문서 임베딩 및 키워드 추출을 통한 신문기사 추천 및 큐레이션 시스템*

오원식, 조세은, 송지윤, 안혜빈, 이오준†
가톨릭대학교 인공지능학과
e-mail : {dodow, seeun1197, sara4423, ahb020303, ojlee}@catholic.ac.kr

News Article Recommendation and Curation System based on Document Embedding and Keyword Extraction

Won-Sik Oh, Se-Eun Cho, Ji-yoon Song, Hye-Bin Ahn, O-Joun Lee
Dept. of Artificial Intelligence
The Catholic University of Korea

요 약

본 논문은 기사와 그 댓글들을 수집하는 기능과 수집된 데이터를 효과적으로 탐색할 수 있도록 하는 내부 데이터에 의한 Doc2Vec 기반 기사 추천과 TF-IDF를 사용한 키워드 추출 기능, 기존 시스템의 단점을 개선한 신규 추천 기사 UI, 마지막으로 처리된 내부 정보를 표시하는 웹 인터페이스를 구현하는 통합 체계를 구현하였다. 최근 코로나19 확산으로 인터넷 뉴스를 이용하는 사용자가 증가함[1]에 따라 뉴스 포털의 필요성이 드러나고 있다. 하지만 유사 시스템을 제공해주는 대표 서비스 중 네이버 뉴스, 다음 뉴스는 기사, 주제별 탐색의 기능이 부족하다. 따라서 본 논문에서는 사용자가 간단한 입력으로 원하는 기사들을 수집하고, 수집된 기사들의 키워드, 카테고리, 유사 기사들을 분석해 다양한 방식으로 탐색할 수 있도록 하는 웹 인터페이스를 구현하고자 하였다. 본 논문에서 기사별 대표 키워드 분석과 지역별 키워드 분석은 전처리를 거친 TF-IDF 방법을 사용하였다. 특정 기사에 관한 상위 기사 추천은 Doc2Vec의 기사 간 벡터 유사도 비교로 구현하였다. 본 논문의 결과로 사용자의 관심 있는 기사들을 수집해 이를 분석하고, 다양한 탐색 기능을 제공하는 뉴스 인터페이스를 구현할 수 있다.

1. 서 론

최근 인터넷 뉴스를 이용하는 사용자가 증가하는 추세를 보이고 있다[1]. 또한 각 분야와 주제별로 많은 양의 인터넷 뉴스가 매일 업로드되고 있어 사용자의 선택지 또한 넓어지고 있다. 이런 상황에서 사용자는 효율적인 선택이 가능하게 해 주는 기제들을 적극적으로 활용하고자 할 것이고[2], 따라서 여러 신문 기사별 뉴스를 모아주는 뉴스 포털의 필요성이 부각되고 있다. 그러나 유사 서비스인 '네이버 뉴스'의 경우, 뉴스별 키워드 기능은 존재하지 않고, 또한 추천 시스템도 뉴스 내용 관련 추천보다는 현재 인기 있는 뉴스 위주로 제공하고 있다. 그리고 '다음 뉴스'의 경우, 모든 뉴스에 대해 같은 내용의 추천 뉴스를 제공하고 있다. 그러므로 사용자는 사건별, 키워드별로 뉴스를 찾아볼 때 불편함을 느낄 것이다. 따라서 본 연구에서는 뉴스당 3개의 맞춤 추천 뉴스와 추천 뉴스별 중요 댓글을 같이 표시해 사용자에게 추천 기사의 효율적인 선택이 가능해지도록 하였으며, 사용자가 간단한 입력을 통해 사건별, 키워드별로 원하는 뉴스를 수집할 수 있도록 설계했다. 이와 더불어 수집한 뉴스별 키워드와 지역별 키워드들을 분석하고 다양한 방식으로 뉴스를 탐색할 수 있는 웹 인터페이스를 구현하였다.

