ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2023 September: 29(3): 1~17 http://dx.doi.org/10.13088/jiis.2023.29.3.001

확장된 사용자 유사도를 이용한 CF-기반 건강기능식품 추천 시스템*

홍세인
경희대학교 대학원 빅데이터응용학과
(h110505@khu.ac.kr)

정의주
경희대학교 대학원 빅데이터응용학과
(euiju1011@khu.ac.kr)

김재경
경희대학교 경영대학 & 대학원 빅데이터응용학과
(jaek@khu.ac.kr)

정보통신기술의 발전과 디지털 기기의 대중화로 인해, 온라인 시장의 규모가 커지고 있다. 그 결과 고객들은 상품을 선택하는데 많은 시간과 비용이 소요되는 정보 과부하(Information Overload) 문제에 직면하고 있다. 따라서 고객이 선호할 만한 상품을 추천해 주는 추천 시스템은 필수적인 도구가 되었으며 협업 필터링(Collaborative Filtering) 기법은 가장 널리 쓰이는 추천 방법이다. 전통적인 추천 시스템은 평점과 같은 정량적인 데이터만을 사용하기 때문에 추천의 정확도는 높지 않다. 이와 같은 문제를 해결하기 위해 요즘에는 사용자 리뷰와 같은 정성적 데이터를 반영하는 연구가 활발히 진행되고 있다. 협업 필터링의 일반적인 절차는 사용자-상품 행렬 생성, 이웃 집단 탐색, 추천 목록 생성 3단계로 구성되며 코사인 같은 사용자 유사도를 사용하여 목표 고객의 이웃을 탐색하며, 추천 상품 목록을 생성한다. 본 연구에서는 이웃 집단 탐색 및 추천 목록 생성 단계에서 사용하는 사용자 간의 유사도를 기존의 사용자 평점을 이용한 유사도에 고객의 리뷰 데이터를 사용하는 확장된 사용자 유사도를 제시한다. 리뷰를 정량화 하기 위해 본 연구에서는 텍스트 마이닝을 활용한다. 즉, 리뷰 데이터에 TF-IDF, Word2Vec, 그리고 Doc2Vec 기법을 사용하여 두 사용자 간의 리뷰 유사도를 구한 후 사용자 평점을 사용한 유사도와 리뷰 유사도를 결합한 확장된 유사도를 생성하는 것이다. 이를 검증하기 위해 전자상거래 사이트인 Amazon의 'Health and Personal Care'의 사용자 평점과 리뷰 데이터를 사용하였다. 실험 결과, 사용자 간 유사도를 산출할 때 기존의 평점에 기반한 유사도만을 사용하는 것보다, 사용자 리뷰의 유사도를 추가로 반영한 확장된 유사도를 사용하면 추천의 정확도가 높아진다는 것을 확인했다. 또한, 여러 텍스트 마이닝 기법 중에서 TF-IDF 기법을 사용한 확장된 유사도를 이웃 집단 탐색 및 추천 목록 생성단계에서 사용할 때의 성능이 가장 좋게 나타났다.

주제어 : 리뷰 텍스트,협업 필터링, TF-IDF, Word2Vec, Doc2Vec

논문접수일 : 2023년 2월 10일 논문수정일 : 2023년 4월 17일 게재확정일 : 2023년 4월 20일
원고유형 : Regular Track 교신저자 : 김재경

1. 서론

최근 정보통신기술의 발전으로 온라인 쇼핑이 대중화되면서 전자 상거래 또한 매우 빠르게 발전하고 있다. 고객들이 제공 받을 수 있는 상품 정보가 기하급수적으로 늘어남에 따라 고객은 자신의 선호도에 맞는 상품과 서비스를 고르는 것에 어려움을 겪는 정보 과부하(Information Overload) 문제가 발생할 수 있다(김병만 et al., 2004). 이러한 문제를 해결하기 위해 개인에게 맞춤형 서비스를 제공하는 것이 중요해지면서 추천 시스템의 중요성이 대두되고 있다. 특히 Amazon, Netflix, Youtube, Google 등 대형 온라인 플랫폼은 개인의 선호도를 고려하여 상품 및 서비스를 추천 해

* 본 논문은 연구재단 4단계 BK21 사업으로부터 지원받은 연구임.

주는 시스템을 적극적으로 활용하고 있다(Lee & Hosanagar, 2019). 아마존(Amazon)의 경우, 사용자에게 제공되는 추천 목록의 60%가 최종 구매 행동으로 이어졌으며, 이는 전체 매출의 35% 향상에 기여 했다고 밝혔다. 또한 넷플릭스(Netflix)의 사용자들이 시청한 영화 중에서 75%가 추천 서비스로 이루어지고 있다(MacKenzie et al., 2013).

협업 필터링(CF, Collaborative Filtering) 기법은 전자상거래 분야에서 개인화 추천 서비스를 제공 할 때 가장 널리 사용되고 있으며 추천 시스템 알고리즘 중 가장 우수한 성능을 보여준다. 사용자 기반 협업 필터링은 고객의 평점, 구매 여부, 클릭 여부 등, 정량적 선호도 정보를 바탕으로 유사한 이웃 고객을 선정한다. 이는 상품에 대한 정량적 선호도가 유사한 사용자들은 다른 상품에 대해서도 비슷한 선호도를 가진다는 전제를 바탕으로 한다(Goldberg et al., 1992). 이후 선정된 이웃 목록의 구매 기록으로부터 타겟 고객의 선호도를 예측하여 상품 및 서비스를 제공한다.

전통적인 협업 필터링에서는 고객의 정량적 선호도만을 사용하여 추천을 하기 때문에 사용자의 정성적 선호도가 충분히 반영되지 않는 한계가 발생한다(Zhang et al., 2014). 이러한 문제를 해결하기 위해 사용자의 정성적인 선호도를 추천시스템에 반영하려는 연구가 활발히 진행되었다. 특히 사용자가 작성한 리뷰 텍스트 정보를 반영한 연구가 많은 주목을 받고 있다(김유영 & 송민, 2016; Choeh et al., 2013; Zheng et al., 2017). 리뷰 데이터는 사전에 정해진 기준 없이 본인의 의견을 자유롭게 작성할 수 있다는 특징 때문에, 특정 상품에 대해 구체적이고 상세한 구매이유, 제품의 사용후기 등이 담겨있다.

본 연구에서는 정량적인 평점 정보 뿐만 아니라 사용자가 남긴 리뷰 데이터를 활용하고자 한다. 새롭게 제안하는 방법론은 다음과 같다. 첫째, 평점 정보에 나타난 사용자의 정량적인 선호도를 기반으로 사용자 간의 유사도 행렬을 생성한다. 둘째, 다양한 텍스트 마이닝 기법을 사용자가 남긴 리뷰 데이터에 적용하여 정성적인 선호도를 도출하고, 이를 바탕으로 사용자 간의 유사도 행렬을 생성한다. 셋째, 두 유사도 행렬을 결합한 확장된 유사도를 생성한다. 마지막으로 유사 이웃 집단을 탐색할 때 사용하는 유사도와 평점을 예측하여 추천 목록을 생성할 때 사용하는 유사도를 앞서 구한 평점, 리뷰 및 확장된 유사도를 사용하여 추천 알고리즘을 생성한다. 본 연구에서는 Mean Absolute Error(MAE), Root Mean Squared Error(RMSE) 평가지표를 사용하여 제안하는 추천 방법론의 성능을 평가하였다.

