The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.5, pp.221~226
eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.5.221

Search Re-ranking Through Weighted Deep Learning Model

Gi-Taek An† · Woo-Seok Choi†† · Jun-Yong Park†† · Jung-Min Park††† · Kyung-Soon Lee††††

ABSTRACT

In information retrieval, queries come in various types, ranging from abstract queries to those containing specific keywords, making it a challenging task to accurately produce results according to user demands. Additionally, search systems must handle queries encompassing various elements such as typos, multilingualism, and codes. Reranking is performed through training suitable documents for queries using DeBERTa, a deep learning model that has shown high performance in recent research. To evaluate the effectiveness of the proposed method, experiments were conducted using the test collection of the Product Search Track at the TREC 2023 international information retrieval evaluation competition. In the comparison of NDCG performance measurements regarding the experimental results, the proposed method showed a 10.48% improvement over BM25, a basic information retrieval model, in terms of search through query error handling, provisional relevance feedback-based product title-based query expansion, and reranking according to query types, achieving a score of 0.7810.

Keywords : Information Retrieval, Deep Learning Model, DeBERTa, Product Search

검색 재순위화를 위한 가중치 반영 딥러닝 학습 모델

안 기 택† · 최 우 석†† · 박 준 용†† · 박 정 민††† · 이 경 순††††

요 약

정보검색에서 질의는 다양한 유형이 존재한다. 추상적인 질의부터 구체적인 키워드를 포함하는 질의까지 다양한 형태로 구성되어 있어서 사용자의 요구에 정확한 결과 도출은 어려운 과제이다. 또한 검색시스템이 오타, 다국어, 코드와 같은 다양한 요소를 포함하는 질의를 다뤄야 하는 특징이 존재한다. 본 연구에서는 질의 유형을 분석하고, 이에 따라 딥러닝 기반 재순위화의 적용 여부를 결정하는 방법을 제안한다. 최근 연구에서 높은 성능을 보인 딥러닝 모델인 DeBERTa를 이용하여 질의에 대한 적합 문서의 학습을 통해 재순위화를 수행한다. 제안 방법의 유효성을 평가하기 위해 국제정보검색 평가대회인 TREC 2023의 상품 검색 트랙(Product Search Track) 테스트컬렉션을 이용하여 실험을 하였다. 실험 결과에 대한 정규화된 할인누적이득(NDCG) 성능측정 비교에서 제안 방법이 정보검색 기본 모델인 BM25 에 비해 질의 오류 처리를 통한 검색, 잠정적 적합성피드백을 통한 상품제목 기반 질의확장과 질의유형에 따른 재순위화에서 0.7810으로 BM25 대비 10.48% 향상을 보였다.

키워드 : 정보검색, 딥러닝모델, DeBERTa, 상품검색

1. 서 론

온라인 쇼핑은 등장 이후 다양한 요인과 기술의 발전으로 계속 성장하고 있다. 온라인을 통해 음식을 배달하고, 중고 물품을 거래하고, TV홈쇼핑은 라이브커머스로 온라인 플랫폼화 되었다. 대형 온라인 쇼핑몰에서는 이미 오프라인의 매장에서 판매하는 거의 모든 물건을 구매할 수 있다. 소비자는 이렇게 많은 물건 중에서 내가 원하는 물건을 찾아야 하고 판매자는 본인의 물건을 소비자에게 노출시켜야 판매할 수 있다. 정보검색 기술은 이러한 문제를 해결하기 위해 사용된다. 키워드 중심의 검색 방법에서 최근에는 딥러닝을 적용한 검색 방법이 연구되고 있고 좋은 성능을 보여준다.

