JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 21, No. 3, pp. 479-486, Mar. 2020

Check for updates

텍스트와 이미지 태그 데이터에 기반한 뉴스기사 추천시스템

정 인 정 · 김 보 미 · 김 수 경 · 유 견 아*
덕성여자대학교 컴퓨터공학과

News Recommendation System Based on Text and Image Tag Data

In-Jeong Jeong · Bo-Mi Kim · Su-Kyung Kim · Kyeonah Yu*
Department of Computer Engineering, Duksung Women's University, Seoul, Korea

[요 약]

오늘날 온라인에서는 매일 엄청난 양의 뉴스가 생성되고 있으며, 따라서 개인 맞춤형 뉴스 서비스의 중요성은 증대하고 있다. 본 논문에서는 뉴스 기사의 속성을 반영한 뉴스 기사 추천시스템을 개발한다. 이 시스템에서는 기사 자체의 선호도가 아닌, 기사의 태그에 대한 사용자의 선호도를 계산하여 뉴스를 추천하며, 태그 구독 서비스를 제공하여 기존의 시스템과의 차별화를 꾀한다. 또한, 텍스트에서 추출한 태그뿐만 아니라 기사 이미지에서 추출한 태그를 함께 활용하고, 모델 기반 협업 필터링인 SVD 방식으로 사용자의 선호 태그를 추천한다. 이처럼 태그에 기반한 추천 방식은 변화가 빠른 기사 추천의 어려움을 보완하는 동시에 사용자-기사 관계의 특징인 데이터 희소성 문제를 해결해 준다. 또한, 텍스트가 아닌 기사의 사진에서도 태그를 추출함으로써 한국어 형태소 분석 시 발생하는 오류로 인한 문제를 완화하여 태그 추천의 정확도를 향상시킬 수 있음을 보인다.

[Abstract]

With huge amounts of news being generated online every day today, the importance of personalized news services is increasing. In this paper, we develop a news article recommendation system that reflects the properties of news articles. This system recommends news by calculating the user's preference for the article's tag, not the article's own preference, and provides a tag subscription service to differentiate itself from the existing system. In addition, the tag extracted from the article image as well as the tag extracted from the text are used together, and the user's preferred tag is recommended by the SVD method, which is a model-based collaborative filtering. This tag-based recommendation complements the difficulty of fast-changing article recommendations, while addressing the data scarcity problem, a feature of user- article relationships. It is also shown that extracting tags from pictures of articles as well as text can improve the accuracy of tag recommendations by mitigating problems caused by errors occurring during the Korean morphological analysis.

색인어 : 추천시스템, 이종 데이터 활용, 자동 태깅, 뉴스 기사 추천, 기사 요약
Key word : Recommendation System, Using Heterogeneous Data, Automatic Tagging, Recommending News Articles, Article Summarization

http://dx.doi.org/10.9728/dcs.2020.21.3.479

Received 23 January 2020; **Revised** 15 March 2020
Accepted 25 March 2020

***Corresponding Author; Kyeonah Yu**

Tel: +82-2-901-8346
E-mail: kyeonah@duksung.ac.kr

Ⅰ. 서 론

추천시스템이란 정보 필터링의 일종으로 사용자와 상품의 정보를 기반으로 하여 특정 고객이 좋아할 만한 상품을 찾아주는 것을 뜻한다[1]. 추천시스템은 도서, 의류, 영화, 음악 등의 전자상거래에서의 상품 추천뿐만 아니라 블로그, 사람, 논문 등으로 그 활용 범위가 넓어졌으며 온라인 뉴스도 그중의 한 분야이다. 특히 최근에는 온라인으로 뉴스 기사를 읽는 것이 보편화함에 따라 추천시스템을 활용한 뉴스 추천 서비스의 중요성은 더욱 강조되고 있다 [2].

뉴스 추천은 다른 아이템들의 추천과 구분되는 여러 가지 특징이 있다 [3]. 우선 뉴스라는 아이템은 매우 변화가 빠르므로 실시간으로 생겨나는 기사를 모두 보여 주는 것은 거의 불가능하다는 것이며 두 번째 특징은 기사들은 서로 연관성이 높고 사용자들의 선호도만 학습하는 것은 불충분하다는 것이다. 마지막으로 기사의 수는 엄청 많고 사용자와 뉴스 기사 행렬의 엔트리가 비어 있는 데이터 희소성 문제(data sparsity problem)가 특히 심하다는 특징이 있다. 이를 해결하기 위해 본 논문에서는 SNS에서 사용하는 메타데이터인 해시태그 개념을 활용한 뉴스 기사 추천시스템을 제안하고자 한다. 해시태그는 단어나 구절 앞에 기호 #을 붙여서 만든 메타데이터로 동일한 해시태그를 이용하여 게시물을 빠르게 검색하고 주제별로 그룹화하여 볼 수 있게 한다. 이때 중요한 것은 게시물에 적절한 해시태그를 다는 것이다.

