온라인 가공식품의 수량과 중량에 따른 최저가격 검색 모델

최태민 [1], 임희석 [2]

[1] 고려대학교 인공지능융합학과 석사과정
[2] 고려대학교 컴퓨터학과 교수

never7653@korea.ac.kr, limhseok@korea.ac.kr

A Model for Minimum Price Search of Processed Food Items on Online Platforms Based on Quantity and Weight

Tae-Min Choi[1], Heui-Seok Lim[2]
[1]Dept. of Applied Artificial Intelligence, Korea University
[2]Dept. of Computer Science, Korea University

요 약

가공식품이라는 특정 도메인에서는 기존 검색엔진에서 많이 활용되는 BM25 만을 가지고 최저가 검색하는 데는 어려움이 있다. 본 논문에서는 BM25 외에도 검색의 정확성을 높이기 위해 HuggingFace 에 공개되어 있는 KoELECTRA 를 활용하여 개체명 인식(Named Entity Recognition 과 이진 분류모델(Binary Classification)을 Fine-tuning 하고 BM25 와 연계하여 구축한 검색시스템을 제안한다. 기존의 BM25 대비 성능 평가를 통해 효과를 검증하였다.

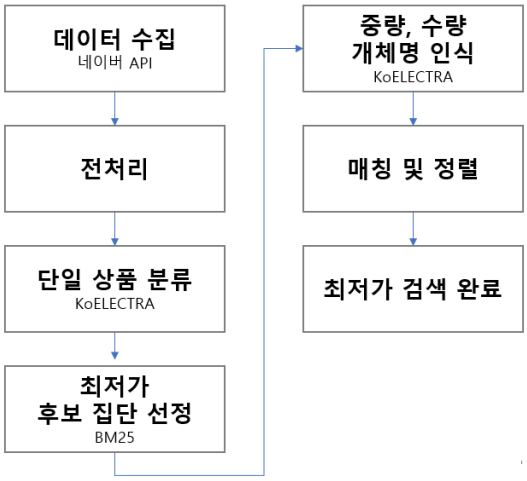
1. 서론

가공식품을 제조/판매하는 판매자들은 본인의 판매 채널(D2C mall, 소셜커머스 등)에서 판매하고 있는 제품이 다른 오픈 도메인에서 시장가격이 어떻게 형성되고 있는지 파악하고 가격경쟁력 있는 가격정책 설정이 필요하다. 별도 시스템이 구축되어 있지 않은 판매자들은 수기로 몇몇 경쟁사이트에 접속하여 시간 제약에 의해 일부 핵심 상품만 시장 가격을 조사하고 있다. 경쟁력 있는 가격정책 수립을 위해 판매하고 있는 상품을 각 경쟁 사이트에서 주기적으로 웹 크롤링을 이용하여 데이터를 수집하고 수집된 데이터를 가공하여 최저가격을 탐색할 수 있는 검색시스템의 구축이 필요하다. 가공식품의 최저가를 검색하는데 문서검색에서 우수한 성능을 보이는 BM25 를 사용할 수 있지만 문제점이 있다. 가공식품의 경우 일반 의류 등 과는 다르게 같은 상품이지만 중량, 수량에 따라 가격이 다르게 형성되어 있어 구체적인 중량, 수량으로 검색 가능해야 한다. 하지만 띄어쓰기 및 중량, 수량 표기가 일정한 패턴이 없고 판매자가 제품명을 등록할 때, 규칙 준수에 대한 의무가 없기 때문에 오픈도메인에서 검색이 잘 되도록 하기 위해 다양한 형태로 제품 등록을 하게 된다. 일반적으로 검색엔진에서 많이 사용되는 엘라스틱서치 기반의 BM25 는 토큰 빈도수 기반이기 때문에 검색 쿼리의 토큰과 명시적으로 일치해야 한다는 한계가 있다. 최근 검색시스템에는 이런 기존 BM25 의 한계를 극복하기 위해 의미적 검색이 가능하도록 개체명 인식, 이미지, 지식그래프 등을 활용한 딥러닝 연구가 많이 진행되고 있다. 본 논문에서는 추가 정보 없이 제품명 데이터만을 대상으로 BM25 와 KoELECTRA 를 Fine-tuning 한 개체명 인식과 이진분류 모델을 연계한 검색시스템을 제안한다.