본 연구에서 새롭게 제안하는 리뷰를 반영하는 확장된 유사도를 사용하는 추천 알고리즘이 단일 평점 정보만 반영한 기존의 추천 알고리즘보다 우수한 결과를 나타냈다. 또한, TF-IDF, Word2Vec, Doc2Vec 기법을 사용하여 텍스트를 임베딩 하는 방법을 다르게 적용하고 협업 필터링 단계에 사용하는 유사도가 달라짐에 따라 변화하는 추천 알고리즘 성능을 비교한 결과, 이웃 집단 탐색 및 추천 목록 생성 단계에서 TF-IDF 기법을 사용하여 리뷰 유사도를 반영할 때의 추천 정확도가 가장 높은 것을 확인하였다.

본 논문의 구성은 다음과 같다. 제2장에서는 협업 필터링과 데이터 마이닝에 관련 연구에 대해 살펴본다. 제3장에서는 본 연구에서 제안하는 방법론에 대해 구체적으로 살펴보고, 제 4장에서는 제안한 추천 알고리즘의 성능 검증을 위한 데이터 소개 및 파라미터 세팅, 실험 설계 및 실험 결과에 대해 기술한다. 마지막으로 제 5장에서는 본 연구의 결론과 시사점, 한계점 및 향후 연구

계획에 대하여 기술한다.

2. 관련연구

2.1. 협업 필터링

추천 시스템은 사용자가 상품에 부여한 평점, 과거 구매 기록, 클릭 여부 등의 정량적 정보를 기반으로 사용자가 선호할 만한 상품이나 서비스를 제공하는 기법이다(이승우 et al., 2022). 온라인 플랫폼이 제공하는 정보가 기하급수적으로 증가하고 있기 때문에 이러한 추천 시스템의 중요성이 대두되고 있다. 특히 기업은 추천 시스템을 활용하여 효과적인 자원 관리 및 고객 관리, 매출 증대 등의 다양한 혜택을 얻을 수 있다. 따라서 글로벌 기업 및 각계 산업 분야에서 추천 서비스를 도입하고 있으며, 이에 관한 연구들도 활발히 진행되고 있다(Marlin et al., 2012; Paradarami et al., 2017).

추천 시스템 알고리즘 중 가장 우수한 성능을 보여주는 협업 필터링(CF, Collaborative Filtering)은 상품 혹은 사용자 간 유사성을 기반으로 추천한다. 구매 이력이 유사한 사용자에 대한 선호도를 기반으로 사용자와 상품 간의 유사도를 계산한 다음 구매하지 않는 상품에 대한 목표 사용자의 선호도를 예측하고 상품을 추천하는 알고리즘이다(Kim & Ahn, 2009; Resnick et al., 1994). 일반적으로 협업 필터링의 절차는 사용자-상품 행렬 생성, 이웃 집단 탐색, 추천 목록 생성 3단계로 구성된다. 사용자 기반 협업 필터링과 아이템 기반 협업 필터링은 각 단계에서는 기준이 사용자인지 상품인지만 다르고 기본 개념은 같기 때문에 사용자 기반 협업 필터링을 기준으로 설명한다. 첫번째 단계는 사용자가 상품에 남긴 평점을 사용하여 사용자가 남긴 평점을 기반으로 사용자-상품 행렬을 생성한다. 두번째 단계는 협업 필터링 단계에서 가장 중요한 단계로 사용자 간의 유사도를 계산하여 이웃 집단을 탐색하는 과정이다. 흔히 사용되는 유사도 측정 방법은 피어슨 상관계수(Pearson correlation coefficient), 코사인 유사도(Cosine similarity) 등을 사용한다. 사용자 간 유사도가 산출되면 이를 기반으로 추천 타겟 고객과 상품 구매 패턴이 유사한 고객들로 이웃 집단을 형성한다. 마지막 단계는 추천 목록 생성단계로, 타겟 사용자가 구매하지 않은 상품들 중에서 구매할 가능성이 가장 높을 것이라고 예상되는 상위 K개의 상품을 선택하여 추천 목록을 생성한다. 두번째 단계에서 구한 평점 유사도를 사용하여 구매 가능성 점수(PLS, Purchase Likelihood Score)를 계산한다. PLS값이 클수록 상품을 구매할 확률이 높은 것을 의미한다.

2.2. 사용자 리뷰를 사용한 추천 서비스

기존 협업 필터링 관련 연구에서는 구매 기록 및 평점 등의 정량적 선호도만을 사용하여 추천하였다. 그러나 정량적 선호도만을 사용하여 추천할 경우, 사용자의 정성적 선호도는 반영하지 못하기에 추천 성능이 떨어지는 문제가 발생한다(Lu et al., 2015; Srifi et al., 2020). 최근에는 이러한 문제점을 개선하기 위해 다양한 추가 정보를 사용하는 연구들이 진행되고 있으며, 대표적으로 사용자가 작성한 리뷰를 사용한 연구가 활발히 진행되고 있다(Choeh et al., 2013; Zheng et al., 2017). 리뷰는 사용자가 상품에 대한 정보와 사용 후기 등을 텍스트로 남긴 정성적 정보이다. 이러한 리뷰에는 사용자의 상품에 대한 솔직하고 상세한 정보다 담겨 있다.

전병국과 안현철(2015)은 추천 시스템에서 사용자 간의 유사도를 계산할 때 리뷰 데이터의 유사도를 같이 고려한 새로운 추천 알고리즘을 제안했다. 리뷰 유사도를 산출하는 기법으로는 TF-IDF 방식을 적용했으며, 단어의 등장 빈도 가중치를 고려하는 것이 결과 향상에 기여했다고 밝혔다. Terzi et al.(2014)는 평점이 아닌 리뷰의 유사도를 기반으로 사용자 간의 유사도를 계산하는 사용자 기반 기법의 변형을 제안했다. 장예화 등(2021)은 리뷰 데이터를 이용하여 기존 방식보다 빠르고 정확하게 유사도를 계산 하는 2단계 하이브리드 추천 알고리즘을 제안했다. 강부식(2018)은 상품별로 구매 사용자 집합을 문장으로 대응해서 Word2Vec 기법을 적용했다. 추출된 사용자 벡터를 기반으로 유사도를 구하고, 이를 사용자 기반 협업 필터링에 활용하였다. 현지연 등(2019)은 감성분석을 통해 리뷰 텍스트를 정량화 하고 평점과 결합하여 새로운 가중 평점을 생성하여 추천 알고리즘에 적용하였다. 이처럼 정성적 데이터인 리뷰 데이터를 정량화 하여 추천 알고리즘에 적용하는 것이, 정량적인 평점 정보만 고려한 추천 알고리즘보다 우수한 성능을 보여주는 것을 증명하였다. 리뷰에는 사용자가 상품을 구매한 이유 및 상품에 대한 선호도 등이 구체적으로 반영되어 있기 때문에 개인화 추천 서비스를 제공할 때 유용하게 사용된다.