본 논문에서는 기사 자체의 선호도가 아닌, 기사의 태그에 대한 사용자의 선호도를 계산하여 뉴스 추천과 태그 구독 서비스를 제공하는 추천시스템을 제안한다. 기사의 키워드를 이용한 기존의 콘텐츠 기반 추천과는 다르게[4] 기사 텍스트의 키워드뿐만 아니라 연관검색어 태그 및 이미지로부터 얻은 태그를 함께 활용하여 뉴스 기사에 효율적인 해시태그 역할을 수행할 수 있도록 한다. 사용자의 선호 태그를 추천하는 방식은 가장 보편적으로 많이 사용되는 협업 필터링(Collaborative Filtering) 방식 중에 모델 기반인 절사형 SVD(truncated Singular Value Decomposition) 추천 알고리즘을 이용하여 사용자와 아이템 간의 희소 데이터 문제를 해결하는 동시에 사용자의 잠재적인 선호 특징을 찾아낼 것을 제안한다. 또한, 특정 뉴스 기사를 추천하는 것이 아니라 태그를 추천함으로써 뉴스 기사 추천에서 문제점으로 지적되고 있는 정보의 편향적 제공 문제를 피할 수 있게 한다. 마지막으로 추천한 기사에 대한 요약 서비스를 제공하여 일일이 기사를 클릭하여 기사 페이지를 확인하지 않고도 기사 내용을 알 수 있도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 협업 필터링 방식과 뉴스 추천시스템에 대한 관련 연구를 살펴보고 3장에서는 제안하는 뉴스 기사 추천시스템의 개요와 이종 데이터로부터 기사의 태그를 추출하는 방법을 설명한다. 4장에서는 추천시스템을 구현한 방법과 실행 결과의 예시들에 관해 설명하고 실제 사용자 테스트를 통해 텍스트만 이용한 기존 방식 대비 성능 향상 결과를 제시하며 5장에서는 본 시스템의 장단점을 분석하고 향후 연구에 대한 제언과 함께 논문을 마무리한다.

Ⅱ. 관련 연구

현대 대부분의 추천시스템은 협업 필터링 기법을 주로 사용한다. 협동 추천이라고도 불리는 협업 필터링 방식은 영화 추천 시스템인 GroupLens 프로젝트[5]에서 처음 사용한 이래 다양한 방식으로 발전해왔다. 협업 필터링은 일반적으로 사용자 기반 필터링과 아이템 기반 필터링으로 나뉜다. 사용자 기반 필터링은 유사한 성향을 지닌 사람들을 구분하고, 해당 성향의 사람들이 좋아하는 것을 이용해 추천하는 방식이며 아이템 기반 필터링은 사람이 아닌 아이템 간의 유사도를 이용해 추천하는 방식이다 [6].

추천시스템은 여러 기업에서도 활용되고 있다. 추천시스템을 적극적으로 이용하기 시작한 최초의 온라인 사이트인 아마존의 경우 평점, 구매행위, 검색행위 정보들을 이용해 추천시스템을 운영하고 있으며, 특히 평점을 명시적 평점(Explicit Rating)과 암묵적인 평점(Implicit Rating)으로 구분해 9가지 온라인 추천 방식에 활용하고 있다. 비디오 스트리밍 회사인 넷플릭스 역시 추천 서비스를 적극적으로 활용하고 있는 회사로서 추천시스템을 이용해 사용자의 성향을 파악하여 로그인 순간 좋아할 만한 영화로 전체 페이지를 구성한다. 페이스북에서는 가입자들의 교류를 증대시킬 목적으로 전통적인 상품 추천과는 다른 친구 추천이라는 분야를 개척한다 [7].

인터넷을 통해 뉴스 기사를 읽는 것이 보편화함에 따라 추천시스템을 활용한 뉴스 추천 서비스가 생겨나기 시작했다. 구글 뉴스에서는 MinHash와 PLSI(Probabilistic Latent Semantic Indexing)를 통해 사용자 집단을 클러스터링하였으며 아이템 간 상호방문을 이용해 개인화된 뉴스 추천시스템을 구축하였으며[8], [9]에서는 이를 발전시켜 사용자의 클릭 기록(Click Log)을 분석해 사용자의 뉴스 관심사를 파악하였는데 여기서는 사용자의 관심사 예측을 위해 베이지안 프레임워크(Bayesian Framework) 개발하였고 내용 기반 방식과의 하이브리드 방식을 이용하였다. 추천시스템에서는 사용자와 아이템의 개수가 늘어날수록 사용자가 실제로 이용한 아이템의 비율이 점점 줄어들어 추천 피드백 정보가 희박해지는 데이터 희소성 문제가 발생한다. [10]에서는 딥러닝 모델을 이용한 해결을 시도하였으며 RNN(Recurrent Neural Network)을 이용해 사용자의 뉴스 소비 패턴을 분석하고 사용자가 볼 뉴스를 예측하여 추천한다. 가장 많이 사용되는 모델 기반 협업 추천시스템 방식에서는 전통적인 SVD를 사용하는 것이 아니라 확률적 그레디언트 디센트 방법에 따른 점진적 SVD 방식으로 구현한다 [11]. 현재까지 개발된 대부분의 추천시스템들은 뉴스 기사의 텍스트 데이터에 기반하여 사용자와 아이템을 구분하고 있다. 본 논문에서는 뉴스 기사의 텍스트 데이터 뿐만 아니라 기사 이미지 데이터를 CNN(Convolutional Neural Network)으로 분석한 결과를 함께 추천시스템 구축에 활용하여 기사 자체뿐 아니라 선호 태그를 추천하는 방법을 제안한다.

