추계학술대학생-2023-06호

상 장

입상 부문 : 은상

논문 제목 : 상품 검색을 위한 딥러닝 기반 재순위화 모델
저　　자 : 박준용, 최우석, 안기택, 이경순(전북대학교)

위 논문은 2023년 한국디지털콘텐츠학회 추계종합학술대회 대학생 논문경진대회 발표논문 중 은상 논문으로 선정되어 이 상을 수여함.

2023년 11월 10일

(사)한국디지털콘텐츠학회 회장 김영철

상품 검색을 위한 딥러닝 기반 재순위화 모델

박준용(*), 최우석(*), 안기택(**), 이경순(**)
(*) 전북대학교 컴퓨터인공지능학부(공동 1저자), {pjy010608, ccwwsss}@jbnu.ac.kr
(**) 전북대학교 컴퓨터인공지능학부(공동 교신저자), {gt, selfsolee}@jbnu.ac.kr

Deep Learning based Reranking Model for Product Search

Jun-Yong Park, Woo-Seok Choi, Gi-Taek An, Kyung-Soon Lee
Jeonbuk National University, Department of Computer Science & Artificial Intelligence

요 약

온라인 상품 검색은 질의가 추상적인 유형과 구체적인 상품을 요구하는 유형이 있어 사용자의 요구에 적합한 결과를 제시하기가 일반 정보검색에 비해 어려운 점이 있다. 또한 질의에 오타와 다국어, 상품코드를 포함하고 있어 이를 다뤄야 하는 문제를 포함하고 있다. 본 연구에서는 질의 유형을 분석하는 방법을 제안하고, 그 분석에 따라 상품 문서 결과에 대한 딥러닝 기반 재순위화 적용 여부를 판단하는 방법을 제안한다. 최근 우수한 성능을 보이는 딥러닝 모델 DeBERTa 을 이용하여 질의와 적합 문서에 대한 학습을 통해서 재순위화를 수행한다. 상품의 속성정보를 특별 정보로 처리함으로써 학습 효과를 높이도록 하였다. 국제정보검색 평가대회인 TREC2023 상품 검색 트랙에서 제공한 데이터를 활용한 평가에서 제안한 방법이 정보검색 기본 모델(BM25)에 비해 ndcg 기준 12.4% 성능이 향상됨을 확인하였다.

1. 서 론

온라인 상품 검색의 질의 유형은 “$10 Candles”, “$5 items”와 같이 추상적인 질의와 “boys purple under armour shirt”와 같이 구체적인 상품 정보를 포함하여 질의하는 유형이 있다. 다양한 질의 유형은 사용자에게 적합한 결과를 제시하기가 일반 정보검색에 비해 어려운 점이다. 또한 다양한 국적의 사용자가 검색하기 때문에 다국어 질의의 처리, “battery operated outsoor(outdoor) lamps”와 같이 질의에서 발생하는 오타, “0307339459”와 같이 상품의 코드를 직접 검색하는 경우를 다루어야 한다.

본 연구에서는 상품 검색에 특화된 질의어 처리 방법과 질의 유형에 따라 딥러닝을 적용하여 재순위화 하는 방법을 제안한다. 관련 연구로는 KDD Cup 2022 챌린지와 TREC 2023에서 아마존 상품 검색 데이터를 이용한 부문이 있었고 결과가 공개된 KDD Cup의 상위 순위의 연구팀에서 모두 딥러닝을 활용하였다. 다수의 연구에서 BERT를 기반으로 하는 DeBERTa, RemBERT, RoBERTa 등의 모델을 사용하였고[1,2] 그 외에도 infoXLM 등을 사용한 연구[3]가 있었다.

2. 상품 질의 분석을 통한 개선된 검색 모듈

본 연구에서는 상품 검색 성능 향상을 위해 질의 및 문서를 분석한 결과 “다국어 질의”, “오타가 있는 질의”, “상품 코드 형식의 질의” 3가지를 처리하는 방법을 제안한다.