2.3. 텍스트 마이닝

텍스트 마이닝은 자연어로 구성된 대량의 비정형 텍스트 데이터에서 숨겨진 패턴 또는 관계를 추출하여 의미 있고 활용가치가 높은 정보 또는 지식을 추출하는 일련의 분석 기법을 의미한다(Hearst, 1999). 정성적인 리뷰 텍스트를 정량화하기 위해서는 벡터(Vector) 형식으로 표현하는 임베딩 방식이 사용된다. 임베딩 방식을 이용한 방법으로는 TF-IDF, Word2Vec, Doc2Vec기법 등이 있다(정지수 et al., 2019).

TF-IDF(Term Frequency-Inverse Document Frequency) 기법은 여러 문서 내에서 단어가 특정 문서내에서 얼마나 중요한지를 평가하는 방식이다. TF(Term Frequency)값은 한 문서 내에서 특정 단어가 출현한 빈도수를 의미한다. 문서에서 출현한 빈도수가 큰 단어일 수록 해당 문서에서의 중요도가 높은 것을 기본 전제로 한다. 하지만, 단순히 단어 출현 빈도가 높다고 문서의 연관성을 높게 판단하기에는 오류가 있다. 관사와 같은 불용어는 문장과 연관성이 낮지만 자주 출현하기 때문에 문서간 연관성이 높다고 판단 할 수도 있다. 이런 오류를 줄이기 위해 IDF(Inverse Document Frequency) 값을 활용한다. IDF값은 문서 집합에 포함되어 있는 문서 수를 특정 단어가 나타난 문서의 수로 나눈 값이다. 많은 문서에서 나타나는 보편적인 단어의 IDF값은 작아지게 되고, 문서 내에서 중요한 의미를 가지는 단어의 IDF값은 커지게 된다. TF-IDF값은 TF값과 IDF값을 곱한 값으로, TF-IDF값이 클수록 적은 수의 문서 집합에 등장하지만 각 문서내에서의 빈도수는 높다.

자연어 처리를 위해 사용된 예측 기반 방법은 딥러닝(Deep Learning) 기법에 기반한 Word2Vec과 Doc2Vec이 있다. 2013년 구글 연구 팀이 발표한 기법인 Word2Vec은 단어의 문맥적 의미를 보존하여 단어를 벡터로 표현하는 임베딩(embedding) 방식이다. 특정 차원의 벡터 공간상으로 변환하여 표현하는 기법으로 의미적으로 유사한 단어 및 문서는 벡터 공간 상에서 가깝게 매핑(mapping)된다(Mikolov et al., 2013). Word2Vec의 학습 방법은 CBOW(Continuous Bag of Words)와 Skip-Gram

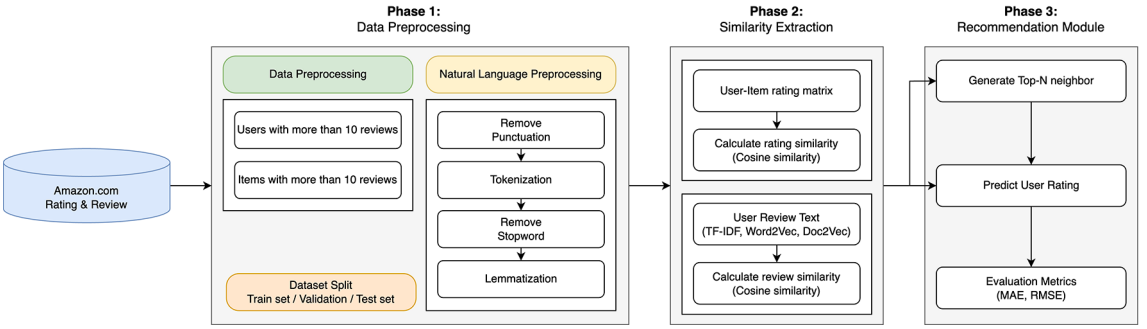
두가지 방법이 있다. CBOW는 주변에 있는 단어들을 기반으로 타겟 단어를 예측하고, Skip-Gram은 타겟 단어를 기반으로 주변 단어들을 예측한다. 이처럼 Word2Vec은 단어들의 전후 관계를 학습하고 단어 자체의 의미를 벡터 형식으로 표현하기 때문에 복잡한 개념 표현 및 의미론적 추론이 가능하다. Yin et al.(2018)은 영화 데이터인 IMBD Dataset을 사용하여 사용자가 작성한 리뷰의 감성점수를 파악하여 영화를 평가하였다. 이때, 리뷰의 벡터 값을 구하기 위해 Word2Vec 기법을 적용하여 리뷰 임베딩을 하였다. 영화 데이터(감독, 배우, 제작연도 등)로 Word2Vec 학습 기법을 활용한 영화 추천 시스템을 제안하였다. Yoon & Lee.(2018)은 영화 데이터인 MovieLens를 사용하여 평점 및 메타데이터(감독, 배우, 제작연도 등)을 사용하여 Word2Vec 기법을 활용한 영화 추천 시스템을 제안하였다. 강부식(2019)은 영화 데이터인 filmtrust를 사용하여 사용자 벡터와 영화 벡터를 생성하고, 사용자 벡터를 입력받아 평점을 예측하는 합성곱 모델과 영화 벡터를 입력 받아 평점을 예측하는 합성곱 모델을 구성한 후, 두 모델을 하나로 연결하는 앙상블 합성곱 신경망 모델을 제안하여 향상된 추천 정확도를 보여주었다. 임민아 등(2022)은 Word2Vec 기법을 사용하여 텍스트를 임베딩 하여 텍스트 유사도 값을 이용해 추천하는 알고리즘을 제안하였다.

Doc2Vec은 Word2Vec의 확장된 개념으로 문서를 벡터로 표현하는 임베딩 방식이다. 입력 부분에 문서 번호(paragraph id)가 추가 되어 각 문서에 고유 번호를 받아 새로운 단어처럼 모델에 추가한다. Doc2Vec의 학습 방법은 DM(Distributed Memory)과 DBOW(Distributed bag of Words) 두가지 방법이 있다. PV-DM(Paragraph Vector-Distributed Memory)은 문서 번호와 이전 단어들로 특정 단어를 예측하며 학습을 하고, CBOW(Paragraph Vector-Distributed Bag of Words)는 문서 번호와 문서 내 단어들을 예측하면서 학습을 한다. 김영수 등(2020)은 자기소개서 데이터에 Doc2Vec 기법을 활용하여 구직자의 합격 여부를 분류하는 모델을 제안하였다. 이윤주 등(2021)은 사용자의 검색 키워드와 사용자가 구매한 상품 정보에 Doc2Vec 기법을 사용하여 추천 알고리즘을 생성하였다.

본 연구는 임베딩 기법 중 TF-IDF, Word2Vec의 CBOW 기법과 Doc2Vec 기법을 사용하여 리뷰 데이터에서 유사도를 계산하고 추천 알고리즘에 적용하였다.