Ⅲ. 텍스트와 이미지 데이터를 활용한 뉴스 기사 추천시스템

3-1 전체 시스템 개요

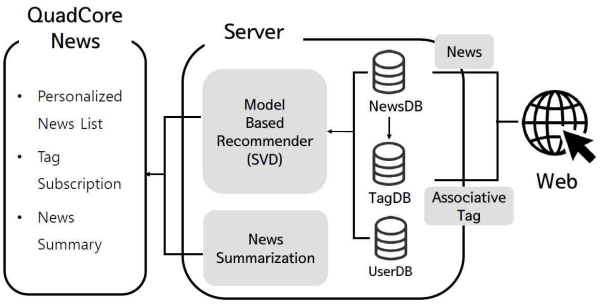

그림 1. 뉴스 추천시스템 구성도
Fig. 1. System Diagram

전체 시스템 구성도는 그림 1과 같다. 쿼드코어 뉴스는 웹 브라우저에서 사용할 수 있으며 웹 서버에서 사용자에게 추천하는 뉴스 및 태그 추천 서비스와 기사 요약 서비스를 제공한다. 웹 서버에는 추천에 사용하는 3종류의 태그 데이터베이스가 있는데 기사에서 태그를 추출하는 과정은 그림 2와 같다.

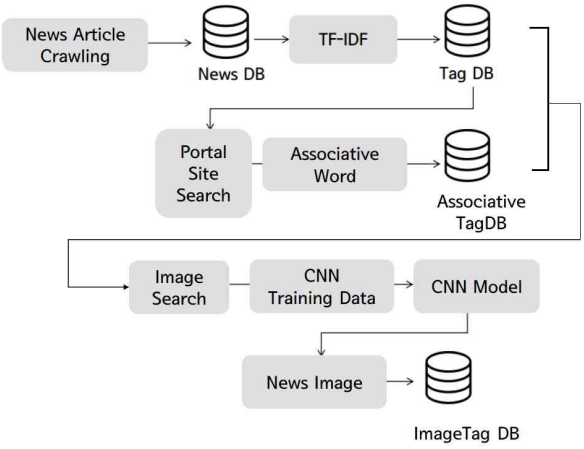

그림 2. 태그 추출 과정
Fig. 2. Tag Extraction Process

뉴스 기사는 네이버 뉴스에서 크롤링해 데이터를 수집하였고 TF-IDF(Term Frequency-Inverse Document Frequency)를 이용해 각 기사에서 중요도가 높은 명사를 태그로 추출한다. 이렇게 추출한 태그는 다시 네이버에 검색해 연관검색어를 가져와 '연관검색어 태그'를 구성하며 'TF-IDF 태그'와 '연관검색어 태그'를 구글에 검색해 얻은 이미지를 CNN 모델의 학습데이터로 활용하였다. CNN 모델을 기사 이미지에 적용해 기사 이미지에서 '이미지 태그'를 추출한다. 추천시스템 구현을 위해 모델 기반 협업 필터링 방식 중 하나인 SVD 추천 알고리즘을 이용하여 사용자가 읽은 태그 기록을 기반으로 추천 알고리즘에 반영한다.

3-2 텍스트로부터 태그 추출

본 논문은 기사의 텍스트로부터 태그를 추출하기 위해 TF-IDF(Term Frequency-Inverse Document Frequency) 알고리즘을 사용하였다. TF-IDF는 텍스트 마이닝에 주로 사용되는 알고리즘으로 여러 개의 문서가 있을 때, 문서 내에 있는 단어들에 가중치를 적용하여 상대적으로 중요한 단어를 알아낼 수 있다 [12]. TF(Term Frequency)는 문서 내에 해당 단어의 빈도수를 나타낸다. IDF(Inverse Document Frequency)는 DF의 역수를 의미하는데 DF는 전체 문서에서 특정단어 t가 포함된 문서의 수의 비율이기 때문에 IDF는 이의 역수인 [전체 문서의 수/단어 t가 포함된 문서의 수]가 된다. 본 논문에서 전체 문서의 수는 카테고리별 총 기사의 수로 정하였다. 이렇게 구해진 TF 값과 IDF 값을 곱한 값이 TF-IDF 값이 된다. 한글 뉴스의 TF-IDF 계산을 위해서는 한국어 정보처리 파이썬 패키지인 KoNLPy (Korean Natural Language Processing in Python)를 적용하여 형태소 분석을 하고 문서 내 명사들을 추출하였다 [13].

