검색환경 개선을 위한 자연어 처리 기반 맞춤형 추천 검색시스템

**승현수, *박지윤, *우다현, *오승민
*아주대학교 소프트웨어학과
**아주대학교 전자공학과

todd06@ajou.ac.kr, jiyoon0043@ajou.ac.kr,
wdh112139@ajou.ac.kr, mangusn1@ajou.ac.kr

Recommender system for web search based on NLP to improve user search environment

Hyeon-Su Seung**, Ji-Yun Park*, Da-Hyun Woo*, Seung-Min Oh*
*Dept. of Software and Computer Engineering, Ajou University
**Dept. of Electronic Engineering and Computer Science, Ajou University

요 약

일반적인 검색엔진을 가진 포털 환경에서 정보검색 시 사용자가 원치 않는 수많은 검색결과가 동반되기도 하고 자신의 취향에 맞는 글을 검색하지 않았다는 이유만으로 원하는 정보를 놓치는 상황도 일어난다. 이러한 검색환경의 문제를 개선하기 위해 본 논문에서는 사용자들의 검색환경 개선을 위한 맞춤형 검색결과 정렬, 검색어 추천, 게시글 추천의 추천 시스템을 설계하고 제작한다. 이러한 추천 시스템은 워드 임베딩 모델과 추천 시스템 모델을 포함한다. 기존에 존재하던 워드 임베딩 모델의 성능을 실험을 통해 비교 및 분석하고, 크롤링을 통해 모은 데이터로 성능을 24.98%P 개선하였다. 추천 시스템 모델은 RMSE 비교를 통해 최적이 알고리즘을 제안한다. 해당 기술을 통해 사용자 스스로 자신의 검색환경을 개선할 수 있도록 구현하는 것이 이 시스템의 목표이다.

1. 서론

원하는 정보를 찾기 위한 인터넷 검색은 현대 사회에서 나이와 직업을 막론하고 불가피하다. 하지만 인터넷의 방대한 자료 속에서 자신이 원하는 정보만 골라내기는 어려운 일이다. 사용자는 원하지 않는 글을 접하거나, 무엇을 검색해야 할지 몰라 검색에 시간을 낭비하거나, 정확한 검색어를 검색하지 않았다는 이유로 원하는 글을 놓치기도 한다.

이러한 문제를 해결하기 위해서 사용자 맞춤형 추천 검색시스템이 필요하다. 우리는 NLP 기술과 추천 기술을 활용해서 검색결과를 사용자의 취향에 맞게 재 정렬하고, 맞춤 검색어를 추천해주며, 맞춤 게시글 또한 추천해주는 추천 검색시스템을 제안한다.

현재 여러 포털사이트의 뉴스, 카페 글, 블로그 글을 모아서 사용자가 평소에 관심 있는 분야의 글을 추천해주는 플랫폼은 존재하지 않는다는 점에서 이러한 추천 검색시스템의 필요성은 더욱 대두된다.

본 논문에서 제안하는 추천 시스템을 통해 사용자가 스스로 자신의 검색환경을 개선해나가는 것을 기대해볼 수 있다.

2. 관련 연구

2.1. NLP

NLP(자연어 처리)는 텍스트에서 의미 있는 정보를 분석, 추출하고 이해하는 일련의 기술 집합이다. 사용자에게 검색어와 게시글을 추천해주기 위하여 검색한 검색어와 게시글의 특징을 찾아낼 수 있어야 한다. 텍스트에서 특징을 추출하기 위해 워드 임베딩 모델이 필요한데 이 모델은 설계에 따라 종류가 다양하다. 본 논문에서는 4가지 설계 방식의 워드 임베딩 모델을 소개하고 비교해보며 콘텐츠 카테고리화와 추천 시스템에 알맞은 모델을 확인했다.

<표 2> 4가지 임베딩 모델

Model	Architecture
Baseline	LSTM
Pororo[1)]	zero-shot learning
FastText[2)]	n-gram
KoBERT[3)]	BERT

1) https://github.com/kakaobrain/pororo
2) https://github.com/facebookresearch/fastText

최근에는 신조어가 많이 등장하기 때문에 단어 사전에 없는 단어도 임베딩 값을 찾을 수 있도록 zero-shot learning[1] 기반의 모델과 일반적으로 많이 쓰이는 n-gram[2] 기반의 모델, 그리고 최근 NLP에서 가장 핫한 BERT[3] 모델의 성능을 비교했다.