- 오타 교정 모듈을 통한 질의 교정
- 번역 API를 통한 영문 번역
- 상품 코드 질의일 경우, 일치되는 상품명으로 변환
- 잠정적 적합성 피드백을 통한 질의 확장

질의 교정에 pyspellchecker를 상품에 특화하여 학습 과정을 거친 후 사용하였다. 기본 사전을 사용하면, “anker iphone chargeer”와 같은 문장이 교정되면서 “anger

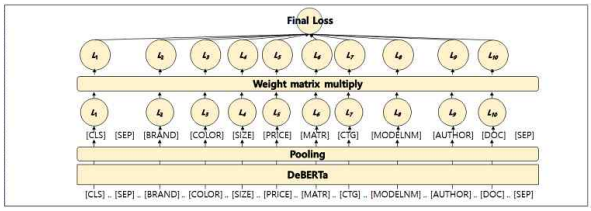
iphone charger"로 오타 이외에도 브랜드명이 변경되는 것을 보완할 수 있다. 이후, 질의에 대해 잠정적 적합성 피드백을 진행함으로써 질의를 확장했다.

"$5 items"와 같이 추상적인 질의의 경우, 상품과 관련된 구체적인 속성을 특정할 수 없어 질의와 관련된 상품 사이의 관계성이 모호한 부분이 있다. 따라서 모든 질의에 대하여 재순위화를 진행하는 것보다 구체적인 상품 질의에 대해 딥러닝을 적용하여 재순위화하는 방법을 제안한다.

3. 질의 유형에 따른 딥러닝 기반 재순위화

본 연구에서는 구체적인 속성이 있는 질의에 대하여 문서의 속성정보를 활용한 DeBERTa 기반의 딥러닝 재순위화 방법을 (그림 1)과 같이 제안한다. 상품 문서의 속성 데이터 중 질의에서 많이 나타나는 브랜드, 색상, 크기, 가격, 재료, 카테고리, 모델명, 저자, 제목, 세부설명 10개를 스페셜 토큰과 함께 사용하고 손실함수 계산에서 [질의에 스페셜 토큰의 값이 포함될 경우 해당 토큰의 가중치를 줘 연산 과정에서 해당 정보가 최종 손실 계산에서 크게 영향을 줄 수 있도록 하는 방법으로 상품 검색에 영향을 끼치는 상품 속성을 학습에 적용하는 방법을 제안한다.

(그림 1) 딥러닝 기반 재순위화 모델



(Figure 1) Deep Learning-Based Re-ranking Model

4. 실험 및 분석

실험에는 30,734개(훈련셋 20,888개/평가셋 9,846개)의 질의와 1,118,658개의 문서 데이터를 사용하였다. 문서는 Exact(정확), Substitute(대체가능), Complement(보완가능), Irrelevant(무관)로 분류되어 있다. 딥러닝 하이퍼파라미터는 학습률 1e-5, 배치 사이즈 7, 토큰 길이 512, Cross Entropy Loss 와 Adam을 사용하였다. <표 1>의 실험 결과에 따르면 본 연구에서 제안하는 방법이 BM25 검색 결과 대비 ndcg 기준 12.4% 향상됨을 확인하였다.

<표 1> 실험 결과

		ndcg	ndcg@15	P@5
BM25		0.3740	0.2244	0.2624
질의 교정 모듈 적용		0.3686	0.2321	0.2633
+ 질의 확장(PRF) 적용		0.4099	0.2456	0.2739
+ 질의 유형 기반 Reranking	전체 질의	0.4163	0.2544	0.2788
	구체적 속성 질의만	**0.4204**	**0.2601**	**0.2823**

<Table 1> Experiment Result

5. 결론

본 연구에서는 상품 질의 분석에 따라 딥러닝 기반 재순위화 적용 여부를 판단하는 방법을 제안하였다. 상품 질의의 오타를 교정하는 특화된 모듈의 사용과 번역, 상품코드 변환, 질의 확장, 구체적인 속성이 있는 질의에 대하여 문서 속성정보를 활용하는 딥러닝 기반 재순위화 방법을 실험을 통해 확인하였다. 본 연구에서 제안하는 방법이 ndcg 기준 12.4% 향상되어 검색 성능 향상에 효과적임을 확인하였다.

감사의 글

본 연구는 2023과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음 (2022-0-01067)

참고 문헌

[1] Q Zhang, Z Yang, Y Huang, Z Chen, Z Cai, K Wang, J Zheng, J He, J Gao, "A Semantic Alignment System for Multilingual Query-Product Retrieval", arXiv preprint arXiv, 2208.02958, 2022.

[2] X Qin, N Liang, H Zhang, W Zou, W Zhang, "Second place solution of Amazon KDD Cup 2022: ESCI Challenge for Improving Product Search", 2022.

[3] J Lin, L Xue, Z Ying, C Meng, W Wang, H Wang, X Wu, "A Winning Solution of KDD CUP 2022 ESCI Challenge for Improving Product Search", 2022.