2. 최저가 검색 시스템

본 논문에서 제안하는 시스템은 (그림 1) 과 같다. AI 모델은 HuggingFace 에 공개되어 있는 KoELECTRA 모델을 개체명 인식, 복합상품분류 모델로 각각 Fine-

tuning 하였다.



(그림 1) 시스템 구성도

2.1. 데이터 수집 및 전처리

데이터는 네이버 쇼핑 API 를 통해서 주요 소셜미디어 및 오픈마켓에서 데이터를 수집하였다. 훈련용 데이터는 제한된 환경에서 실험을 진행하기 위해 총 102 개의 상품을 지정하여 검색어를 별도로 구성하였다. 정규식을 활용하여 특수문자를 제거하고 대문자를 소문자로 치환하고 중량과 수량의 띄어쓰기를 일정한 패턴으로 변환하여 (예시 1)과 같이 될 수 있도록 전처리하였다.

(예시 1) '참치 150G50 개 → 참치 150g 50 개'

2.2. 단일 상품 분류

하나의 단일 상품이 아닌 성격이 다른 상품과 결합하여 하나의 상품으로 판매하는 것을 복합상품이라 한다. 예를 들어, 햄버거와 피자를 결합하여 판매하거나 불고기버거와 치즈버거를 결합하여 판매하는 것 또한 복합 상품으로 분류할 수 있다. 본 논문에서는 사용자들이 단일 상품을 검색하는 경우의 수가 많다고 가정하고 단일 상품에 대한 검색 성능을 높이기 위해 HuggingFace 에 공개된 KoELECTRA 를 17,942 개의 데이터로 Fine-tuning 하여 이진 분류 모델을 개발하였다. 복합상품이 아닌 단일 상품만이 시스템의 다음 순서로 진행될 수 있게 된다.

2.3. 최저가 후보 집단 선정

수집한 데이터들 중 검색 대상 상품과 수량과 중량이 동일한 상품의 최저가를 탐색하기 위해서는 우선 크롤링 데이터에 대한 후보집단 선정이 필요하다. 본 논문에서는 정보 검색 분야에서 이미 성능이 입증된 BM25 를 사용하여 후보 집단 선정에 활용하였다. 실제 연구에서는 BM25 가 내장되어 있고 검색시스템에 많이 활용되는 엘라스틱서치를 활용하여 구현하였다. 또한, 한국어 데이터의 효율적인 처리를 위해 분석기로 엘라스틱서치의 플러그인 중 하나인 노리 토크나이저를 사용하였다. 기존 시스템에서 BM25 가 직접적으로 최적가격 상품을 검색하는 것이 목적이라면 본 시스템에서는 동일 제품 품목만을 분류하여 최저가 후보 집단을 선정하는데 목적이 있다.

2.4. 개체명 인식

후보집단에서 바로 가격순으로 정렬하면 정확히 매칭된 결과를 얻기 어려운 경우가 있다. 후보집단에서 최저가 상품을 정확히 검색하기 위해서는 검색 대상과 데이터들의 중량과 수량을 각각 파싱해서 매칭해야 한다. (표 1) 과 같이 중량의 단위가 모호하거나 수량의 단위가 모호한 경우는 정규식을 활용하여 별도로 구분해내기 어렵다. 그래서 본 연구에서는 개체명 인식을 통해 의미적으로 수량, 중량을 파싱하여 검색의 정확도를 향상시켰다.

제품명	정규식	개체명 인식
참치 100g 10 개	중량:100g 수량:10 개	중량:100g 수량:10 개
참치 100g 10 (옵션 : 100 그람 10)	중량:100g, 수량:	중량:100g, 100 그람 수량:10, 10

(표 1) 정규식과 개체명 인식 결과 비교

단일 상품 분류와 동일하게 KoELECTRA 모델을 사용하였고 '<weight>, <quantity>' 2 가지 개체명으로 17,942 개의 데이터로 Fine-tuning 진행하였다.

2.5. 매칭 및 정렬

Weight 와 Quantity 을 매칭하고 매칭되는 데이터를 가격순으로 정렬하여 가격이 가장 낮은 데이터를 최종 검색 결과로 도출하였다.

3. 실험 결과

본 연구의 효과를 검증하기 위해 테스트용 데이터를 수집한 뒤 수집된 데이터가 100 개 이상인 상품, 총 54 개를 대상으로 실험을 진행하였다. 기존의 BM25 만 사용하였을 때, BM25 와 단일 상품 분류기를 같이 사용하였을 때, BM25, 단일 상품 분류기 및 정규식을 활용한 중량, 수량 인식기를 사용하였을 때, BM25, 본 연구의 시스템으로 실험하였을 때 이렇게 네 가지 케이스로 분류하여 실험 결과를 도출하였다.