본 논문에서는 TF-IDF 태그와 더불어 연관검색어를 추가로 활용한다. 앞서 TF-IDF를 이용해 추출한 상위 k개의 단어를 포털사이트에 검색하고 결과로 나오는 연관검색어를 직접 개발한 크롤러를 통해 수집한다. 이렇게 수집한 검색어를 '연관검색어 태그'로 사용한다. 이를 통해 기사 본문에는 포함되지 않았던 연관 단어 또한 추천시스템에 활용하게 됨으로써 다른 단어나 표현을 사용했지만, 내용상 관련 있는 기사들까지 찾아서 추천할 수 있다.

3-3 이미지로부터 태그 추출

본 논문에서는 기사의 텍스트뿐만 아니라 이미지 데이터에서 시각적 이미지를 분석하는 데 가장 일반적으로 사용되는 인공신경망의 한 종류인 CNN 딥러닝 모델을 이용해 의미 있는 값을 추출한다. CNN은 일반적으로 합성곱 계층(Convolutional Layer)과 풀링 계층(Pooling Layer), 완전 연결망 등으로 구성되는데 본 논문에서는 구글의 인셉션 v3 모델을 사용하여 학습하였다. 인셉션 v3 모델은 이미지넷의 데이터에 대해 훈련된 컨볼루션 신경망으로서 기존의 CNN 모델들이 같은 크기의 필터를 여러 층 사용했다면 인셉션은 한 번에 여러 크기의 필터를 복합적으로 사용하여 이미지의 특징을 더 잘 잡아낸다고 알려져 있다. 인셉션 v3는 그림 3과 같은 인셉션 모듈의 여러 버전으로 총 48개의 계층으로 이루어져 있으며 2014년에 이미지넷 대회에서 우승한 이래 버전을 업그레이드하며 우수한 정확도를 나타내고 있는 모델이다 [14].

연관검색어 태그를 이용하여 관련 이미지 데이터를 수집하였고, 카테고리 수와 데이터의 수를 조정하기 위해 연관 검색어 내의 최상위 태그만을 이용하여 데이터 수집을 시행하였다. 최상위 태그의 이름을 영어로 변환하여 폴더 이름을 설정하였고

그 아래 수집 데이터를 저장했다. 수집 데이터를 이용해 학습 후 기사에 첨부되어 있는 기사 이미지에 적용하여 이미지 기반 태그를 추출한다. 이미지 태그를 추출함으로써 기사 본문에는 포함되어 있지 않은 단어를 태그로 얻어 활용할 수 있다. 예를 들어, 그림 4와 같이 기사 본문에는 포함되어 있지 않은 단어인 '음악회', '퍼포먼스', '공연'을 기사 이미지를 통해 추출하여 태그로 활용할 수 있다.

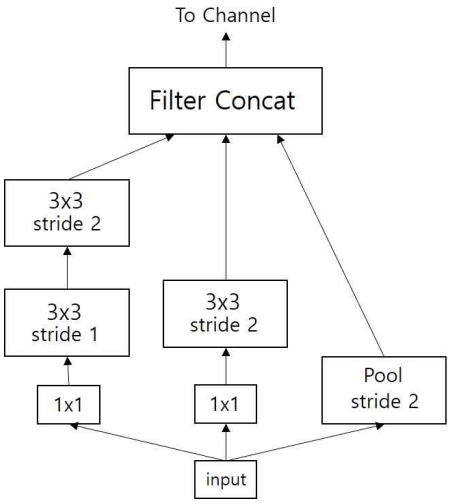

그림 3. 인셉션 v3 모듈 구성도
Fig. 3. Inception v3 Module Diagram

그림 4. 이미지 태그
Fig. 4. Image Tag

3-4 사용자 추천에의 적용

본 연구에서는 추천시스템에 많이 사용되는 모델 기반 CF 중 행렬 분해(Matrix factorization) 방법인 SVD 알고리즘을 사용하였다.