최근 정보검색 분야의 주요 국제 평가 대회를 보면 온라인 상품 검색의 중요성과 필요성을 알 수 있다. 2022년에는 KDD Cup 2022 챌린지에서 상품 데이터를 이용한 정보검색 대회가 있었고, 2023년에는 TREC(Text Retrieval Conference)에 상품검색(Product Search) 태스크가 추가되었다. 두 대회 모두

※ 이 논문은 2024년도 과학기술정보통신부 재원으로 한국식품연구원의 지원(기본사업 E0220700)을 받아 수행된 연구성과입니다.
※ 본 연구는 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 지원을 받아 수행되었음(2022-0-01067).
† 정 회 원 : 전북대학교 컴퓨터공학전공 박사과정
†† 비 회 원 : 전북대학교 컴퓨터인공지능학부 학사과정
††† 비 회 원 : 한국식품연구원 지능화정책팀장(책임연구원)
†††† 비 회 원 : 전북대학교 컴퓨터인공지능학부 교수
Manuscript Received : March 12, 2024
Accepted : April 11, 2024
* Corresponding Author : Jung-Min Park(parkjm@kfri.re.kr), Kyung-Soon Lee(selfsolee@jbnu.ac.kr)

온라인 쇼핑몰인 아마존의 상품 문서 데이터와 상품을 찾는 질의 데이터를 제공하여 질의에 따른 순위를 적용하거나 문서와 질의의 적합을 분류하는 것을 목표로 하고 있다.

상품검색 평가대회의 질의를 분석해보면 “$10 Candles”, “$5 items”와 같은 추상적인 질의, “boys purple under armour shirt”와 같은 구체적인 상품 정보를 포함하여 질의하는 유형으로 구분할 수 있다. 질의에 나타나는 현상으로는 일본어, 스페인어 등 다양한 언어를 사용하여 검색하는 다국어(multi-lingual) 질의, “battery operated outsoor lamps”와 같이 질의 입력에서 발생하는 오타가 있는 경우(‘outdoor‘), “0307339459”와 같이 상품의 코드를 검색하는 질의 등 검색성능에 영향을 미치는 부분들을 다루어야 한다. 이와 같이 상품검색에서 나타나는 다양한 질의 특성은 일반정보검색에 비해 적합한 결과를 검색하는데 있어서 성능 저하의 요인이 된다.

본 연구에서는 다양한 특징이 있는 상품 검색을 위해 질의 분석 결과에 따른 유형을 검색 단계별로 분리하여 적용하여 성능을 향상하는 방법을 제안한다. 최근 정보검색에서는 검색, 재순위화 두 단계를 활용한 모델을 많이 활용하고 있다. 딥러닝을 이용한 재순위화는 좋은 성능을 보여주고 있지만 딥러닝을 사용하기 위해서는 GPU와 같은 많은 연산자원이 필요하여 모든 문서를 대상으로 재순위화를 적용하기에는 무리가 있다. 따라서 재순위화 대상을 축소해야 할 필요가 있고 기존의 검색 단계를 통한 상위 검색 결과를 대상으로 딥러닝 단계를 통해 재순위화하는 두단계 모델이 활용되는 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개한다. 3장에서는 본 연구에서 질의 분석을 통한 검색 단계에서의 성능 하락 요인 개선 방법을 제안하고, 4장에서는 질의 유형에 따른 특성을 반영한 딥러닝 기반의 재순위화 방법을 제안한다. 5장에서는 실험 결과를 비교 분석하고 6장에서는 결론을 맺는다.

2. 관련 연구

2.1 두 단계 검색 아키텍처

최근 많은 정보검색 관련 연구에서는 다단계 아키텍처(Multi-stage architecture)를 사용하고 있다[1-4]. 기존의 단일 단계 검색에서는 키워드 기반 검색, 문서 유사도 기반 검색 등이 있었다. 다단계 아키텍처 연구에서는 주로 후보 검색 단계와 재순위화 단계로 이루어진 두 단계 파이프라인을 사용한다. 첫 번째 단계에서는 검색 모듈을 사용하여 대규모 풀에서 관련된 문서를 검색한다. 이때, TF-IDF[5], BM25[6]와 같은 정보검색 기본 모델을 이용한다. 두 번째 단계에서는 주로 밀집 검색(Dense Retrieval)을 이용한 딥러닝 방법을 사용한다. 이때, 딥러닝 모델을 이용한 검색 또는 재순위화를 하는 딥러닝 모델을 적용하여 순차적으로 정렬한다. 딥러닝 기반 방식의 밀집 검색은 기존의 검색 모델인 어휘 기반 방식 보다 좋은 성능을 보여주고 있다[7-10]. 하지만 딥러닝 기반 방식은 많은 연산시간과 비용이 필요하므로 두 가지 방식을 병렬 또는 독립적으로 구성한 두 단계 검색 아키텍처가 표준 관행이 되었다[11,12].