웹 크롤링으로 수집한 상품들 중 검색 대상과 최종적으로 매칭된 상품이 실제 최저가일 경우를 최저가 탐색의 성공으로 정의하였다. 본 논문에서는 (수식 1)과 같이 54 개의 대상 상품(TI) 중 최저가 탐색을 성공한 상품 수(SI)의 비율인 성공률(SR)로 성능 지표로 평가하였다. (수식 1)

$$SR = \frac{SI}{TI} * 100\% \qquad (1)$$

실험 구분	탐색 성공 상품 수(SI)	성공률 (SR)
BM25	7	12.9%
BM25 + 단일 상품 분류기	8	14.8%
BM25 + 단일 상품 분류기 + 정규식	30	55.5%
BM25 + 단일 상품 분류기 + 개체명 인식	36	66.7%

(표 2) 성능 지표

(표 2)와 같이 BM25, 단일 상품 분류기, 정규식 조합 대비 제안한 시스템이 11.2%p 더 나은 검색 결과를 도출함을 알 수 있다.

4. 결론 및 향후 연구 과제

본 논문에서는 가공식품이라는 특정 도메인의 상품 검색에서 제품명만을 가지고 최저가를 검색할 때, 중량과 수량의 표기 및 수식어 등에 따른 검색 성능 정확도 저하를 자연어처리의 개체명 인식과 이진분류기로 향상시킬 수 있는 시스템을 제안하였다.

실험을 통해서 BM25 만으로 단일 시스템 대비 제안한 시스템이 더 나은 효과가 있음을 보였다.\

추가적인 성능 개선을 위해서는 최저가 후보 집단 중 중량과 수량은 검색 대상과 같지만 품목 분류가 다른 상품이 있어 성공률에 부정적인 영향을 주는 경우가 있는데 최저가 후보 집단 선정 모델을 키워드 기반의 BM25 가 아닌 semantic search 알고리즘으로 변경하면 성능 향상을 예상할 수 있다. 또한, 데이터를 늘려서 개체명 인식 모델을 학습하거나 다른 한국어 사전학습 모델을 사용한다면 개체명 인식율을 개선하여 검색 성능을 높일 수 있을 것이다.

향후 연구과제로는 본 논문의 시스템처럼 별도 모델 학습을 하지 않고 최근 자연어처리 분야의 주목받는 연구 분야인 OpenAI 의 ChatGPT 같은 LLM(Large Language Model)을 활용하는 방법론이 있다. Fewshot 을 이용한 Prompt Engineering 으로 더 나은 결과를 가져올 수 있을 지에 대한 실험이 필요하다.

본 논문은 가공식품 제조/판매하는 판매 채널들의 가격 경쟁력 확보를 위한 가격 정책 수립에 긍정적인 영향을 끼칠 수 있음을 기대한다. 더 나아가 검색된 가격을 기준으로 내부 가격정책을 정한다면 경쟁사에 따른 자동 가격 변동 시스템을 구축할 수 있다. 또한, 일부 판매자들을 대상으로는 수기 업무를 줄여 업무 환경 개선에 기여할 수 있음을 예상한다.

참고문헌

[1] Clark, K., Manning, C. D., Luong, M.-T., & Le, Q. V. “Electra: Pre-Training Text Encoders as Discriminators Rather Than Generators.” 8th International Conference on Learning Representations, ICLR, 2020.

[2] Zhen Yin, Scott Uk-Jin Lee. “ Detecting Cross-Site-Script Attacks using BM25 Algorithm.” 한국정보과학회 2022 한국컴퓨터종합학술대회, 대한민국, 2022, p. 1,303-1,305.

[3] 정현정, 김현희 “개체명 인식을 이용한 소셜 미디어에서의 약물 부작용 표현 추출 및 분”, 한국정보처리학회 학술대회논문집, 28(1), 2021, p. 443

[4] J. W. Park. “KoELECTRA: Pretrained ELECTRAModel. for Korean.” Github Repository. https://github.com/monologg/ KoELECTRA.

[5] 김재원, 신우성, 김영섭. “의류 검색 프로그램을 위한 Cnn 기반 의류 분류 모델”, 한국통신학회 학술대회논문집, 2022(2), p. 947–948.