2. 관련 연구

기사 데이터셋에 대해 학습한 Word2Vec을 사용해 LDA 과정을 통해 분류된 토픽별 대표 키워드와 해당 토픽으로 분류된 기사별 단어 유사도, 단어 빈도수를 곱해 나온 수치로 기사를 추천한 김선미의 논문[3]의 경우 단순 단어별 Word2Vec 유사도 분석 스코어는 기사적 단어 사용 문맥이 고려될 수 없고, 또한 LDA에 의해 같은 토픽으로 분류되지 않은 기사에는 유사도 계산이 진행되지 않으며 새로운 기사에 대한 유사도를 계산하려면 다시 전체 기사 데이터에 대한 학습 프로세스를 진행해야 한다. 따라서 본 논문에서는 Doc2Vec[4]을 통해 문서를 벡터화 시키는 방법을 사용하였다. 프로세스가 더 간단해졌을 뿐만 아니라 새로운 문서에 대해서도 바로 유사도를 측정할 수 있고 모든 문서에 대한 유사도 측정이 가능하게 되었다.

3. 구현

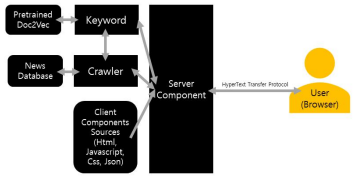
3.1 시스템 구조

크롤러(Crawler) 컴포넌트는 터미널 입력을 통해 기사를 크롤링하고 기사별로 자체적인 ID를 매겨 관리한다. **키워드(Keyword)** 컴포넌트는 Doc2Vec 분석, TF-IDF 키워드 추출 메소드들이 구현된 컴포넌트이다. **클라이언트(Client)** 컴포넌트는 크롤러, 키워드 컴포넌트의 정보와 결합하여 사용자의 브라우저에서 인터프리트(Interpret)될

---

* 학문후속세대(학부생) 논문
† 교신저자

HTML, Javascript, CSS 소스 코드들을 얘기한다. **서버(Server)** 컴포넌트는 이러한 키워드와 크롤러 컴포넌트의 연산 결과물을 클라이언트 컴포넌트와 결합해 사용자에게 전송시켜 주고, 사용자가 요청하는 정보를 처리한다. 이 과정은 HTTP 통신상에서 이루어진다.



(그림 1) 본 시스템의 작동 과정 도식화

3.2 Doc2Vec 기반 기사 추천

크롤러 컴포넌트에 저장돼있는 기사 데이터를 데이터셋으로, 기사별 제목과 본문 내용을 더해 하나의 문서로 구성하고 이를 크롤러 컴포넌트에서 사용하는 내부 기사 Id로 태깅시켜 학습을 진행하였다. 기사 추천은 기사별 벡터 간 유사도를 구해 상위 3개를 추천하는 방식으로 이루어졌다.

3.3 키워드 추출

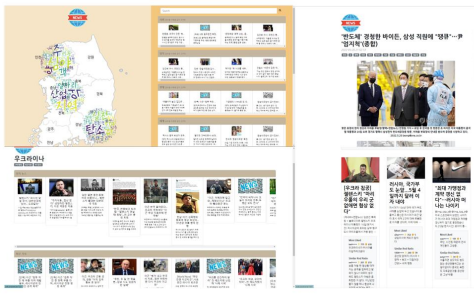
기사별 키워드 추출은 기사 본문의 명사들을 추출하고 그 뒤 추출된 명사 리스트들에 대해 TF-IDF 계산을 진행했다. 그 뒤 계산된 문서 단어별 Frequency에 대해 상위 10개를 키워드로 선정하여 기사별 데이터셋에 저장하였다.

지역별 키워드 추출 또한 검색어 ['서울', '대구', '인천'] 등을 선정해 검색어에 관한 상위 기사를 크롤링하고 TF-IDF를 통해 키워드를 추출했다. 추출된 키워드별 Frequency를 Wordcloud의 글자 크기 값으로써 사용해 사용자에게 제공하였다.