2.2 상품 검색 관련 연구

상품 검색과 관련한 최근 연구로 KDD Cup 2022 챌린지와 TREC 2023에서 아마존 상품 검색 데이터를 이용한 평가대회가 있다. 이 평가대회들에서 딥러닝 기반 검색 기법을 적용한 연구팀의 성능이 상위 순위를 나타냈다. 다수의 연구에서 BERT를 기반으로 하는 DeBERTa, RemBERT, RoBERTa 등의 모델을 사용하였고[13,14] 그 외에도 infoXLM 등을 사용한 연구[15]가 있다. 본 연구에서는 기존연구[16, 17]를 확장하여 실험을 수행하였다.

3. 상품 질의 분석을 통한 개선된 검색 모듈

본 연구에서는 상품 검색에서 성능저하 요인이 되는 질의를 분석하여 처리함으로써 기본 검색 성능을 높이고, 재순위화 단계에서 구체적인 질의 유형에 대한 딥러닝 기반 학습 개선을 통해 검색 성능을 향상시키는 방법을 제안한다. 기본 검색 단계에서는 확률 기반 정보검색 모델인 BM25를 이용하였고, 재순위화단계에서는 기존 연구에서 좋은 성능을 보여주었던 DeBERTa 모델을 적용하였다. 모든 문서를 대상으로 딥러닝을 적용하여 검색하는 것은 연산에 많은 자원과 시간을 요구된다. 시간과 연산량을 줄이기 위하여 첫 단계 검색에서는 엘라스틱서치(Elasticsearch) 엔진을 이용한 BM25 기법으로 1,000개의 후보 문서를 검색한다. 상품 검색 테스트컬레션의 문서에 포함된 모든 정보를 색인하고, 항목은 제목과 내용만을 구분하여 색인하였다.

3.1 성능이 낮은 질의 유형의 처리를 통한 검색 개선 모듈

본 연구에서는 TREC 2023 상품검색 테스트컬렉션의 학습 문서 집합에 대하여 질의별 BM25 검색 성능 결과를 분석하였다. 상품검색에서 성능 하락의 요인이 되는 3가지 질의 유형은 다국어 질의, 오타가 포함된 질의, 그리고 상품코드가 질의인 경우이다.

Table 1은 성능 하락의 요인이 되는 주요 3가지 질의 유형과 그에 대한 질의 예를 나타낸다. 다국어 질의에서 질의 언어가 전체 대비 11.1%가 영어가 아닌 언어로 되어 있다. 상품코드로 된 질의는 관련 문서를 검색하기가 어렵다.

상품검색 질의 유형 분석을 통해 본 연구에서는 위 성능 하락 요인에 대한 처리 방법을 다음과 같이 제안한다.

1) 영어가 아닌 다국어 질의에 대해 번역 모듈을 적용하여 질의를 영어로 변환
2) 상품 정보를 활용하여 학습한 질의 교정 모듈의 적용
3) 상품 코드 정보를 포함하는 질의를 구분하여 상품명으로 변환

Table 1. Three Types of Queries and Example that Cause Performance Degradation

non-English query	Abrigo de invierno para jovens 자전거 트레일러 ダイソン V11 Fluffy コードレス 掃除機
typo query	!awnmower tires without rims battery operated outsoor lamps
product code query	B01GHS8MA8 0307339459

본 연구에서는 상품 검색 도메인에 적합한 오타 교정을 위해 질의 교정 모듈을 개선하였다. 상품 검색에서의 질의에 발생하는 오타는 오픈 도메인 질의에서 발생하는 오타와는 다른 문제가 있다. 첫 번째는 질의가 짧고 문맥이 없이 키워드 중심의 질의라는 점과 두 번째는 제품명, 브랜드명 등 사전에 없는 단어를 질의에 사용하는 경우에 발생한다.