Zero-shot learning 기반의 모델은 Kakao brain에서 제작한 Pororo 라이브러리의 zero-shot classification을 사용했고, n-gram 모델은 Facebook에서 제작한 FastText 라이브러리의 text classification을 활용했다. BERT모델은 한국어로 추가 학습한 SKTBrain의 KoBERT를 이용하고, Baseline모델로 LSTM기반의 가벼운 임베딩 모델을 설계했다. 이 임베딩 모델을 기반으로 콘텐츠 카테고리화를 구현한다.

2.2. Ranking System

사용자가 선택한 선호 카테고리(C), 스크랩 이력(Sc), 검색 이력(S), 좋아요 이력(H) 등을 바탕으로 미리 지정해둔 32개의 카테고리에 대해 사용자의 선호 rating(R)을 식(1)의 형태로 계산한다.[5] 이후, 검색결과의 제목으로 콘텐츠 카테고리 분류 모델을 이용하여 각 카테고리에 대한 점수(P)와 검색결과의 카테고리와 사용자의 카테고리별 rating(R)을 통해 예상 선호 순위(Expected Ranking)가 높은 순으로 검색결과를 정렬하여 사용자 취향에 맞게 보여준다. ER을 통해 검색결과 추천을 구현한다.

$$R = 5C + 0.01Sc + 0.005S + 0.001H \quad \cdots (1)$$

$$ER = R \times P \quad \cdots (2)$$

2.3. Item Based Collaborative Filtering

사용자에게 적절한 검색어와 게시글을 추천해줄 때, 사용자의 취향을 반영하여 위하여 Item Based Collaborative Filtering[4] 기법을 적용한다. 이 기법은 사용자의 콘텐츠에 대한 선호도를 바탕으로 사용자가 관심 있어 할 만한 콘텐츠를 추천해주는 방법이다.

검색어와 게시글 추천은 카테고리와 사용자의 아이템 선호도에 따른 Item Based Collaborative Filtering의 결과를 최대 평가 예측값으로 계산하여 사용자에게 추천한다. Collaborative Filtering의 과정은 3가지의 과정으로 이루어진다. 첫 번째로, 평가치 matrix를 <표 2>와 같이 구성한다.

<표 2> User-Item Matrix

User \ Item	Item1	Item2	Item3
User A	$R_{A,1}$	ϕ	ϕ
User B	$R_{B,1}$	$R_{B,2}$	ϕ
User C	ϕ	ϕ	$R_{C,3}$

두 번째로, 사용자에 대한 아이템의 평가치를 아이템의 차원으로 벡터화 시킨 후 식(3)을 거쳐서 유사도에 따른 최 근접 이웃을 구성한다.

$$\cos(\vec{A},\vec{B}) = \frac{\vec{A} \bullet \vec{B}}{|A|^2 \times |\vec{B}|^2} \quad \cdots (3)$$

마지막으로 식(4)를 통해 최근접 이웃 간의 가중평균 후 Top-N 기법으로 추천목록을 생성한다. 생성된 추천목록은 검색어 추천과 게시글 추천구현을 위해 사용된다.

$$R_{A,i} = \overline{R_A} + \frac{\sum_{j=1}^{k} w(A,j)(R_{j,i} - \overline{R_j})}{\sum_{j=1}^{k} |w(A,J)|} \quad \cdots (4)$$

3. 구현

3.1. 구현 환경

학습을 시키는데 필요한 데이터는 데이터 크롤링으로 직접 수집하였으며, 32개의 카테고리에 대해 카테고리별로 약 2000개의 데이터를 모아 약 65000개의 데이터셋을 구축했다. 학습은 train set, test set으로 나누어 각각 61,800개, 3200개로 학습 및 평가한다.

콘텐츠 카테고리화 모델은 워드 임베딩 모델에 fully connected layer를 추가하여 만든 모델이다. 워드 임베딩 모델의 성능을 실험해보기 위해 콘텐츠 카테고리화 모델을 기준으로 성능을 측정했다. 워드 임베딩 모델로 사용한 모델은 총 4가지로 LSTM으로 구성된 Baseline, Pororo, FastText, KoBERT 이다. KoBERT는 pretrained 모델을 사용하였고 나머지는 scratch로 진행했다. 모델별 word embedding 및 fully connected layer를 통하여 콘텐츠 카테고리화의 정확도를 평가했다. 모델에 사용한 파라미터는 <표 3>과 같다.