$$R = U\Sigma V^T \quad (1)$$

$$\hat{U}\hat{\Sigma}\hat{V}^T = \hat{R} \approx R \quad (2)$$

행렬 R은 m 사용자와 n 태그 사이 카운트 값이며 이 숫자가 클수록 태그의 중요도가 높다고 할 수 있다. 기본 SVD 방법은 R을 (1)과 같이 세 행렬 U, Σ, V의 곱으로 나타낸다. 여기서 U는 m×m, V는 n×n의 행렬로 역행렬이 대칭인 행렬이며 Σ는 대각선이 특이치(singular value)로 이루어지고 나머지는 0인 m×n 행렬이다. 본 논문에서는 데이터 희소성을 고려하여 전체 특잇값 중에 가장 값이 큰 k개의 특잇값만을 사용하는 truncated SVD로 (2)와 같이 근사행렬을 구한다. 추천에 사용할 때 평점이 많지 않은 희소 데이터의 차원 축소가 가능하다. (2)의 식에서 $\hat{U}$은 가장 큰 k개의 특이치에 대응하는 k개의 특징만을 남긴 m×k 크기의 행렬이고 $\hat{\Sigma}$ 와 $\hat{V}$는 각각 k×k, k×n 크기의 행렬이다.

사용자가 기사를 클릭하면 기사의 TF-IDF 태그, 기사 이미지의 이미지 태그, 연관검색어 태그와 함께 카운트 값이 저장된다. 파이썬 Surprise 패키지를 사용하여 태그를 항목(item) 값으로, 카운트를 평점(rate) 값으로 입력해 SVD 알고리즘을 실행한다. SVD에 의해 구해진 특이치를 이용해 기존의 사용자와 태그의 평점 데이터의 근사행렬을 구성하고 이를 통해 평점 데이터의 상위 n개의 태그의 값을 추천한다. 이렇게 예측된 태그들이 포함된 기사들로 메인 페이지를 구성하며 다양한 서비스를 제공하게 된다.

Ⅳ. 시스템 구현 및 테스트

4-1 시스템 구현

'QuadCore News' 시스템 구현 과정은 그림 5와 같다. Spring과 Bootstrap, HTML/CSS(cascading style sheets)/JavaScript를 이용해 웹으로 구현하여 웹 브라우저에서 사용할 수 있도록 하였다. 웹 서버는 AWS(Amazon Web Service) 서버를 사용했고 데이터베이스는 MySQL을 사용하였다. 뉴스 기사는 네이버 뉴스에서 크롤링해 데이터를 수집하였고 TF-IDF를 이용해 각 기사에서 중요도가 높은 명사를 태그로 추출하였다. 이렇게 추출한 태그는 다시 네이버에 검색해 연관검색어를 가져와 '연관검색어 태그'를 만들었으며 'TF-IDF 태그'와 '연관검색어 태그'를 구글에 검색해 얻은 이미지를 CNN 모델의 학습데이터로 활용하였다. CNN 모델은 텐서플로를 이용해 제작했다.

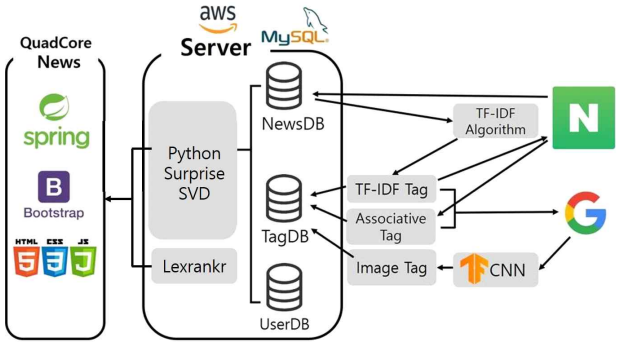

그림 5. 서비스 구현 과정
Fig. 5. Service Implementation Process

사용자가 메인 페이지에 리스트된 기사 중 하나를 선택하면 해당 기사의 페이지로 연결되면서 사용자-태그 테이블이 수정된다. 사용자-태그 테이블은 사용자 테이블과 태그 테이블의 관계 테이블로써 사용자가 읽은 태그를 저장한다. 저장되는 값으로는 사용자 ID, 태그 ID, 카운트 값이 있고 카운트는 태그를 읽은 횟수를 의미한다. 만약 현재 사용자가 선택한 기사의 태그가 테이블에 존재하지 않는다면 태그를 추가하고 카운트 값을 1로 설정한다. 그림 6은 사용자-태그 테이블이 갱신되는 방법을 보여 준다. 기존의 카운트 값에 곱해지는 α는 이전 데이터의 반영률을 나타내며 $0 < \alpha < 1$로 새로 추가되는 카운트의 영향이 가장 크게 유지되도록 한다. 이렇게 사용자가 기사를 선택할 때마다 업데이트된 사용자-태그 테이블은 기사 추천을 위한 SVD 계산에 정해진 주기마다 이용된다.

