# 웹 미디어 데이터를 이용한 이슈 예측 시스템 설계

윤현노*, 문남미*
*호서대학교 컴퓨터공학부
e-mail:zxcv9153@naver.com

# Designing issue prediction system using web media data

Hyun-Noh Yun*, Nammeee Moon*
*Dept of Computer Science, Hoseo University

요 약

IT 기술의 발달에 따라 다양한 웹 미디어의 데이터가 기하급수적으로 증가하고 있으며 이는 비정형 형태의 빅 데이터로 활용도가 매우 높다. 그 중 인터넷 뉴스나 SNS 등은 시간의 흐름에 따라 다양한 이슈들이 서로 영향을 주며 발생, 결합, 분화, 소멸된다. 본 논문에서는 인터넷상에서 발생하는 비정형 데이터들을 수집하여 텍스트 마이닝을 통해 글의 주요이슈 키워드, 카테고리, 날짜 등을 추출한다. 추출한 데이터를 일정 기간별로 나누어 이슈 매핑을 통해 이슈간의 상관관계를 분석한다. 나아가 LSTM 또는 GRU를 이용한 딥러닝을 통해 앞으로의 이슈를 예측하는 시스템 설계를 제안한다.

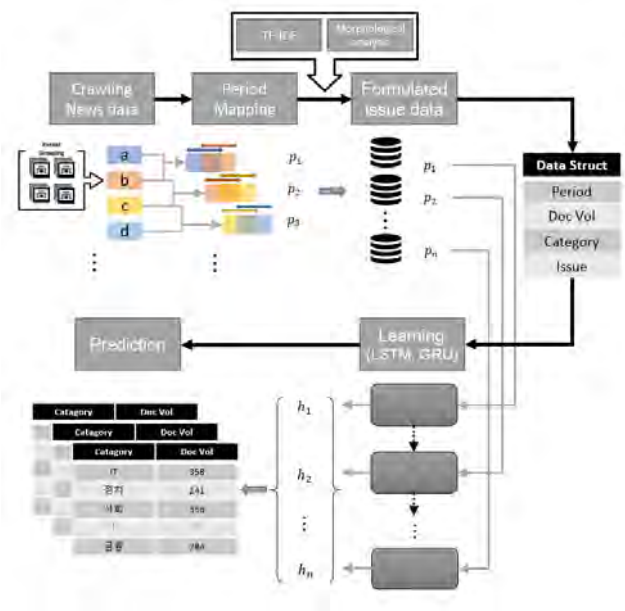
## 1. 서론

현대 사회에서 스마트 기기의 보급률이 높아짐에 따라 사람들은 뉴스 사이트, SNS, 커뮤니티 등 웹 미디어에서 정보를 얻고 공유한다. 이렇게 생성된 데이터들은 대부분 비정형 형태의 데이터로 이루어져있다. 특히 빅데이터의 중요성이 부상하고 있는 가운데 비정형 데이터를 정형화, 분석을 통해 가치 있는 데이터로 바꾸기 위한 노력이 꾸준히 이루어지고 있다[1].

빅데이터 중 텍스트 데이터는 사람들이 가장 널리 사용하는 표현의 수단으로 텍스트 마이닝(Text Mining)을 통해 분석되어 왔다. 텍스트 마이닝은 비정형 데이터를 정형화 한 뒤 특징을 추출한다. 추출된 특징을 통해 유의미한 정보를 얻는 기술이다. 그중 최근 포털 사이트의 검색어나 SNS에서 언급되는 주요 검색어나 키워드에 대한 분석을 통해 사회적 이슈나 동향을 파악하려는 시도가 이루어지고 있다[2-3]. 이런 모습은 네이버, 구글 등에서 네이버 랩, 구글 애널리틱스와 같은 형태를 예로 들 수 있다.

본 논문에서는 해당 기간의 이슈나 사건을 잘 보여주는 뉴스 사이트의 기사를 통해 분석하고자 한다. 뉴스 기사 데이터를 수집하여 일정 기간으로 나눈 다음 인위적으로 기간을 중첩시켜 매핑한다. 매핑된 기사들에서 형태소 분석 및 TF-IDF를 통해 빈도수를 기반으로 이슈를 추출한다[4]. 추출한 이슈를 이슈 간 매핑을 통해 이슈들의 관계를 분석한다. 이후 시계열 데이터에 적합한 RNN 기반의 LSTM 또는 GRU를 이용한 학습을 통해 이전 이슈들을 통해 앞으로의 이슈가 어떤 카테고리의 기사로 도출될지 예측하는 모델을 제안한다.