3) https://github.com/SKTBrain/KoBERT

<표 3> 모델 파라미터

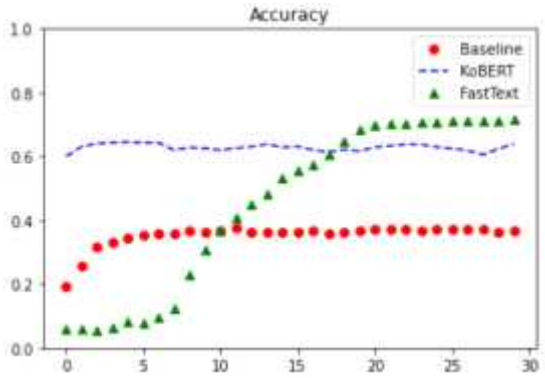
Parameter	
batch size	128
drop out rate	0.3
epochs	30
bidirectional	True
shuffle	True

3.2. 성능 측정

3.2.1. 콘텐츠 카테고리화

콘텐츠 카테고리화 모델을 훈련한 결과는 (그림 1)과 같다. Pororo 모델은 학습할 수 있는 함수를 라이브러리에서 제공하지 않아서 학습에 포함하지 않았다. 10 epoch당 정확도는 <표 4>와 같고, 모델별 속도 비교는 <표 5>에서 알 수 있다.

실험 결과 FastText가 0.7151의 정확도를 보이며 가장 높았고 속도 역시 비교 모델 중 가장 빨랐다. 따라서 FastText로 categorizing 및 유사단어 추출을 진행했다.



(그림 1) 모델별 정확도

<표 4> 모델별 epoch에 따른 정확도 비교

모델	Epoch에 따른 정확도(%)		
	10 epoch	20 epoch	30 epoch
Baseline	36.33	36.63	36.49
FastText	30.74	**64.83**	**71.51**
KoBERT	**62.91**	62.86	68.37

<표 5> 모델별 속도와 정확도

모델	속도(초)	정확도(%)
Baseline	2.321	36.49
FastText	**0.183**	**71.51**
KoBERT	9.832	68.37
Pororo	1920.826	12.625

3.2.2. 검색결과 추천

검색결과 추천은 정확도를 측정할 방법이나 데이터셋이 명확하지 않아서 사용자별 관심 분야를 다르게 설정하고 검색결과가 어떻게 나오는지 직접 확인했다. 실험 결과는 <표 6>과 같으며, 같은 검색어를 입력하더라도 관심 분야에 따라 알맞은 검색결과가 나타나는 것을 볼 수 있다.

<표 6> 관심사에 따른 검색어 '레드벨벳'의 검색결과

관심 분야	검색 결과 (Top-3)
공연/전시, 방송	1. 내 곁에 미술관 HAPPINESS 후기 레드벨벳 슬기와 함께하는... 2. 내 곁에 미술관 SLEEP 후기 레드벨벳 슬기 미술책 추천해요 3. 온앤오프-레드벨벳-로꼬... MZ세대 겨냥한 DJ들의...
스타/연예인	1. 연예인 평판 1위는 방탄소년단... 2위 임영웅 3위 블랙핑크 2. [단독] 레드벨벳 조이♥크러쉬, 핑크빛 열애... '고막커플... 3. 이무진의 '신호등', 멜론 주간 국내종합 1위... 아이유 8곡, 임영웅 곡, 방탄...
요리/레시피, 상품리뷰	1. 대전 관저동 :: 윤달/컵케익(레드벨벳, 딸기) 2. [광주 운암동] 레드벨벳 컵케이크 맛집 아담한 카페 버터플레이버 3. 큐브, 레드벨벳 크림 치즈 케이크, 부드러운 생크림 카스텔라...
방송, 음악	1. MBC 쇼! 음악중심 / 레드벨벳 컴백 무대 2. 레드벨벳 서머송으로 완전체 컴백! 기다렸노라! 3. 레드벨벳 퀸덤 컴백 기념 스밍 무나
사진, 패션/미용	1. 레드벨벳 예리 인스타그램 여성 원피스 코디, 연예인 옷... 2. 레드벨벳 웬디 귀걸이 추천 에끌라스튜디오로 여자 가을 코디에... 3. 레드벨벳 퀸덤 슬기 숄더백 가방 패션 @마지셔우드

