기 위해서는 단어 임베딩 과정을 거쳐야 한다. 여러 임베딩 모델 중 중심 단어를 통해 주변 단어를 예측하는 모델인 Skip-Gram 모델을 사용하여 임베딩을 수행하기로 선정했다. Skip-Gram 모델을 사용하는 경우, 단어 단위의 Semantic Accuracy가 다른 모델들에 비해 높은 것도 확인할 수 있다.

| Model | Semantic-Syntactic Word Relationship test set | | MSR Word Relatedness |
|---|---|---|---|
| Architecture | Semantic Accuracy [%] | Syntactic Accuracy [%] | Test Set [20] |
| RNNLM | 9 | 36 | 35 |
| NNLM | 23 | 53 | 47 |
| CBOW | 24 | 64 | 61 |
| Skip-gram | 55 | 59 | 56 |

Fig. 3. 각 임베딩 모델별 정확도 비교

수집한 키워드의 등장 횟수를 카운트하여 가장 많이 등장한 키워드를 기준으로 해당 키워드가 포함된 기사 제목에서 연관 단어를 탐색하고, 관심 단어로 예측되는 각 단어마다 가중치를 부여하여 연관도를 분석한다. 단어 임베딩 수행 라이브러리는 구글 Word2Vec을 사용했다. 연관도가 높은 순으로 리스트를 정렬하고, 추출된 키워드 리스트는 관심 단어와 연관 단어를 웹 상에서 출력하기 위해 DB에 저장했다.

### 3.2 API 요청을 통한 Front/Back-end 시스템 구축

사용자들에게 출력되는 Front-end 화면은 React를 사용하여 구현했고, DB와의 연동 및 웹 Front-end에서 사용자가 입력한 데이터를 받아 API 요청을 수행하기 위해 Spring-Boot(스프링 부트)를 사용했다. 스프링 부트의 컨트롤러에서 정의한 URL을 이용하여 API 요청을 보내면 해당 부분의 함수를 수행한다.

### 3.3 채팅 서버 구현을 위한 미들웨어 구축

사용자의 실시간 채팅 환경과 빠른 속도의 데이터 처리를 보장하기 위해 GraphQL 기반의 미들웨어 서버를 구축했다.

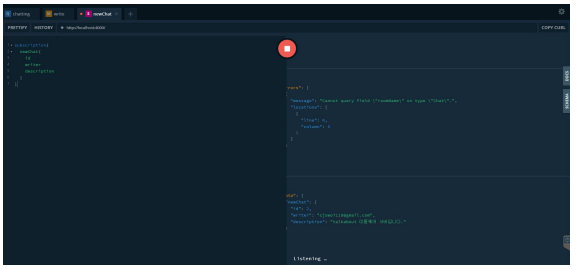

Fig. 4. GraphQL 대시보드

graphQL에서는 SQL 대신 GQL 명령어를 사용하여 데이터를 가져온다. SQL은 Back-end 시스템에서 작성 및 호출되지만, GQL은 클라이언트 시스템에서 작성 및 호출되므로 웹소켓 통신 구현에 용이하다는 장점이 있다.

실시간 채팅 로그를 기록하기 위해 기존 Front-end에서 동작하는 React 환경에 Apolo 라이브러리를 임포트한 뒤, 데이터가 변경되면 해당 영역을 자동으로 리렌더링하도록 구현했다.

### 3.4 사용자 정보 수집 및 유효성 검사, 정보 수정을 위한 모듈 구현

회원가입에서 DB에 저장하기 전에 사용자의 입력 오류를 검출하기 위해 유효성 검사를 수행한다. 입력받는 데이터 항목은 사전에 정의한 요구사항명세서를 기반으로 각 조건에 맞게 유효성 검사를 수행하도록 반영했다. 입력한 E-mail에 대해 인증 URL을 생성하여 해당 URL을 클릭하지 않으면 회원가입이 완료되지 않도록 2차적으로 유효성 검사를 진행한다. 악의적인 접근을 방지하기 위해 SHA256 암호화를 적용하여 URL을 생성한다. 입력한 사용자 정보는 추후 수정이 가능하도록 메뉴를 추가로 구성하며, 해당 기능에 대해서도 유효성 검사를 실시한다.

## IV. Implementation

사이트 첫 진입시 사용자에게 E-mail과 비밀번호를 입력받는 로그인 화면이 뜬다. 사이트를 이용하기 위해서는 회원가입이 필요하다. 회원가입은 사용자의 E-mail 및 비밀번호를 포함한 개인정보를 입력받은 후, 각 입력에 대한 유효성 검사를 진행한다.