3. 확장된 사용자 유사도를 사용한 CF 방법론

전통적인 협업 필터링 연구에서는 평점과 같은 정량적인 데이터만을 반영한 추천 알고리즘을 제안하였지만, 최근에는 사용자의 구매동기 및 정성적인 선호도를 정교하게 반영한 연구가 활발하게 진행되고 있다. 본 연구는 다양한 텍스트 마이닝 기법을 활용한 리뷰 유사도를 추가로 반영하여, 협업 필터링의 성능을 개선하는 새로운 추천 알고리즘을 제안하다. 본 연구에서 제안하는 알고리즘은 <Figure 1>과 같이 먼저 수집된 데이터에 대해 전처리를 수행한다. 이때 고객이 작성한 리뷰 텍스트는 토큰화, 불용어 제거 등 자연어 처리 기법을 적용한다. 이후 사용자의 평점 정보를 바탕으로 사용자 간의 유사도를 산출한다. 동시에 사용자의 리뷰 데이터를 TF-IDF, Word2Vec 및 Doc2Vec 기법으로 임베딩 하여 리뷰



〈Figure 1〉 Proposed Methodology Framework

유사도를 산출하고 사용자 평점 유사도와 리뷰 유사도를 결합한 확장된 유사도를 생성한다. 마지막으로 사용자 기반 협업 필터링의 3단계 중에서 두번째 단계인 유사 이웃 집단을 탐색할 때 사용하는 유사도와 마지막 단계인 평점을 예측하여 추천 목록을 생성할 때 사용하는 유사도를 다르게 조합하여 추천 알고리즘을 생성하였다. 이러한 결과를 테스트 세트에 적용하여 성능을 비교하였다.

3.1. 데이터 전처리

데이터 내에 결측치와 중복된 데이터가 존재한다면 편향된 결과를 초래할 수 있기 때문에 가장 먼저 결측치 및 중복 데이터를 제거하였다. 텍스트 데이터 전처리에서는 HTML, 태그, 특수기호 등을 제거하였고, 축약형 형태를 원래 형태로 복구 하였다. 이후에는 토큰화(Tokenization)하여 단어 단위로 분류했다. 분류된 단어에 별도의 정규 표현식을 정의해서 관사, 대명사 등의 불용어를 제거했으며, 모든 단어를 소문자로 변환했다. 최종적으로 표제어 추출(Lemmatization) 기법을 활용하여 리뷰에서 같은 의미를 나타내는 단어를 하나로 통일하였다.

3.2. 평점 데이터를 사용한 사용자 간 유사도 계산

먼저 데이터로부터 사용자-상품 평점 행렬 생성한다. 생성된 행렬의 각 행은 사용자의 정량적인 선호도를 의미한다. 이후 사용자 벡터 간의 코사인 유사도(Cosine Similarity)를 계산하여 사용자간 유사도 행렬을 생성한다. 평점을 사용한 유사도 계산에 대한 자세한 과정은 아래와 같다.

타겟 사용자의 선호도를 예측하기 위한 단계로 사용자 평점을 사용하여 사용자-상품 평점 행렬을 생성해야 한다. 5에 가까울 수록 사용자는 상품을 매우 선호 한다는 뜻이다. 이때 사용자가 평가하지 않은 상품에 대해서는 Nan 값을 부여한다.

다음으로 도출된 사용자-상품 평점 행렬을 기반으로 타겟 사용자와 다른 사용자들 간의 평점 유사도를 산출하였다. 대표적인 점수를 기반으로 하는 유사도 산출 방법으로는 피어슨 상관계수(Pearson correlation coefficient)와 코사인 유사도(Cosine Similarity)가 있다. 피어슨 상관계수는 두 속성간의 상관성을 계산해 -1과 1사이의 값이 산출되며, 1에 가까울 수록 양의 상관관계, -1에 가까울 수록 음의 상관관계가 있으며, 0은

상관관계가 없음을 나타낸다. 코사인 유사도는 두 벡터간 각도의 크기(θ)를 이용하여 유사도를 측정하는 방법이다. 0과 1사이의 값이 산출되며, 두 벡터의 방향이 같을 경우 코사인 값은 1이며, 1에 가까울수록 두 벡터간 유사도가 높다. 본 연구에서는 식(1)과 같이 코사인 유사도를 사용했다.

$$similarity = \cos(\theta) = \sqrt{\frac{r_a \cdot r_u}{\| r_a \| \| r_u \|}} = \frac{\sum_{i=1}^{n}(r_{a,i} \cdot r_{u,i})}{\sqrt{\sum_{i=1}^{n} r_{a,i}^2}\sqrt{\sum_{i=1}^{n} r_{u,i}^2}} \quad (1)$$

$r_{a,i}$는 사용자 a가 상품 i에 부여한 평점이고, $r_{u,i}$는 사용자 u가 상품 i에 부여한 평점이다.

<Table 1>은 사용자-상품 평점 행렬에서 코사인 유사도를 사용한 사용자 간의 유사도 행렬이며 C로 표기한다.

〈Table 1〉 Similarity matrix between users using Cosine

	User0	User1	User2	User3	User4
User0	1.000000	0.180702	0.16966	0.227550	0.094037
User1	0.180702	1.000000	0.164896	0.132997	0.134159
User2	0.169667	0.16489	1.000000	0.112230	0.259890
User3	0.227550	0.132997	0.112230	1.000000	0.179507
User4	0.094037	0.134159	0.259890	0.179507	1.000000

3.3. 리뷰 데이터를 사용한 사용자 간 유사도 계산

앞 단계에서 사용자-상품 평점 행렬을 기반으로 사용자들 간의 평점 유사도를 산출하여 사용자 유사도 행렬을 생성했다면, 다음 단계에서는 사용자의 리뷰 데이터로부터 유사도를 계산하여 리뷰 유사도 행렬을 생성한다. 이때, 사용자 리뷰의 벡터 값을 산출하는 방법으로는 다양한 텍스트 마이닝 기법 중 TF-IDF, Word2Vec 및 Doc2Vec 기법을 사용하였다. 각 리뷰에 대한 벡터 값은 상품에 대한 사용자의 정성적인 선호도를 의미한다.

3.3.1. TF-IDF 기법을 사용한 리뷰 유사도 측정

TF-IDF 기법은 여러 문서에서 특정 단어가 얼마나 중요한 것인가를 나타내는 카운트 기반 방법이다. TF(Term Frequency)는 단어가 한 문서 내에서 얼마나 등장했는지를 나타내는 단어 빈도를 나타내는 값이며, IDF(Inverse Document Frequency)는 전체 문서내에서 해당 단어 등장한 문서의 역수인 역문서 빈도를 나타내는 값이다. 두 사용자가 공통으로 사용한 단어들의 TF-IDF의 가중치 합을 유사도로 사용하는 방법이다. <Table 2>는 리뷰 데이터에 TF-IDF 기법을 사용하여 구한 리뷰 유사도 이며 T로 표기하고, <Table 3>은 리뷰 데이터에 TF-IDF 기법을 사용하여 리뷰 유사도와 평점 유사도를 결합한 확장된 유사도 이며 CT로 표기한다.