3.2.3. 검색어, 게시글 추천

검색어 추천과 게시글 추천은 같은 추천 시스템을 사용했고 이를 평가하기 위해서 Movie Lens

Dataset을 이용하여 RMSE[4] 값을 추출했다. Movie Lens Data의 Movie ID를 ratings로 변환하고 사용자와 아이템 기준으로 matrix를 만들어 User Based Collaborative Filtering[6], Item Based Collaborative Filtering 기법을 식(5)를 통해 평가했다. <표 7>에서 Item Based Collaborative Filtering이 RMSE값이 더 낮은 것을 확인했고, 이를 사용한 추천 시스템을 개발했다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \cdots (5)$$

<표 7> RMSE 평가 결과

Type	RMSE
Item Based	0.81415
User Based	1.69484

<표 8> 사용자별 선호 카테고리의 일부

User Idx	Category Idx	Rating
1	20	2
2	35	1
3	2	4

Item Based Collaborative Filtering을 활용해 <표 8>과 같이 임의로 18명의 선호 카테고리를 설정하여 이를 기반으로 희소 행렬을 추출하고 결측치를 0으로 설정했다. 카테고리를 기준으로 정렬된 행렬을 코사인 유사도 행렬로 분리하여 아래와 같이 타 카테고리에 대한 예측값이 있는 행렬 <표 9>를 계산한다. 그리고 이를 통하여 사용자가 선택하지 않은 항목의 아이템에 대한 선호도가 나오는 것을 확인할 수 있다.

<표 9> 사용자별 선호 카테고리 예측 일부

User Idx	6	7	8	9	10	11	12	13
1	0	1.5	0	0	0	16.	1.5	1.8
2	1.0	0	0	1.6	0	0	0	0
3	0	0.5	0	0.6	0.6	0	0	0

4. 결론 및 향후 연구 계획

사용자가 검색에 많은 시간을 소비하고 원하는 글을 쉽게 찾아볼 수 없는 검색환경 문제의 개선을 위해서 검색결과 정렬, 검색어 추천, 게시글 추천 시스템을 설계 및 개발했다. 이러한 추천 시스템은 워드 임베딩 모델과 사용자 추천 시스템 모델을 포함한다.

본 논문에서는 FastText의 n-gram 모델과 크롤링 데이터를 통해 학습을 진행했고, 워드 임베딩 모델은 FastText에서 제공해주던 pretrained 모델의 정확도 46.53%보다 24.98%P 개선한 71.51%를 달성했다. 추천 시스템 모델은 RMSE 값을 비교하여 User Based Collaborative System보다 Item Based Collaborative System이 본 논문에서 지적한 문제를 해결하는데 더 적합한 것으로 판단했다.

본 연구를 통해 기존 모델을 개선하고 직접 모델을 설계하고 제작해보면서 다양한 NLP 기술과 추천 시스템 기술에 대해 알아볼 수 있었다.

또한, 개발한 모델을 사용자가 쉽게 사용할 수 있도록 실제 애플리케이션에 적용했다. 사용자는 맞춤형 서비스를 통해 검색에 필요한 노력과 시간을 단축할 수 있을 것이며 웹서핑을 더욱 수월하게 할 수 있을 것이다. 더하여, 애플리케이션을 출시하고 사용자 데이터가 더 많아진다면 성능 향상을 위한 연구를 추가로 진행할 예정이다.

감사의 글

본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물이다.

참고문헌

[1] Richard Socher, et al, “Zero-shot Learning Through Cross-Modal Transfer”, In *CVPR*, (2013)

[2] Armand Joulin, etal. “Bag of Tricks for Efficient Text Classification”, In arXiv, (2016)

[3] Vaswani, et al. "Attention is all you need." Advances, In NIPS, (2017)

[4] Sarwar, et al. "Item-based collaborative filtering recommendation algorithms." In WWW, (2001)

[5] 이재식, 박석두 , “장르별 협업필터링을 이용한 영화*추천 시스템의 성능 향상”, 한국지능정보시스템학회논문지, 제13권 제4호 67~68p

[6] Robin, et al. "Using content based filtering for recommendation." In MLnet/ECML2000, (2000)

4) Root Mean Square Error: 평균 제곱근 오차