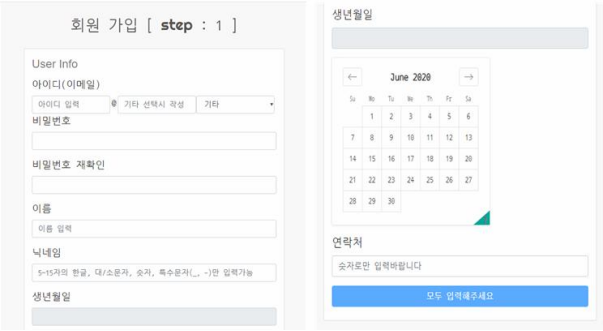


Fig. 5. 회원가입 입력 화면

페이지 하단 버튼을 클릭하면 사용자가 입력한 메일 주소로 인증 메일을 발송하고, 메일 내에 있는 인증 URL을 클릭하면 인증이 완료되고 해당 계정의 사이트 로그인 권한이 부여된다.

Fig. 6. 회원가입 인증 메일 내용

로그인을 수행하면 기입한 입력에 대해 유효성 검사를 진행한다. 문제가 발생하지 않는다면 서비스 화면으로 이동한다.

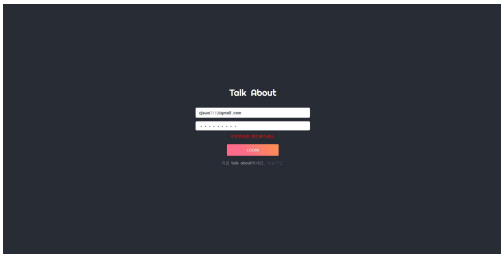


Fig. 7. 로그인 유효성 검사 수행 화면

화면 좌측에는 순서대로 채팅, 사용자 정보, 통계 메뉴로 이루어져 있다. 기본 화면인 채팅 메뉴는 상단에 오늘 날짜와 그에 해당하는 관심 키워드 및 연관 키워드로 구성되며, 하단에는 채팅에 참여할 수 있는 화면과 입력창으로 구성되어 있다.

Fig. 8. 홈 및 채팅 화면

사용자 정보 메뉴는 회원가입 시 입력한 E-mail, 이름, 닉네임 등을 조회하고 비밀번호와 생년월일, 연락처 수정이 가능하다.

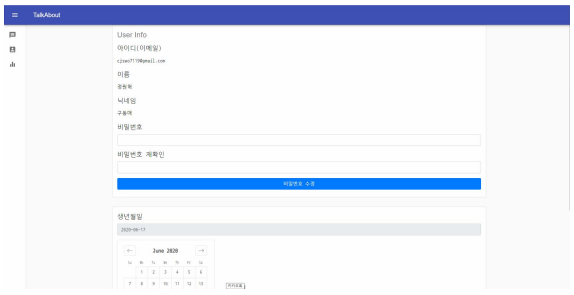


Fig. 9. 사용자 정보 조회 및 수정 화면

마지막으로 통계 메뉴는 오늘 날짜를 기준으로 가장 많이 언급된 상위 10개 키워드를 파이 그래프로 볼 수 있다. 각 단어의 분포도를 통해 현재 이슈되고 있는 키워드를 확인할 수 있으며, 이를 채팅에 활용할 수 있다.

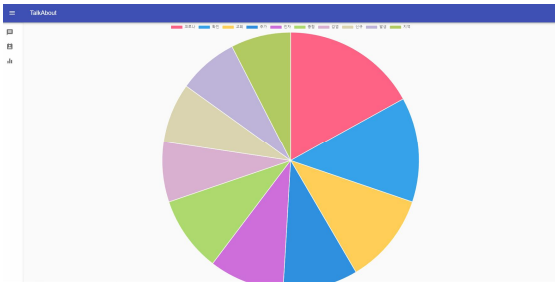


Fig. 10. 이슈 키워드 분포 파이 그래프

## V. Conclusions

현대 사회에서의 SNS는 사용자간의 관계를 형성하고 개인의 생각을 자유롭게 공유할 수 있는 성격을 갖는다. 본 논문은 전통적인 SNS 중의 하나인 채팅 시스템 구현을 통해 사회 이슈에 대해 사용자들의 의견을 공유하고 동참 의지를 전파할 수 있는 실시간 플랫폼을 구축한다. 작은 힘이 세상을 움직인다는 말처럼 소수의 의견이 모이면 큰 파장을 일으킬 수 있다. 국민 의식을 기르고 더 나아가 건강한 사회 운동을 주도할 수 있도록 소통의 장을 형성하는 것이 위 연구의 목표이다.

향후 사용자와 관리자 기능을 분리하여 채팅방 내의 불량 사용자들을 제제하고 관리할 수 있는 기능 추가와 함께 키워드별로 채팅방을 따로 두어 자정마다 채팅방이 추가 및 삭제가 이루어지도록 하기 위한 연구가 필요할 것으로 생각된다.

## REFERENCES

[1] https://arxiv.org/pdf/1301.3781.pdf
[2] https://kiise.or.kr/e_journal/2018/5/JOK/pdf/04.pdf
[3] https://pdfs.semanticscholar.org/9867/5d11d8004cc48b569a42ed85142376ba265e.pdf
[4] https://www.dbpia.co.kr/pdf/pdfView.do?nodeId=NODE07531033&mark=0&bookmarkCnt=1&ipRange=N&language=ko_KR