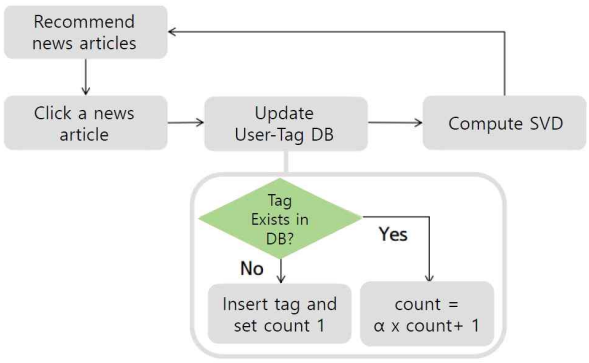

그림 6. 태그 갱신과 기사 추천
Fig. 6. Tag Update and News Recommendation

추천을 위해서는 파이썬 Surprise 라이브러리를 사용하여 SVD를 실행하였으며 사용자-태그 테이블을 입력 행렬 데이터로 사용하였다. 한글 요약 서비스를 위해 lexrankr[15]을 이용하였다. lexrankr은 한국어 요약을 위해 Lexrank 알고리즘을 한국어에 적합하도록 구현한 파이썬 패키지로 한국어 다중 문서에서 가장 높은 성능을 낸다고 알려진 기본 설정으로 이용하였다. 기사 요약을 위해 형태소 분석에 사용된 품사는 동사, 형용사, 명사이고 기사 문장 간의 유사도는 코사인 유사도를 사용하였다. 문장들을 노드로, 문장들 간의 유사도를 링크 값으로 그래프를 만들어 그래프 클러스터링을 수행한 후, 페이지 랭크(PageRank) 적용하여 각 클러스터에서 선택된 문장들의 일정 비율을 남기는 식으로 문장을 추출하였다.

4-2 시뮬레이션 결과

웹사이트로 구현한 본 논문의 시스템은 웹 브라우저를 통해 접속하여 로그인하면 사용자의 취향을 반영하여 추천된 기사들을 메인 페이지에 보여 준다 (그림 7). 웹 페이지 상단에는 메뉴가 있으며 메뉴에는 카테고리, 태그, 신문사, MyPage가 있다. 카테고리 메뉴를 선택하면 뉴스 기사를 카테고리 별(정치, 경제, 사회, 생활문화, IT 과학)로 선택하여 볼 수 있다. 태그 메뉴를 선택하면 추천 태그와 태그가 포함된 기사들만 따로 확인할 수 있다. 신문사 메뉴를 선택하면 신문사(조선일보, 중앙일보, 동아일보)를 선택하여 해당 신문사의 기사만 볼 수 있다. MyPage에서는 사용자가 스크랩한 뉴스의 목록과 구독한 태그 목록을 확인할 수 있다.

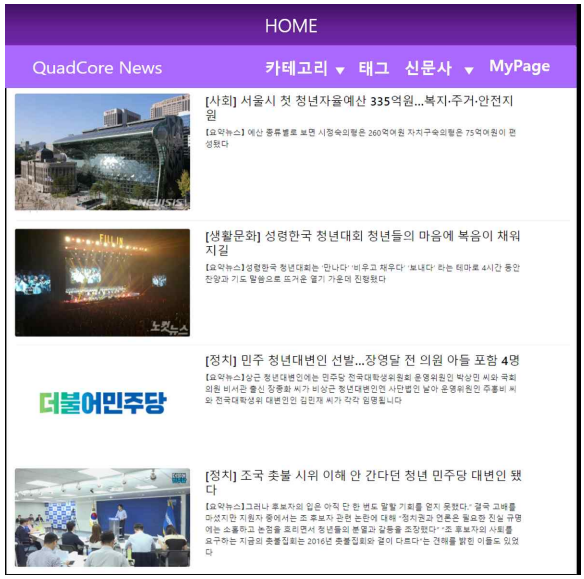

그림 7. 사용자 메인 화면
Fig. 7. User Main Screen

카테고리별로 기사를 보여 주는 방식은 사용자가 현재까지 읽은 기사의 태그 데이터를 기반으로 추천된 기사들을 기사의 태그와 요약문과 함께 표시해 주는 것이다. 그림 8은 '정치' 카테고리를 선택했을 때의 추천 화면이다. 읽은 기사의 태그들을 중심으로 추천된 태그인 '청년', '조국', '입시' 등이 포함된 기사들이 추천되었는데 다양한 카테고리에서 이들 태그가 포함된 기사가 추천된다.