질의가 짧고 문맥이 없는 경우 Seq2Seq 모델들을 사용할 수 없고, 최근 좋은 성능을 보여주는 transformer[18] 기반의 모델들에서도 그러하다. 또한 공개된 사전기반의 오타 교정을 하는 경우 사전에 없는 단어를 오타로 판단하여 수정하는 경우가 발생한다. 예를 들어 “anker iphone chargeer”와 같은 문장이 교정되면서 “anger iphone charger”로 오타 이외에도 브랜드명이 변경되는 경우이다. 질의 전처리 과정에 대한 순서를 Fig. 1의 순서도로 표현하였다.

본 연구에서는 질의 교정을 위해 pyspellchecker[19]를 상품 정보에 특화하여 사용하였다. 문서 데이터셋에서 색상, 브랜드, 모델명, 소재를 구분하여 상품에 특화된 사전데이터를 구축하여 학습하였다.

원본 질의를 이용하여 오타 교정 기본모듈과 학습된 모델

Table 2. Correction Comparison of Typographical Error Correction Module

original query	default typo	our typo
airpod case **steipes**	**airport** case **stripes**	**airpod** case **stripes**
anker iphone **chargeer**	**anger** iphone **charger**	**anker** iphone **charger**

간의 차이를 Table 2와 같이 비교하였다. 첫 번째 질의인 “airpod case steipes”의 경우 오타인 “stripes”는 모두 잘 교정되었다. 하지만 제품명인 “airpod”을 기본모듈에서는 “airport”로 사전에 있는 단어로 교정하였다. 상품명을 인식하지 못하고 잘못된 교정을 한 것이다. 우리가 제안하는 학습된 오타 교정 모듈은 상품명을 인식하고 불필요한 교정을 발생하지 않는다.

검색 성능 저하에 영향을 줄 수 있는 요인인 번역과 오타 교정 등 질의 처리를 통해 딥러닝 기반 재순위화 대상이 되는 후보 문서를 검색한다.

3.2 잠정적 적합성 피드백을 이용한 질의 확장 모듈

본 연구에서는 상품 제목 정보가 질의 확장에 유효한 정보임을 확인하고, 잠정적 적합성 피드백(Psuedo Relevance Feedback) 방법을 이용하여 질의를 확장하였다.

학습 질의에 대한 PRF 성능 분석을 통해, 질의에 대한 문서 제목 필드의 검색을 하고, 그 결과 상위 4개의 상품 제목을 질의로 확장하였다. 확장된 질의를 이용하여 상품 제목, 카테고리, 상세설명 등에 대해서, 질의가 제목에 매칭된 경우, 카테고리에 매칭된 경우, 상세 설명에 포함된 경우에 가중치를 다르게 적용하여 (엘라스틱서치의 boost 값을 각각 10, 5, 1로 설정함) 1,000개의 후보 문서를 검색하였다.