3.4 추천 기사 UI

비교 대상 서비스인 네이버 뉴스와 다음 뉴스는 추천 기사에 대해 섬네일 이미지와 제목만을 제공한다. 따라서 우리는 해당 디자인이 사용자가 추천 기사를 선택하는데 충분한 정보를 담고 있지 않다고 결론을 내리고, 본 논문의 추천 기사 섹션은 각 추천 기사별 제목, 이미지, 본문 내용의 요약과 기사별 제일 좋아요가 많은 댓글과 좋아요, 싫어요 비율이 비슷한 댓글 또한 표시해 탐색을 더 쉽게 하였다.

3.5 웹 인터페이스 구현



(그림 2) 실제 시스템의 구현 화면

웹 인터페이스는 총 4가지 화면으로 지역별 키워드와 카테고리별 기사 표시 그리고 키워드 검색 기능이 있는 **메인화면**(그림 2 왼쪽 위 끝), 키워드별 인기 기사와 최신 기사를 확인할 수 있는 **키워드 화면**(왼쪽 하단), 기사 내용과 댓글, 해당 기사에 관한 추천 기사들을 확인할 수 있는 **기사별 화면**(오른쪽 상단)들이 있다. 오른쪽 하단은 **추천 기사 섹션**이다.

4. 실험 및 검증

본 논문에서 제안하는 기사 추천 시스템과 키워드 추출, 새로운 추천 기사 UI의 선호도 및 정확도를 검증하기 위해 설문조사를 진행했다. 설문 대상은 국내 재학 학부생 18명이었다. 문항은 1~5점 척도로 숫자가 클수록 긍정적이다. 해당 설문 진행을 위해 구현한 소스 코드는 Github 리포지터리 형태로 공개되어 있다.*

질문(x>=4)	1	2	3	4	5	계
기사 추천 정확도(77.8%)	0	1	3	8	6	18
네이버 뉴스 추천 시스템과의 비교(61.1%)	3	3	1	3	8	18
키워드 정확도(94.4%)	0	0	1	11	6	18
키워드 화면별 기사 정확도(77.8%)	0	0	4	6	8	18
추천 기사별 대표 댓글 표시의 유능성(77.8%)	0	1	3	6	8	18
지역별 키워드의 정확도(38.9%)	1	1	9	3	4	18

(표 1) 설문조사 결과 (괄호: 4점 이상 표시한 참가자의 비율), n=18

네이버 뉴스의 기사 추천 시스템보다 본 논문의 기사 추천 시스템을 선호하였고(61.1%) 키워드 추천 시스템에 대한 정확도를 높게 평가해(94.4%) 본 논문이 제안하는 시스템이 유사 뉴스 포털 시스템으로써 사용되기에 적합하다고 판단할 수 있다.

5. 결론

기존 뉴스 포털 시스템은 사건별, 키워드별로 뉴스를 찾아볼 때 불편함이 존재한다. 본 논문에서는 이를 해결하기 위해 기사별 맞춤 기사 추천 시스템과 기사별, 지역별 키워드 추출 방법, 새로운 추천 기사 UI 설계를 진행하고 이를 실제로 구현해 설문을 통해 검증하였다. 한계점은 실험 검증이 자가 설문으로만 이루어졌기 때문에, 향후 연구에서는 기사별 Label을 매겨 객관적인 정확도를 매겨 실험을 검증할 것이다.

참 고 문 헌

[1] 한국언론진흥재단, "전통매체 뉴스 인터넷 뉴스 이용률 추이 (2011~2020)", http://hannun.or.kr/2021/3-2/

[2] 양정애.(2011).뉴스 기사의 현저성과 이용자의 선택적 노출.한국방송학보, 25(2),77-117.

[3] 김선미, 나인섭, 신주현.(2019).단어 연관성 가중치를 적용한 연관 문서 추천 방법. 멀티미디어학회논문지,22(2),250-259.

[4] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International conference on machine learning. PMLR, 2014.

* https://github.com/dnjstlr555/CountryWideTopics