〈Table 2〉 Similarity matrix between users using TF-IDF

	User0	User1	User2	User3	User4
User0	1.000000	0.339334	0.355116	0.286543	0.401847
User1	0.339334	1.000000	0.412036	0.379414	0.381515
User2	0.355116	0.412036	1.000000	0.367379	0.422189
User3	0.286543	0.379414	0.367379	1.000000	0.338920
User4	0.401847	0.381515	0.422189	0.338920	1.000000

〈Table 3〉 Extended similarity matrix between users combining C and T

	User0	User1	User2	User3	User4
User0	2.000000	0.520035	0.524783	0.514093	0.495884
User1	0.520035	2.000000	0.576932	0.512410	0.515674
User2	0.524783	0.576932	2.000000	0.479609	0.682080
User3	0.514093	0.512410	0.479609	2.000000	0.518426
User4	0.495884	0.515674	0.682080	0.518426	2.000000

3.3.2. Word2Vec 기법 및 Doc2Vec 기법을 사용한 리뷰 유사도 측정

본 연구에서는 Python Gensim 패키지를 사용하여 Word2Vec, Doc2Vec 기법을 적용해 각 리뷰 별로 벡터 값을 추출했다. Word2Vec은 단어에 의미를 특정 벡터로 표현하는 기법이다. 이때 의미가 비슷한 단어들의 벡터 유사도를 유사하게, 의미가 다른 단어들의 유사도는 멀어지게 학습한다. Word2Vec의 단어 표현 방식에는 CBOW(Continuous Bag of Words)와 Skip-Gram이 존재한다. CBOW는 중간에 있는 단어에 인접한 단어들을 통해 의미를 학습하는 방식이다. 반대로, Skip-Gram은 중간에 있는 단어들을 통해 주변 단어들을 예측하는 방법이다. Doc2Vec 기법은 각 단어의 벡터 값을 구하는 Word2Vec과 달리 문서 자체를 벡터화 하는 기법으로 의미가 비슷한 리뷰 데이터는 유사한 벡터 값을 갖도록 학습한다. 이후 생성된 문서의 벡터들의 코사인 유사도를 분석을 함으로써 두 문서의 유사도 값을 계산한다.

3.4. 이웃 집단 탐색 및 최종 추천 목록 생성

추천 알고리즘은 이웃 사용자의 선호도 정보를 사용하여 타겟 사용자가 구매할 가능성이 높은 상품을 기준으로 추천 목록을 생성하는 것이다. 우선 타겟 사용자와 다른 사용자들 간의 평점 및 리뷰 유사도를 사용하여 선호도과 가장 유사한 이웃 사용자 N명을 도출해낸다. 이후 이웃 사용자의 선호도에 따라 예측 평점을 계산하고 최종 추천 목록을 생성한다. 타겟 사용자의 최종 구매 가능성 점수(PLS, Purchase Likelihood Score)는 다음 식(2)을 통해 계산된다. $w_{a,u}$는 사용자 a와 사용자 u 간의 유사도에 따른 가중치이며, $PLS_{a,i}$는 사용자 a가 상품 i를 구매할 가능성을 의미하며 0과 1사이의 값을 가진다. PLS값이 클수록 상품을 구매할 확률이 높은 것을 의미한다.

$$PLS_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n}[(r_{u,i} - \bar{r}_u) * w_{a,u}]}{\sum_{u=1}^{n} w_{a,u}} \quad (2)$$

본 연구에서는 평점 데이터의 코사인 유사도를 구했을 때 C로 표기하고, 고객이 남긴 리뷰 데이터에 TF-IDF 기법으로 임베딩 하여 유사도를 구했을 때 T로, Word2Vec 기법으로 임베딩 하였을 때 W로, Doc2Vec 기법으로 임베딩 하였을 때 D로 정의한다. 평점 데이터의 코사인 유사도와 리뷰 데이터에 Doc2Vec 기법을 적용하여 구한 유사도를 결합한 확장된 유사도를 정의하기 위해서 CD와 같이 표기한다. 각 단계에서 다르게 쓰인 유사도를 표기하기 위해서 C_CD와 같이 표기 했다. C_CD에서 C는 이웃 집단 탐색 단계에서 사용한 유사도 이고, CD는 추천 목록 생성 단계에서 사용한 유사도를 의미한다.

4. 실험 결과

4.1. 실험 데이터

본 연구에서는 전자상거래 사이트 Amazon에서

〈Table 4〉 Notations used in This Paper

	Symbol	Definition
Single Similarity	C	The similarity using rating value via cosine similarity
	T	The similarity for review contents using TF-IDF
	W	The similarity for review contents using Word2Vec
	D	The similarity for review contents using Doc2Vec
Extended Similarity	CT	The extended similarity combining rating value(C) and TF-IDF(T)
	CW	The extended similarity combining rating value(C) and Word2Vec(W)
	CD	The extended similarity combining rating value(C) and Doc2Vec(D)
Similarity used step by step	C_C	The algorithm of using C on the neighbor group generation step and using C on the recommendation list formation step.
	T_T	The algorithm of using T on the neighbor group generation step and using T on the recommendation list formation step.

1996년 5월부터 2014년 7월까지 수집된 Health and Personal Care 데이터를 사용했다. 데이터에는 사용자 ID, 상품 ID, 리뷰, 유용성, 날짜와 다양한 추가 속성을 포함하고 있다. 사용자 수는 38,609명이며 건강기능식품 상품 수는 18,534개이고 전체 리뷰 수는 346,355개이다. 사용자가 평점과 리뷰를 부여한 상품보다 평점을 부여하지 않은 상품의 수가 훨씬 많기 때문에 데이터 희소성 문제가 발생한다. 데이터의 희소성 문제로 인해 사용자의 선호도 예측이 불가능하거나 추천 시스템의 성능을 측정하는데 한계가 존재할 수 있다. 따라서 본 연구에서는 희소성 문제를 해결하기 위해 기존 연구를 참고하여 적어도 20개 이상의 상품에 대한 평점과 리뷰를 남긴 사용자와 20개 이상의 리뷰를 받은 상품을 연구 대상으로 선정하였다(Al-Bashiri et al., 2018; Elahi et al., 2016). 또한 모델 검증을 위해 전체 데이터의 20%는 테스트 세트로 설정했으며, 일정한 실험 결과를 위해 random state는 42로 설정했다.

4.2. 평가지표

본 연구에서는 추천 시스템 관련 연구에서 가장 많이 사용하는 정확도 지표인 MAE(Mean Absolute Error)와 RMSE(Root Mean Squared Error)를 사용하여 추천 성능을 평가했다. MAE는 실제 정답 값과 예측 값의 차이를 절대값으로 변환한 뒤 평균을 구해서 도출할 수 있으며, 식(3)과 같이 계산한다. 값이 작을 수록 모델의 성능이 좋다. y_i는 실제 값을 의미하고, $\hat{y}_i$는 예측 값이며, N은 전체 평가 대상의 개수이다.

$$MAE = \sqrt{\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{n}} \quad (3)$$

RMSE는 각 예측값과 실제 값의 차이(error)의 제곱(squared)의 평균(mean)의 제곱근(root)이다. 식(4)와 같이 계산한다. 예측이 정확할 수록 y_i와 $\hat{y}_i$의 차이가 작다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

4.3. 실험 파라미터 세팅

본 연구에서는 Python 기반의 Gensim 라이브러리를 사용해 Word2Vec, Doc2Vec 기법을 사용하고, Parameter 는 다음 <Table 5>와 <Table 6>과 같이 설정했다.

Vector_size는 각 단어와 문서가 표현될 벡터의 크기(dimension)이며 100으로 설정했다. 따라서 각 단어와 문서는 100차원의 벡터로 임베딩 하였다. 등장 횟수가 낮은 단어는 전체 문서에서 큰 의미를 갖지 못하기 때문에 min_count를 지정하여 특정 횟수보다 낮은 빈도로 등장한 단어를 학습에서 제외하였다. window 파라미터는 단어 학습 시 반영하는 인접 단어의 개수를 의미한다. 이때 CBOW 방식은 중심 단어를 기준으로 앞, 뒤 5개 단어를 통해 의미를 학습한다. Doc2Vec 모델의 파라미터인 Alpha는 모델 학습 시 학습의 정도를 조절하는 학습률(learning rate)이고 초기 값인 0.025를 설정했다. 또한, Doc2Vec의 epoch 파라미터를 5로 설정하여 전체 리뷰에 대해 5회 학습하도록 세팅하였다.