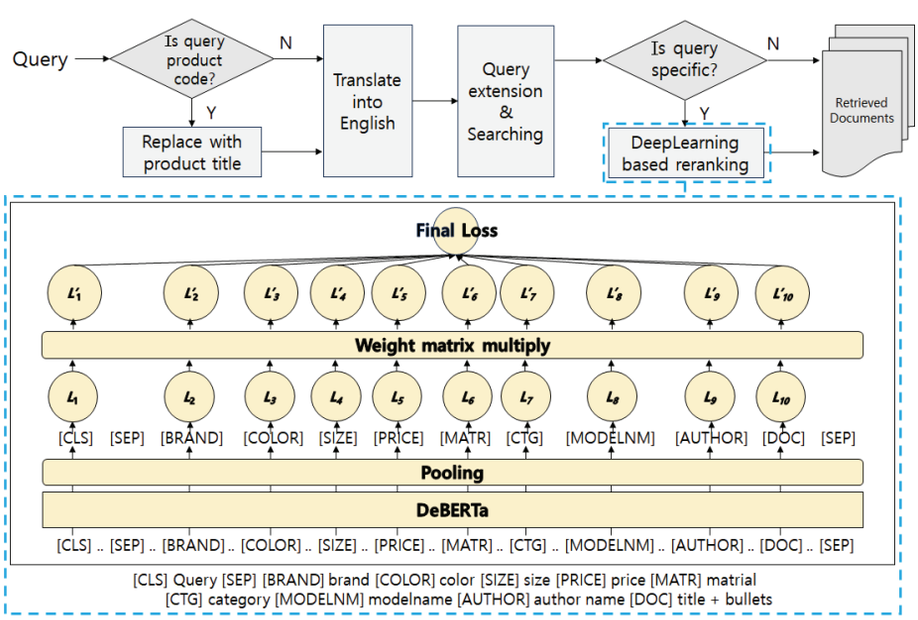


Fig. 1. Deep Learning-Based Re-ranking Model

4. 질의 유형에 따른 딥러닝 기반 재순위화

본 연구에서는 구체적인 속성이 있는 질의에 대하여 문서의 속성 정보를 활용한 DeBERTa[20] 기반의 딥러닝 재순위화 방법을 제안한다. 4장에서는 Fig. 1의 본 연구에서 제안한 두 단계 검색 기법을 설명한다. 순서도의 진행 방향에 따라 질의의 전처리가 이루어지며 딥러닝 기반 재순위화 단계에서는 학습된 모델을 이용하여 재순위화한다. 딥러닝 모델의 학습을 위해서 상품 검색 질의에 많이 나타나는 다음의 10개의 속성을 스페셜 토큰과 함께 사용하였다.

〈브랜드, 색상, 크기, 가격, 재료, 카테고리, 모델명, 저자, 제목, 세부 설명〉

질의에 스페셜 토큰이 포함될 경우 손실함수 계산에서 해당 토큰에 가중치를 줌으로써 연산 과정에서 해당 정보가 최종 손실 계산에서 크게 영향을 줄 수 있도록 하는 방법으로 상품 검색 성능에 영향을 끼치는 상품 속성을 학습에 적용하는 방법이다.

4.1 가중치 반영 DeBERTa 기반 학습 모델

질의와 문서 간의 관계를 학습하기 위하여 질의와 문서정보를 함께 딥러닝 모델의 입력으로 사용한다. 일반적으로 질의와 문서정보를 [SEP] 토큰을 이용하여 연결하고 [CLS]토큰을 활용하여 분류 모델을 구성하게 된다.

학습모델의 입력 형식은 다음과 같다.

“[CLS] query [SEP] [BRAND] 브랜드 [COLOR] 색상 [SIZE] 크기 [PRICE] 가격 [MATR] 재료 [CTG] 카테고리 [MODELNM] 모델명 [AUTHOR] 저자 [DOC] 제목+세부설명 [SEP]”

BERT 모델에서 분류를 위해 사용되는 [CLS] 토큰과 [SEP] 토큰을 제외한 다른 토큰들은 스페셜 토큰으로 추가하였다. 입력 형식에서 스페셜 토큰을 제목과 세부설명 보다 앞에 배치하는 이유는 토큰의 길이 제한이 있어 입력의 내용이 제한되는 경우를 대비하여 앞으로 배치하였다.

입력데이터를 활용하여 딥러닝 모델을 학습한다. 학습에서는 각각의 스페셜 토큰의 출력값을 활용하여 손실함수를 거쳐 계산한다. [CLS], [BRAND], [COLOR], [MATR], [AUTHOR] 토큰의 경우 손실을 계산할 때 다른 토큰보다 더 크게 반영되도록 한다.