## 2. 본론




(그림 1) 시스템 개요도

(그림 1)은 본 논문에서 제안하는 시스템 개요도이다. 시스템은 크게 데이터 크롤링, 기간별 매핑, 데이터 전처리, 학습 하는 부분으로 나누어진다.

### 2-1. 데이터 크롤링

수집하는 데이터는 한국인이 가장 많이 사용하고 있는 포털인 네이버의 기사를 이용한다. 사용할 웹 크롤러는 Python의 Selenium을 사용한다. Selenium은 크롬드라이버를 이용해서 크롬을 제어해 크롤링을 진행한다. 크롤링을 이용해 수집할 데이터는 기사 번호, 날짜, 카테고리, 기사 내용을 수집하여 저장한다.

### 2-2. 기간별 매핑

저장한 데이터는 일정 기간으로 나누어 그룹으로 만든다. 그룹으로 묶인 기사 데이터는 인위적으로 기간을 겹치게 매핑한다. 이는 시간흐름에 따라 이슈의 변화를 관찰하기 위함이다. 인위적으로 중첩되는 기간을 만듦으로써 앞선 기간의 이슈들이 어떤 형태의 이슈로 변화하는지를 분석 할 수 있다.

### 2-3. 데이터 전처리

매핑된 데이터들에서 형태소 분석 및 TF-IDF를 통해 이슈를 추출한다. 형태소 분석에는 Python의 KoNLPy 패키지를 사용한다. KoNLPy를 이용해 불용어와 특수문자 등 분석에 방해되는 글자를 제거하고 품사로 분류한다[5].

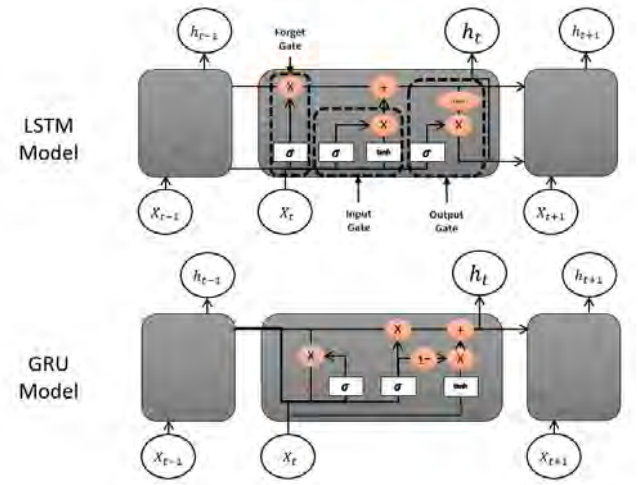
그 다음 TF-IDF를 통해 이슈를 추출한다. TF-IDF는 특정 단어의 중요도가 출현 횟수에 비례하고 단어가 언급된 모든 문서에 반비례하는 것에 기초하고 있다. TF-IDF의 식은 다음 (1)과 같다.

$$w_{x,y} = tf_{x,y} \times \log(\frac{N}{df_x}) \qquad (1)$$

$tf_{x,y}$는 y문서에 x가 얼마나 자주 등장하는지를 의미한다. $df_x$는 x를 포함하는 문서들의 수를 의미하며 N은 총 문서의 수를 의미한다. 따라서 문서에 자주 등장하는 단어들을 추출한 다음 전체 문서에서 자주 등장하는 단어를 제하는 방법이다. 이를 통해 문서 전체에 등장하는 단어들 중 관사, 접속어 등을 제외한 단어 키워드를 추출 할 수 있다. 이슈 키워드 추출은 각 매핑된 데이터와 데이터 전체에서 실시한다. 이를 통해 학습 모델에 들어갈 이슈 키워드들을 벡터화 시킨다.