각 기사 페이지에서는 기사와 함께 기사의 태그들을 보여 준다. 사용자는 기사를 읽고 관심사에 맞는 태그를 클릭하여 태그를 구독할 수 있다. 사용자가 기사 페이지에서 태그를 클릭하면 태그 구독 테이블에 저장된다. 그림 9와 같이 태그 페이지에서는 태그 구독 테이블을 참조해 뉴스 테이블 중 사용자가 구독한 태그가 포함된 기사들을 찾고 페이지에 표시한다. 태그 페이지에는 모든 태그들에 관한 기사를 모두 표시하고 하나의 태그에 관한 기사만 보고 싶다면 태그를 선택해 확인할 수 있다. 이처럼 태그 구독 서비스를 이용하여 사용자가 태그에 관련된 기사를 직접 선택함으로써 추천시스템에 의해 편향된 정보만 추천되는 뉴스 기사 추천시스템의 문제점을 보완할 수 있다.

그림 8. 카테고리별 추천 기사
Fig. 8. Recommended Articles of Each Category

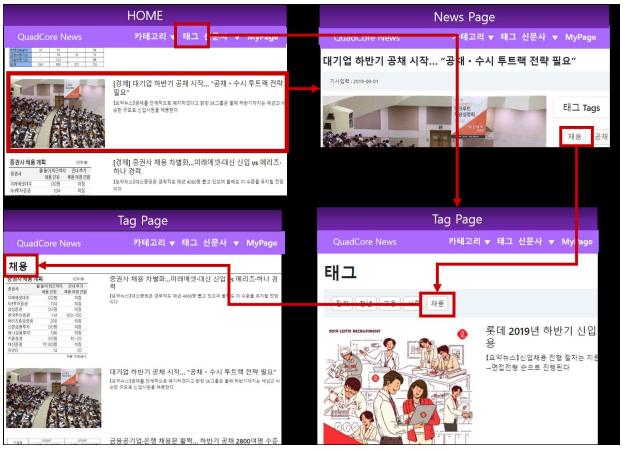

그림 9. 태그 페이지
Fig. 9. Tag Page

4-3 추천 성능 비교

제작한 시스템은 5명의 실험 참가자가 네이버에서 크롤링된 기사들을 기반으로 테스트하였으며 테스트 기간에는 추천된 기사들에 대해 1점에서 5점 사이의 평점을 매기도록 하였다. 네이버에서는 5가지 카테고리별 기사 15,000건 정도가 크롤링 되었으며 본 논문에서 제안한 태그 선정 방식대로 상위 태그만을 취하여 425개의 태그에 대해 학습하였다. 추천 성능을 확인하기 위해 본 논문에서는 사용자 테스트에 대한 평균 제곱근 오차(RMSE, Root Mean Square Error)를 평가 지표로 사용하였다[16]. 평균 제곱근 오차는 사용자-아이템 쌍에 대한 예측 평점(rating)과 실제 평점의 차이를 통해 계산하며 추천시스템 평가에 있어 보편적으로 사용되는 평가 방법이다. RMSE의 오차는 예측값과 실제 값 사이의 차이를 뜻하며, 차이가 작을수록 정확한 예측 시스템이고 시스템의 성능이 좋다는 것을 의미한다. 수식 3은 RMSE 값을 계산하는 식을 나타낸 것이다. 여기서 n은 사용자 수고, x_i는 추천시스템의 예측 평점이며 $\hat{x}_i$는 사용자가 매긴 실제 평점을 가리킨다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \qquad (3)$$

표 1. RMSE 값 비교
Table 1. Comparison of RMSE Value

	Text	Text&Image
RMSE	0.1743	0.1344

표 1에서는 TF-IDF 기반의 텍스트 태그만을 사용했을 경우와 텍스트 태그와 이미지 태그를 함께 사용했을 경우 2가지의 RMSE 값을 비교한 것이다. 먼저 텍스트 태그만 사용했을 경우 RMSE 값은 0.1743이었으며 텍스트 태그와 이미지 태그를 함께 사용했을 경우의 RMSE 값은 0.1344로 이미지 태그를 함께 사용했을 때 추천 성능이 향상된 것을 확인할 수 있다.

Ⅴ. 결 론

본 논문에서는 뉴스 기사의 본문에서 추출한 TF-IDF 태그뿐만 아니라 연관검색어 태그와 기사 이미지 데이터에서 추출한 태그도 함께 고려하는 뉴스 기사 추천서비스인 'QuadCore News' 시스템을 개발하였다. 텍스트뿐만이 아닌 기사의 사진에서도 태그를 추출함으로써 한국어 형태소 분류 시 발생하는 오류로 인한 문제(지칭 대명사, 단위 명사가 상위 단어로 계산되는 문제가 발생함)를 완화할 수 있다. TF-IDF 태그뿐만 아니라 추가로 연관검색어를 활용함으로써 기사와 관련 있는 양질의 데이터 집합을 늘려 기사 본문에는 없지만, 관련성이 높은 데이터를 태그로 제공하였다.