본 연구의 실험을 통해 최적의 반영 배수를 찾았으며 [CLS] 토큰의 경우 손실의 8배, 다른 4개의 토큰은 4배를 반영하는 경우가 가장 성능이 좋았다. 손실함수는 교차엔트로피(cross entropy)를 사용하였고, 최종 손실값은 손실값에 전체 가중치의 합의 평균으로 계산했다.

딥러닝 학습 모델은 상품검색 테스트컬렉션의 질의에 대한 문서의 적합도를 4단계로 표현한 E/S/C/I 카테고리로 모델을 학습한다. 이때, E는 질의에 아주 적합한(Exact) 문서를 나타내고, S는 대체할 수 있는(Substitute) 문서, C는 보완적인(Complement) 문서, I는 관련이 없는(Irrelevant) 문서를 나타낸다.

4.2 학습모델을 활용한 검색 결과 재순위화

기본 검색 단계를 통해 생성된 후보 문서들에 대해 본 연구에서의 가중치를 반영해 학습된 딥러닝 모델을 이용하여 문서를 재순위화한다. 재순위화 단계에서의 질의와 문서 표현은 학습에 사용한 것과 같은 형태로 구성하여 입력으로 사용한다.

$$Score = E\times 100 + S\times 10 + C\times 1 + I\times 0 \qquad (1)$$

각 문서에서 대해 DeBERTa 모델을 적용하여 분류한 결과는 E/S/C/I 카테고리의 확률값이다. 이 값을 수식 1에 적용하여 재순위화를 위한 각 문서의 최종 점수를 계산한다. 분류의 정확도 반영을 하기 위해 E에는 100, S에는 10, C에는 1의 가중치를 적용하여 합산하였다.

5. 실험 및 분석

5.1 실험 데이터

Table 3은 데이터셋의 질의 데이터 개수(총 30,734개)와 문서 데이터(총 1,118,658개)의 질의별 E/S/C/I의 평균 개수이다. 정보검색 분야에서 검색 결과의 평가를 위해 질의와 문서 간의 관계를 ESCI로 표현한다. 하나의 질의에서 검색된 문서에는 E/S/C/I로 표기하여 검색성능을 평가하는 데 사용된다. 상품 검색 트랙의 데이터셋의 질의를 분석하여 질의 유형을 구분하였다.

실험에는 학습을 위해 30,734개의 질의와 1,118,658개의 문서 데이터를 사용하였고, 성능 평가를 위해서 TREC 2023에서 공개한 시험 데이터셋 926개의 질의를 사용하였다. 딥러닝 하이퍼파라미터는 학습률 1e-5, 배치 사이즈 7, 토큰 길이 512, Cross Entropy Loss와 Adam을 사용하였다.

5.2 비교 실험 방법

본 연구에서는 질의 분석과 재순위화 방법의 효과를 확인하기 위하여 단계별로 적용하여 실험하였다.

1) BM25 : 문서 데이터셋 색인 후 키워드 기반 검색
2) 질의 교정(Preprocessing): 오타 교정 모듈을 적용

Table 3. TREC203 Product Search Track Test Colleciton

	# of query	Average number of documents judged as E/S/C/I answer categories			
		Exact	Substitute	Complement	Irrelevant
Train	20,888	8.09	6.6	0.84	3.2
Test	926	7.73	13.64	47.70	551.82

3) 질의 확장(PRF): 1)의 검색 후 상위 4개의 검색 결과 상품 제목을 질의에 추가하여 검색
4) 질의 유형 기반 재순위화(Reranking): 딥러닝 기반 재순위화 모듈을 적용한 실험

실험에서는 각 단계별로 이전 단계의 결과에 방법을 추가하며 성능을 평가하였다. 평가는 TREC 2023에서 제공되는 평가도구와 QREL 파일을 사용하였다.