4.4. 실험 결과

본 연구에서는 추천 서비스에서 사용자의 선호도를 더 정확하게 예측하기 위해 평점과 리뷰 데이터를 모두 고려하여 고도화된 추천 알고리즘을 제안한다. 또한, 협업 필터링의 이웃 집단 탐색 할 때 사용하는 유사도와 추천 목록을 생성할 때 사용하는 유사도 값을 다르게 조합함으로써 유사도를 정의하는 방법 및 적용하는 단계에 따른 추천 성능의 변화를 확인해보았다. 본 연구에서 제안하는 추천 방법론의 성능을 평가하기 위해 평점만을 사용하는 기존 협업 필터링과 비교하였다. 또한 모델의 성능 변화를 파악하기 위해 유사한 이웃의 크기 별로 성능 분석하였다. 이웃 수를 나타내는 N 값을 5 부터 40 까지 5 단위로 설정하고 최종 상품 추천 수는 5 개로 설정하였다.

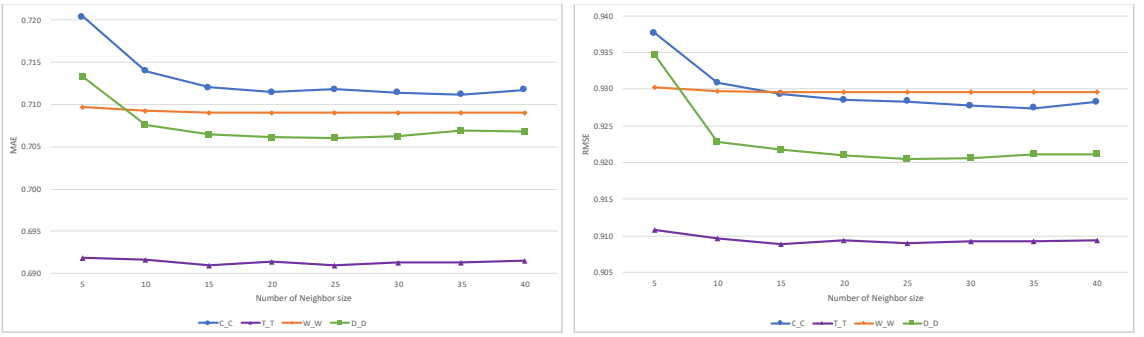
<Figure 2>와 <Figure 3>는 이웃 집단 탐색 및 추천 목록 생성시 같은 유사도를 사용했을 때의 결과이며 이웃 수(N)에 따른 MAE 값과 RMSE 값의 변화이다. MAE 값과 RMSE 값은 작을수록 좋은 성능을 보인다. MAE 결과를 봤을 때, 모든 이웃 사용자의 크기에 관하여 본 연구에서 리뷰의 유사도를 반영한 추천 알고리즘이 기존 평점만 사용하는 유사도인 C_C 추천 알고리즘보다 전체적으로 더 나은 성능을 보여주었다. <Figure 2>은 모든 이웃의 크기에서 T_T 의 성능이 가장 좋은 것을 확인할 수 있었다. 또한 이웃 크기가 5 또는 10 일때는 W_W 의 성능이 더 좋지만, 이웃 크기가 15 보다 클 때는 D_D 의 성능이 W_W 보다 더 나은 것을 확인하였다. RMSE 값

〈Table 5〉 Word2Vec parameter setting

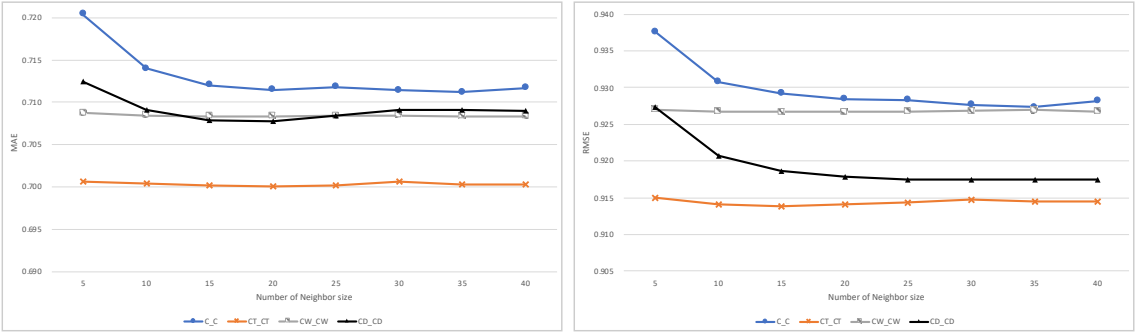
Parameter	Value
Vector_size	100
min_count	5
window	5

〈Table 6〉 Doc2Vec parameter setting

Parameter	Value
Vector_size	100
alpha	0.025
min_count	1
epoch	5



〈Figure 2〉 Recommendation performance comparison when using single similarity



〈Figure 3〉 Recommendation performance comparison when using extended similarity

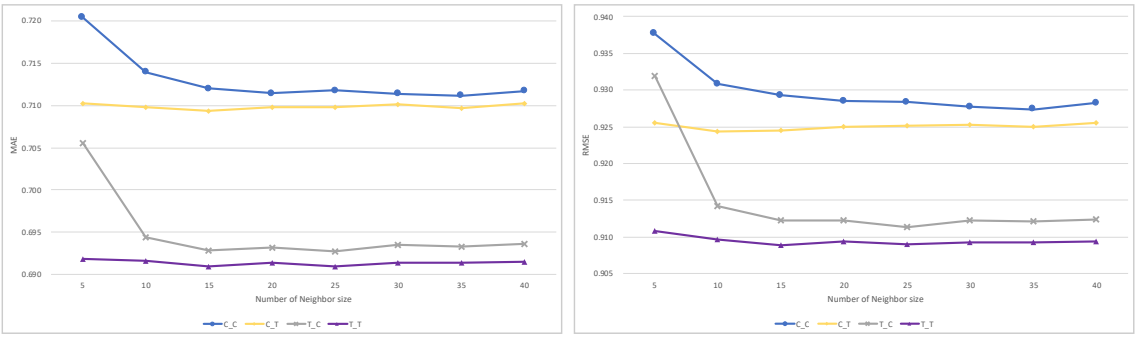
또한 T_T 일 때 성능이 가장 좋은 것을 확인하였다.

<Figure 3>은 확장된 유사도를 사용하여 추천 알고리즘을 생성했을 때의 성능이다. 실험 결과 CT_CT 의 MAE 와 RMSE 값이 가장 작았으며 CT_CT 의 이웃의 크기가 20 일때 추천 성능이 가장 높게 나타났다.

<Figure 2>와 <Figure 3>의 결과에서 T_T와 CT_CT의 성능이 가장 좋은 것을 보아 리뷰 데이터에 TF-IDF 기법을 사용하여 유사도를 구하는 것이 Word2Vec 이나 Doc2Vec 기법보다 효과적인 것을 확인하였다.