넘쳐나는 뉴스 기사들 속에서 자신이 읽고 싶은 기사를 찾는 것은 소모적이고 피곤한 일이 될 수 있다. 제안하는 시스템은 사용자의 사용 기록을 활용해 관심 가질 만한 뉴스 기사를 추천함으로써 현대인들이 더 편리하고 간편하게 뉴스를 소비할 수 있도록 돕는다. 해당 시스템은 현재 뉴스 기사 도메인에 대한 추천시스템을 제작하였지만, 기사가 아닌 다른 텍스트 기반의 추천시스템에도 적용될 수 있을 것으로 기대된다. 또한, 본 논문에서 제안한 학습 방법은 기사 외에도 SNS와 같은 텍스트와 이미지를 동시에 갖는 도메인에 확장될 수 있다.

감사의 글

본 연구는 2019년도 덕성여자대학교 교내연구비 지원에 의해 수행되었다.

참고문헌

[1] M.D. Ekstrand, "Collaborative Filtering Recommender Systems", *Foundations and Trends in Human-Computer Interaction,* Vol. 4, No. 2, pp. 81-173, 2011.

[2] T. Yoneda, S. Kozawa, K. Osone, Y. Koide, Y. Abe and Y. Seki, “Algorithms and System Architecture for Immediate Personalized News Recommendations”, *eprint arXiv:1909.01005*, 2019.

[3] O. Ozgobek, J. A. Gulla, and R. C. Erdur. “A survey on challenges and methods in news recommendation”, *in Proceedings of the 10th International Conference on Web Information System and Technologies* , 2014.

[4] F. Garcin, K.Zhou, B. Faltings, V. Zchickel, “Personalized News Recommendation Based on Collaborative Filtering”, *in Proceedings of the IEEE/WIC/ACM WI-IAT*. pp 437-442, 2012.

[5] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl, “Grouplens: applying collaborative filtering to usenet news,” *Commun. ACM*, vol. 40, pp. 77–87, March 1997.

[6] Charu Aggarwal, Recommender Systems, Springer, 2016.

[7] Open Source Software, Collaborative Filtering – Core Technology of Recommendation System. Available: https://www.oss.kr/info_techtip/show/5419f4f9-12a1-4866-a713-6c07fd36e647

[8] A. Das, M. Datar, A. Garg, S. Rajaram, "Google News Personalization: Scalable Online Collaborative Filtering" *in Proceedings of the 16th International Conference on World Wide Web*, Banff, Alberta, Canada, pp 271-280, May 2007.

[9] J. Liu, P. Dolan, E. R. Pedersen, "Personalized News Recommendation Based on Click Behavior", *in Proceeding of the 15th International Conference on Intelligent User Interfaces*, Hong Kong, pp 31-40, February 2010.

[10] K. Park, J. Lee., J. Cho, "Deep Neural Networks for News Recommendations", *in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management,* pp 2255-2258, 2017.

[11] Y. Ren and S. Gong, "A collaborative filtering recommendation algorithm based on svd smoothing", *in Proceedings of the 2009 Third International Symposium on Intelligent Information Technology Application,* Volume 02, ser. IITA'09, pp. 530-532, 2009.

[12] S. Lee, H. Kim, "News Keyword Extraction for Topic Tracking", *in Proceeding of the 2008 Fourth International Conference on Networked Computing and Advanced Information Management,* Gyeongju, South Korea, pp 554-559, September 2008.

[13] E.L. Park, S. Cho, "KoNLPy: Korean natural language processing in Python", *in Proceeding of the 26th Annual Conference on Human & Cognitive Language Technology*, ChunCheon, South Korea, 2014.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the Inception Architecture for Computer Vision", *in Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp 2818-2826, 2016.

[15] J. Seol and S. Lee, “lexrankr: LexRank based Korean multi-document summarization”, *in Proceedings of the Korea Software Congress(KSC2016)*, pp458-460, 2016.

[16] G. Shani, A. Gunawardana, Evaluating Recommendation Systems. Recommender Systems Handbook, Springer, pp 257-297, 2010.

정인정(In-Jeong Jeong)

2016 - 현 재: 덕성여자대학교 컴퓨터공학과 재학
※관심분야 : 데이터 분석, 웹 시스템

김보미(Bo-Mi Kim)

2016 - 현 재: 덕성여자대학교 컴퓨터공학과 재학
※관심분야 : 웹 서비스, 빅데이터, 텍스트마이닝

김수경(Su-Kyung Kim)

2016 - 현 재: 덕성여자대학교 컴퓨터공학과 재학
※관심분야 : 인공지능, 추천시스템, 자연어처리

유견아(Kyeonah Yu)

1986년: 서울대학교 제어계측공학과 공학사.
1988년: 서울대학교 제어계측공학과 공학석사.
1995년: University of Southern California
컴퓨터학과 공학박사

1996년～현 재: 덕성여자대학교 컴퓨터학과 교수
※관심분야 : 인공지능, 지식기반 시스템과 딥러닝