5.3 실험 결과

Table 4는 비교 실험 결과를 나타낸다. 정보검색 기법에 기본 성능이 높은 BM25 검색 결과와 비교했을 때, 본 연구의 제안 방법을 단계별로 적용했을 때 성능이 향상됨을 확인할 수 있다.

2) 질의 교정 : 성능 하락 요인에 대한 질의 처리를 적용한 결과 0.7540 (6.66% 향상)
3) 질의 확장 : 질의 확장 상위 검색된 상품 제목을 질의로 확장한 결과 0.7711(9.08% 향상)

최종적으로 딥러닝 기반 재순위화까지 모두 적용한 결과 NDCG@1000을 기준으로 0.7810으로 10.48% 향상되었다.

6. 결 론

본 연구에서는 상품 검색에서 발생할 수 있는 성능 저하 요인으로 오타, 다국어, 상품코드 형식의 질의를 분석하였다. 상품 코드 질의 등에 대한 상품 제목을 이용한 질의 확장이 성능을 높일 수 있음을 실험을 통해 확인하였다. 또한 딥러닝 기반 학습 모델에 적합한 질의의 속성을 반영하여 스페셜 토큰을 추가하였다. 추상적인 질의에 대해서는 딥러닝 기반 재순위화가 성능을 높이지 못함을 확인하였고 구체적인 상품 검색 질의에 대해서만 재순위화를 하도록 모델을 구성하였다.

본 연구에서 제안한 질의 유형에 따른 딥러닝 기반 재순위화 방법이 기본 검색 성능에 비해 10.48% 향상을 보임으로써 상품 검색에 유효한 방법임을 확인하였다.

향후연구로는 최신 정보검색 기술과 방법에 본 연구에서 분석한 질의의 특성을 반영한 검색 방법을 연구할 계획이다.

Table 4. Experiment Results

	NDCG@1000	NDCG@15	P@5
1) BM25 (baseline)	0.7069	0.6328	0.6871
2) Query preprocessing	0.7540	0.6953	0.7323
3) + Pseudo Relevance Feedback	0.7711	0.7121	0.7430
4) + Reranking	0.7810	0.7330	0.7774

References

[1] N. Asadi and J. Lin, "Effectiveness/efficiency tradeoffs for candidate generation in multi-stage retrieval architectures," In *Proceedings of the 36th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.997-1000, 2013.

[2] R. C. Chen, L. Gallagher, R. Blanco, and J. S. Culpepper, "Efficient cost-aware cascade ranking in multi-stage retrieval. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.445-454, 2017.

[3] L. Gao, Z. Dai, and J. Callan, "Rethink training of BERT rerankers in multi-stage retrieval pipeline," In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28-April 1, 2021, Proceedings, Part II 43* (pp.280-286). Springer International Publishing, 2021.

[4] Y. Nie, S. Wang, and M. Bansal, "Revealing the importance of semantic retrieval for machine reading at scale," *arXiv preprint arXiv:1909.08041*, 2019.

[5] A. Aizawa, "An information-theoretic perspective of tf-idf measures," *Information Processing & Management*, Vol.39, No.1, pp.45-65, 2003.

[6] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[7] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," *arXiv preprint arXiv:2004.04906*, 2020.

[8] X. Ma, J. Guo, R. Zhang, Y. Fan, Y. Li, and X. Cheng, "B-PROP: bootstrapped pre-training with representative words prediction for ad-hoc retrieval," In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.1513-1522, 2021.

[9] J. Zhan, J. Mao, Y. Liu, J. Guo, M. Zhang, and S. Ma, "Optimizing dense retrieval model training with hard negatives," In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.1503-1512, 2021.

[10] J., Zhan, J., Mao, Y., Liu, M., Zhang, and S. Ma, "Repbert: Contextualized text embeddings for first-stage retrieval," *arXiv preprint arXiv:2006.15498*, 2020.

[11] P. S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp.2333-2338, 2013.