TF-IDF 기법을 사용하는 단계에서 따라 변화하는 추천 성능의 결과는 <Figure 4>와 같다. 이웃 집단을 형성할 때 리뷰에 TF-IDF 기법을 적용한 유사도를 사용하는 것이 기존의 코사인 유사도로 평점 유사도를 사용하는 것보다 성능이 좋은 것을 확인하였다. 또한, 기존의 추천 알고리즘 유사도를 계산하는 방법인 C_C의 평균 MAE값은 0.7119이고 T_C를 사용하여 추천을 진행한 추천 알고리즘의 평균 MAE값은 0.694이며 T_T를 사용한 평균 MAE값은 0.691인 것을 보아 두 단계 모두 리뷰에 TF-IDF 기법으로 임베딩 하여 구한 유사도로 추천하는 것의 기존의 추천 알고리즘 보다 성능이 좋은 것을 확인했다.

본 연구의 실험 결과를 통해 유사도를 구할 때 리뷰를 반영하는 것이 기존 단일 평점만을 고려



〈Figure 4〉 Recommendation performance comparison when using TF-IDF

한 방식보다 모든 이웃의 크기에서 향상된 추천 성능을 보였다. 이는 추천 시스템에서 이웃을 선정할 때, 정량적인 선호도만을 고려하는 것보다 정성적인 선호도를 추가로 반영하면 사용자 특성을 추가로 고려할 수 있어 추천 시스템의 성능을 향상할 수 있음을 시사한다. 또한, 협업 필터링의 이웃 형성 및 추천 목록 단계에서 TF-IDF 기법을 사용하여 임베딩 하여 구한 벡터 값의 유사도를 사용할 때 추천에서 가장 효과적임을 알 수 있었다. 이러한 결과는 'Health and Personal care' 도메인의 특성상 전문어를 사용하여 리뷰를 쓰게 되는데 IDF는 소수의 문서에 사용되는 단어는 높은 가중치를 주기 때문에 전문적인 단어들, 즉 문서에서 중요하게 사용되는 단어들이 그렇지 않은 단어들에 비해 상대적으로 더욱 정확하게 임베딩 되었음을 나타낸다.

5. 결론

본 연구에서는 정량적 선호도를 나타내는 사용자 평점과 정성적 선호도를 나타내는 리뷰 데이터를 결합한 확장된 유사도 기반 추천 알고리즘을 제안했다. 비정형 데이터인 리뷰에 다양한 텍스트 마이닝 기법 중 TF-IDF, Word2Vec, Doc2Vec 기법을 적용하여 그 결과를 비교 분석하였다. 제안한 알고리즘의 성능을 검증하기 위해 세계 최대 전자상거래 사이트 Amazon의 건강기능식품 카테고리에 나타난 평점, 리뷰 데이터를 사용하였다. 실험 결과, TF-IDF 기법으로 구한 리뷰 유사도를 사용한 추천 알고리즘의 성능이 가장 뛰어난 것을 확인하였다. 본 연구에서 다양한 텍스트 마이닝 기법을 사용하여 리뷰 유사도를 산출하고, 추천 알고리즘에 확장된 유사도를 적용했다는 점이 기존 연구와 차별화 되는 점이다. 협업 필터링의 3단계 중에서 이웃 집단 형성 및 추천 목록 생성단계에서 사용하는 유사도를 다르게 설정하여 성능의 변화를 확인하였다.

본 연구의 시사점은 다음과 같다. 첫째, 정량적인 데이터만 고려한 추천 시스템보다 정성적인 데이터를 함께 고려한 추천 시스템의 성능이 향상되었음을 확인했다. 둘째, 지금까지 국내외적으로 연구되지 않았던 주제인 건강기능식품 추천 시스템에 대한 실제적인 개발 방법론을 제시함으로써 유관 분야의 연구를 선도하고 후속 연구의 활성화에 기여할 것으로 판단된다. 셋째,

고객은 건강기능식품 추천 목록을 빅데이터 분석방법에 의하여 제공받음으로써 보다 적은 검색 노력과 통신 비용으로 원하는 건강기능식품 관련 컨텐츠를 손쉽게 얻을 수 있을 것으로 기대된다. 이는 고객을 구매까지 연결하는 구매 전환율을 증가시키게 되며, 이는 매출 증대로 이어져 궁극적으로 기업의 수익성을 향상시키는데 기여할 것으로 기대된다.

본 연구의 한계점과 향후 연구 계획은 다음과 같다. 본 연구에서 개발한 추천 방법론과 전통적인 협업 필터링과 성능을 비교했을 때, 본 연구의 추천 성능이 더 우수했지만, 성능차이는 2% 내외에 지나지 않았다. 리뷰데이터를 이용하여 사용자 간의 유사도를 구하는 것이 효과적이라는 것은 밝혔지만, 더욱 효과적인 리뷰데이터를 반영하는 사용자 유사도를 구하는 것은 추후 연구과제이다. 최근 자연어 처리 분야에서 사전 훈련을 한 다음 목적에 맞게 파인 튜닝(Fine-Tuning)하는 연구가 활발하게 이루어지고 있다. 특히, BERT나 트랜스포머가 다양한 분야에서 높은 성능을 나타내어 이를 적용한 모델들이 계속 제안되고 있다. 추후에는 리뷰에 BERT나 트랜스포머와 같은 자연어 처리 기법 적용한 확장된 사용자 유사도를 사용해서 추천의 성능을 검증하고자 한다. 또한, 추후 연구에서는 보다 큰 규모의 실험 데이터를 사용하여 데이터 규모에 따른 추천 시스템 성능의 변화를 검증하는 것도 적절한 추후 연구 주제라고 판단된다.

참고문헌(References)

김병만, 이경, 김시관, 임은기, & 김주연. (2004). 추천시스템을 위한 내용기반 필터링과 협력 필터링의 새로운 결합 기법. *정보과학회논문지: 소프트웨어 및 응용*, *31*(3), 332-342.

김영수, 문현실, & 김재경. (2020). Doc2Vec 모형에 기반한 자기소개서 분류 모형 구축 및 실험. *한국 IT 서비스학회지*, *19*, 103-112.

김유영, & 송민. (2016). 영화 리뷰 감성분석을 위한 텍스트 마이닝 기반 감성 분류기 구축. *지능정보연구*, *22*(3), 71-89.

이승우, 강경모, 이병현, 이청용, & 김재경. (2022). 사용자의 정성적 선호도와 정량적 선호도를 고려하는 추천 시스템 성능 향상에 관한 연구. *경영과학*, *39*(1), 15-27.

이윤주, 이재준, & 안현철. (2021). 고객의 검색 패턴과 상품 상세정보를 활용한 상품 추천 모형의 개선. *한국컴퓨터정보학회논문지*, *26*(1), 265-274.

임민아, 황승연, & 김정준. (2022). 학습률 향상을 위한 딥러닝 기반 맞춤형 문제 추천 알고리즘. *한국인터넷방송통신학회 논문지*, *22*(5), 171-176.

장예화, 이청용, 최일영, & 김재경. (2021). 리뷰 데이터 마이닝을 이용한 하이브리드 추천 시스템 개발: Amazon Kindle Store 데이터 분석사례. *Information Systems Review*, *23*(1), 155-172.

전병국, & 안현철. (2015). 사용자 리뷰 마이닝을 결합한 협업 필터링 시스템: 스마트폰 앱 추천에의 응용. *지능정보연구*, *21*(2), 1-18.