### 2-4. RNN

본 실험에서는 기간별로 나누어 매핑한 시계열 데이터를 이용해 학습하기 때문에 학습을 진행할 딥러닝 모델은 시계열 데이터에 적합한 RNN(Recurrent Neural Network)기반의 LSTM(Long short-term Memory) 또는 GRU(Gated Recurrent Unit)를 사용한다. RNN은 순환신경망으로 입력을 받아 출력을 하지만 이 출력을 다시 입력으로 받는다. 따라서 과거의 출력 데이터를 이후 학습에 활용하면서 연속적이거나 시계열 데이터에 높은 성능을 보인다. 하지만 타임 스텝이 길어질수록 앞쪽 입력 값에 대한 영향이 최근 학습에 영향을 주지 못하는 문제가 발생하게 된다. 이를 장기 의존성 문제라고 한다.




(그림 2) LSTM 모델 과 GRU 모델 예시

(그림 2)와같이 LSTM과 GRU는 이런 장기 의존성문제를 해결하고 자 나온 모델이다. LSTM은 크게 망각 게이트, 입력 게이트, 출력 게이트 3개로 구성되어있다. 망각 게이트는 이전 출력을 얼마나 학습에 반영할건지 제어한다. 입력 게이트는 새로운 입력 값을 장기기억 셀에 얼마나 추가하는지 결정하는 과정이다. 출력 게이트는 장기기억 셀을 불러와 학습 결과를 출력하는 과정이다. 이런 과정을 통해 장기 의존성을 해결하였다. GRU는 LSTM을 간소화한 모델로 장기기억 셀과 출력 게이트를 없애고 하나의 게이트 컨트롤러가 망각 게이트와 입력게이트를 제어한다. 또한 출력 게이트가 없기 때문에 이전 기억을 제어하는 리셋 컨트롤러가 있다.

GRU가 LSTM에 비해 파라미터 수가 적어서 학습속도가 짧지만 충분한 데이터가 있을 경우 LSTM의 모델링 결과가 더 좋다[6]. 따라서 본 실험에서는 두 모델 모두 실험을 하여 보다 나은 성능을 보이는 모델은 선택한다. 두 모델에 기간, 카테고리, 문서 량, 이슈를 입력 값으로 넣어 학습을 진행한다. 하이퍼파리미터 값으로는 학습 계수는 0.01을 사용하며, 활성화 함수는 tanh보다 안정적인 softsign 함수를 사용한다.

## 3. 결론

본 실험에서는 이슈가 시간의 흐름에 따라 어떤 카테고리의 기사로 도출될지를 예측하는 시스템을 설계한다. 가장 많이 활용하는 네이버 포털 뉴스 기사를 크롤링하여 데이터를 수집하고 매핑과 정규화를 통해 학습에 활용한 입력 값을 만들었다. 이를 LSTM 또는 GRU를 이용해 학습시켜 카테고리 별 문서 량을 예측하는 시스템을 설계하였다. 본 논문에서는 뉴스 기사를 일정 기간으로 나누어 매핑 하여 시간 흐름에 따른 이슈 데이터를 학습 입력 값에 사용함으로써 과거의 이슈가 현제에 어떤 분야 기사에 영향을 미치는지 알아볼 수 있을 것이다.

나아가 본 논문의 학습에 있어 설정 가능한 하이퍼파라미터 값은 본 논문에서 설계하며 임의의 값을 넣어 진행하였지만 추후 실험을 진행하며 최적화된 값과 함수를 찾는 과정을 통해 학습 모델의 성능을 향상시킬 수 있을 것으로 기대된다.

## ACKNOWLEDGEMENT

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2017R1A2B4008886).

## 참고문헌

[1] 이애리, 이주원. 소셜 빅 데이터를 이용한 상권 확장 트렌드 및 소비 트렌드 분석. e-비즈니스연구, 19(6), 401-413, 2018

[2] 임명수, 김남규. "비정형 텍스트 분석을 활용한 이슈의 동적 변이과정 고찰." 지능정보연구, 22.1, 1-18. 2016.03

[3] 강민석, 윤소혜, 조건형, 박석. 시계열 데이터 분석을 통한 뉴스 사이트에서의 이슈 트래킹 시스템. 한국정보과학회 학술발표논문집, 277-279. 2015

[4] 이성직, 김한준. TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법. 한국전자거래학회지, 14(4), 59-73. 2009

[5] 박은정, 조성준. KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지. 제26회 한글 및 한국어 정보처리학술대회 논문집, 133-136. 2014

[6] Jozefowicz, Rafal, Wojciech Zaremba, and Ilya Sutskever. "An empirical exploration of recurrent network architectures." International Conference on Machine Learning. 2015