[12] H. Shan, Q. Zhang, Z. Liu, G. Zhang, and C. Li, “Beyond two-tower: Attribute guided representation learning for candidate retrieval,” In Proceedings of the ACM Web Conference 2023, pp.3173-3181, 2023.
[13] Q., Zhang et al., “A semantic alignment system for multi-lingual query-product retrieval,” *arXiv preprint arXiv: 2208.02958*, 2022.
[14] X. Qin, N. Liang, H. Zhang, W. Zou, and W. Zhang, “Second place solution of Amazon KDD Cup 2022: ESCI Challenge for Improving Product Search,” 2022.
[15] J. Lin, L. Xue, Z. Ying, C. Meng, W. Wang, H. Wang, and X. Wu, “A Winning Solution of KDD CUP 2022 ESCI Challenge for Improving Product Search,” 2022.
[16] J. Park, W. Choi, G. An, and K. Lee, “Deep learning based reranking model for product search,” *Digital Contents Society*, pp.131-132 2023.
[17] G. An, W. Choi, J. Park, and K. Lee, “JBNU at TREC 2023 Product Search Track,” The Thirty-Second Text REtrieval Conference (TREC 2023), 2023.
[18] V. Ashish, “Attention is all you need,” *arXiv preprint arXiv: 1706.03762*, 2017.
[19] T. Barrus, “pyspellchecker,” 2024, accessed: 20.02.2024. [Internet], https://pypi.org/project/pyspellchecker/
[20] P. He, X. Liu, J. Gao, and W. Chen, “Deberta: Decoding-enhanced bert with disentangled attention,” *arXiv preprint arXiv:2006.03654*, 2020.

안 기 택

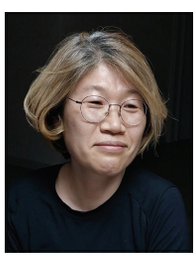
https://orcid.org/0009-0005-7131-8889
e-mail : gt@kfri.re.kr
2011년 남서울대학교 컴퓨터학(학사)
2023년 전북대학교 컴퓨터공학(석사)
2023년 ~ 현 재 전북대학교 컴퓨터공학전공 박사과정
2010년 ~ 2013년 ㈜다나와
2013년 ~ 2016년 ㈜코디웨어
2016년 ~ 현 재 한국식품연구원 선임기술원
관심분야 : 정보검색(Information Retrieval), 기계학습(Machine Learning), 데이터플랫폼(Data Platform) 등

최 우 석

https://orcid.org/0009-0001-0737-9776
e-mail : ccwwsss@jbnu.ac.kr
2021년 ~ 현 재 전북대학교 컴퓨터인공지능학부 학사과정
관심분야 : 자연어처리(Natural Language Processing), 정보검색(Information Retrieval) 등

박 준 용

https://orcid.org/0009-0009-6395-6572
e-mail : pjy010608@jbnu.ac.kr
2020년 ~ 현 재 전북대학교 컴퓨터인공지능학부 학사과정
관심분야 : 자연어처리(Natural Language Processing), 정보검색(Information Retrieval) 등



박 정 민

https://orcid.org/0009-0006-6678-8584
e-mail : parkjm@kfri.re.kr
1991년 이화여자대학교 생물학(학사)
2001년 한남대학교 경제학(석사)
2005년 한남대학교 경제학(박사)
2010년 ~ 현 재 한국식품연구원 지능화정책팀장(책임연구원)
관심분야 : 기술혁신(Technology Innovation), 연구데이터(Research Data) 등

이 경 순

https://orcid.org/0000-0003-2145-3802
e-mail : selfsolee@jbnu.ac.kr
1997년 한국과학기술원 전자전산학(석사)
2001년 한국과학기술원 전자전산학(박사)
2001년 ~ 2003년 일본 국립정보학연구소 연구원
2007년 ~ 2008년 미국 매사추세츠주립대학 방문교수
2004년 ~ 현 재 전북대학교 컴퓨터인공지능학부 교수
관심분야 : 정보검색(Information Retrieval), 기계학습(Machine Learning), 데이터 분석(Data Analysis)