정지수, 지민규, 고명현, 김학동, 임헌영, 이유림, & 김원일. (2019). 문서 유사도를 통한 관련 문서 분류 시스템 연구. *방송공학회논문지*, *24*(1), 77-86.

현지연, 유상이, & 이상용. (2019). 평점과 리뷰 텍스트 감성분석을 결합한 추천시스템 향상 방안 연구. *지능정보연구*, *25*(1), 219-239.

Al-Bashiri, H., Abdulgabber, M. A., Romli, A., & Kahtan, H. (2018). An improved memory-based

collaborative filtering method based on the TOPSIS technique. *PloS one*, *13*(10), e0204434.

Choeh, J. Y., Lee, S. K., & Cho, Y. B. (2013). Applying rating score's reliability of customers to enhance prediction accuracy in recommender system. *The Journal of the Korea Contents Association*, *13*(7), 379-385.

Elahi, M., Ricci, F., & Rubens, N. (2016). A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*, *20*, 29-50.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61-70.

Hearst, M. A. (1999). Untangling text data mining. Proceedings of the 37th Annual meeting of the Association for Computational Linguistics,

Kang, B. (2018). Improving predictive accuracy of user-based collaborative filtering using word2Vec. *Journal of Knowledge Information Technology and Systems*, *13*(1), 169-176.

Kang, B.-S. (2019). A study on the accuracy improvement of movie recommender system using Word2Vec and ensemble convolutional neural networks. *Journal of digital convergence*, *17*(1), 123-130.

Kim, K.-J., & Ahn, H.-C. (2009). User-Item Matrix Reduction Technique for Personalized Recommender Systems. *Journal of Information Technology Applications and Management*, *16*(1), 97-113.

Lee, D., & Hosanagar, K. (2019). How do recommender systems affect sales diversity? A cross-category investigation via randomized field experiment. *Information Systems Research*, *30*(1), 239-259.

Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: a survey. *Decision Support Systems*, *74*, 12-32.

MacKenzie, I., Meyer, C., & Noble, S. (2013). How retailers can keep up with consumers. *McKinsey & Company*, *18*(1).

Marlin, B., Zemel, R. S., Roweis, S., & Slaney, M. (2012). Collaborative filtering and the missing at random assumption. *arXiv preprint arXiv: 1206.5267*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Paradarami, T. K., Bastian, N. D., & Wightman, J. L. (2017). A hybrid recommender system using artificial neural networks. *Expert Systems with Applications*, *83*, 300-313.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. Proceedings of the 1994 ACM conference on Computer supported cooperative work,

Srifi, M., Oussous, A., Ait Lahcen, A., & Mouline, S. (2020). Recommender systems based on collaborative filtering using review texts－a survey. *Information*, *11*(6), 317.

Terzi, M., Rowe, M., Ferrario, M.-A., & Whittle, J. (2014). Text-based user-knn: Measuring user similarity based on text reviews. User Modeling, Adaptation, and Personalization: 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings 22,

Yin, F., Wang, Y., Pan, X., & Su, P. (2018). A Word Vector Based Review Vector Method for Sentiment Analysis of Movie Reviews

Exploring the Applicability of the Movie Reviews. 2018 3rd International Conference on Computational Intelligence and Applications (ICCIA),

Yoon, Y. C., & Lee, J. W. (2018). Movie recommendation using metadata based word2vec algorithm. 2018 International Conference on Platform Technology and Service (PlatCon),

Zhang, Z., Zhang, D., & Lai, J. (2014). urCF: user review enhanced collaborative filtering.

Zheng, L., Noroozi, V., & Yu, P. S. (2017). Joint deep modeling of users and items using reviews for recommendation. Proceedings of the tenth ACM international conference on web search and data mining,

Abstract

A CF-based Health Functional Recommender System using Extended User Similarity Measure

Sein Hong* · Euiju Jeong* · Jaekyeong Kim**

With the recent rapid development of ICT(Information and Communication Technology) and the popularization of digital devices, the size of the online market continues to grow. As a result, we live in a flood of information. Thus, customers are facing information overload problems that require a lot of time and money to select products. Therefore, a personalized recommender system has become an essential methodology to address such issues. Collaborative Filtering(CF) is the most widely used recommender system. Traditional recommender systems mainly utilize quantitative data such as rating values, resulting in poor recommendation accuracy. Quantitative data cannot fully reflect the user's preference. To solve such a problem, studies that reflect qualitative data, such as review contents, are being actively conducted these days. To quantify user review contents, text mining was used in this study. The general CF consists of the following three steps: user-item matrix generation, Top-N neighborhood group search, and Top-K recommendation list generation. In this study, we propose a recommendation algorithm that applies an extended similarity measure, which utilize quantified review contents in addition to user rating values. After calculating review similarity by applying TF-IDF, Word2Vec, and Doc2Vec techniques to review content, extended similarity is created by combining user rating similarity and quantified review contents. To verify this, we used user ratings and review data from the e-commerce site Amazon's "Health and Personal Care". The proposed recommendation model using extended similarity measure showed superior performance to the traditional recommendation model using only user rating value-based similarity measure. In addition, among the various text mining techniques, the similarity obtained using the TF-IDF technique showed the best performance when used in the neighbor group search and recommendation list generation step.

Key Words : Review Text, Collaborative Filtering, TF-IDF, Word2Vec, Doc2Vec

Received : February 10, 2023 Revised : April 17, 2023 Accepted : April 20, 2023
Corresponding Author : Jaekyeong Kim

* Department of Big Data Analytics, Graduate School, KyungHee University
** Corresponding author: Jaekyeong Kim
School of Management & Department of Big Data Analytics, KyungHee University, Seoul 02447,Korea
Tel: +82-02-961-9355, Fax: +82-02-961-9355, E-mail: jaek@khu.ac.kr

저 자 소 개

홍세인
광운대학교 수학과에서 학사학위를 취득하고, 현재 경희대학교 대학원 빅데이터응용학과에서 석사과정에 재학중이다. 주요 관심 연구 분야는 개인화 서비스, 추천 시스템, 자연어 처리 및 딥러닝 등이 있다.

정의주
현재 경희대학교 대학원 빅데이터응용학과에서 석사과정에 재학중이다. 주요 관심 연구 분야는 개인화 추천 서비스, 자연어 처리, 빅데이터 분석, 딥러닝등이 있다.

김재경
서울대학교에서 산업공학학사, 한국과학기술원에서 경영정보시스템 전공으로 석사 및 박사학위를 취득하였으며 현재 경희대학교 경영대학 및 빅데이터응용학과 교수로 재직하고 있다. 미국 미네소타 주립대학교 그리고 텍사스 주립대학교(달라스)에서 교환교수를 역임하였다. 주요 관심 분야로는 개인화 서비스, 추천시스템, 빅데이터 분석 및 딥러닝 등이다. IEEE Transaction on Services Computing, IEEE Transaction on SMC-A, International Journal of Human Computer Studies, International Journal of Information Management, Information and Management, Expert Systems with Applications, Applied Artificial Intelligence, 등 다수의 학술지에 논문을 게재하였다. 현재 4단계 BK21사업 연구단장 (빅데이터 분야) 및 AI 비즈니스 연구센터 센터장을 맡고 있다.