온라인 행동정보를 이용한 협업 필터링

곽지윤, 김가영, 홍다영, 김현희
동덕여자대학교 정보통계학과
jeeyoon3848@gmail.com, ga20171001@gmail.com, dayoung0308@daum.net, heekim@dongduk.ac.kr

BICF : Collaborative Filtering Based on Online Behavior Information

Jee-yoon Kwak, Ga-yeong Kim, Da-young Hong, Hyon Hee Kim
Department of Statistics and Information Science,
Dongduk Women's University

요 약

현재 전자상거래에서 사용되는 협업 필터링은 고객이 입력한 평점 정보를 이용하여 추천 시스템을 구축한다. 하지만 기존의 평점 정보는 고객이 직접 입력해야 하므로 데이터 희소성의 문제가 있고 허위정보를 가려내지 못한다는 문제점 또한 존재한다. 본 논문에서는 기존 평점 정보 기반의 협업 필터링 추천 시스템의 문제점을 해결하기 위해, 온라인 고객 행동 정보를 활용한 협업 필터링 알고리즘을 제안하였다. 실험 결과 본 연구에서 제안한 Collaborative Filtering based on Online Behavior Information (BICF) 알고리즘이 기존의 평점 기반 협업 필터링 방식보다 우수한 성능을 보임을 보여주었다.

1. 서론

온라인 고객 행동 정보란 전자상거래에서 고객이 구매에 이르기까지 정보를 뜻하며, 제품 검색, 장바구니 추가 및 삭제, 구매 시도 등 다양한 온라인 상의 이벤트를 뜻한다. 이러한 온라인 고객 행동 정보는 암시적으로 나타나는 고객의 고유한 소비 패턴을 찾을 수 있다. 온라인 행동 정보를 활용하여 고객의 구매성향을 파악하는 것은 더욱 개인화된 서비스를 제공하고 더 세분화된 추천을 가능하게 한다. 또한 외부적으로 입력을 해야 하는 평점 정보와 달리 전자상거래를 사용하는 모든 사용자의 정보를 활용할 수 있으므로 데이터 희소성의 문제를 해결할 수 있다.

기존의 협업 필터링은 고객이 상품을 구매하고 그에 대한 평점을 매기면 이 평점 정보를 이용해서 상품을 추천한다. 평점 정보를 생성할 권한은 고객에게 있기 때문에 모든 아이템에 대한 평점을 얻을 수 없을 뿐만 아니라 평점 정보를 갖지 않는 신상품이나 인기있는 상품이 아닌 경우는 추천이 되지 않는 문제도 발생할 수 있다. 이런 제품의 경우 고객의 반응을 이끌어내기 위한 많은 노력이 필요하다. 또한 평점 정보는 고객의 주관적인 선호도와 만족도를 나타내고 간혹 거짓 정보를 포함하기 때문에 추천 시스템의 예측 성능을 저하시키는 요인이 된다.

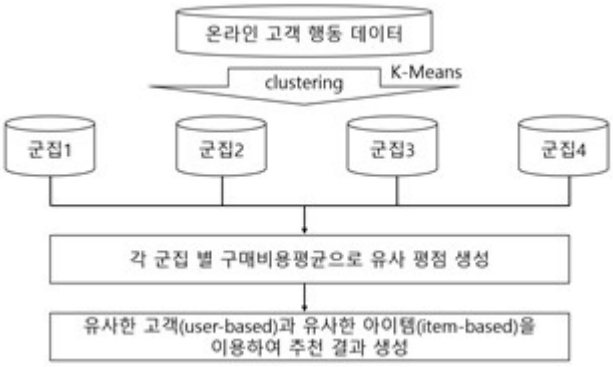
본 논문에서는 온라인 고객 행동 정보를 활용하여 기존의 평점 정보 시스템을 개선한 새로운 추천 시스템인 BICF알고리즘(Behavior Information based Collaborative Filtering)을 제안하였다. 먼저 고객의 온라인 행동을 기반으로 구매 패턴을 파악하고 세분화하기 위해 분석 단위를 세션별로 나누어 전처리를 진행하였다. 다음으로 다양한 온라인 행동 정보를 하나의 압축된 평점 정보로 나타내기 위해 k-means 클러스터링으로 데이터를 군집화하였다. 군집화된 네 가지 소비성향의 구매 금액 평균을 이용하여 유사평점 데이터를 생성하고, 이를 기반으로 협업 필터링을 진행하였다.

협업 필터링 과정에서는 BICF 알고리즘의 성능향상을 증명하기 위해 사용자 기반 협업 필터링과 아이템 기반 협업 필터링 모두 실험에 사용하였다. 마지막으로, 기존의 협업 필터링 시스템과 BICF 알고리즘을 비교하여 모델의 성능을 평가하였고, 그 결과 BICF 알고리즘이 기존의 방법보다 더 높은 정확도를 보임을 보여주었다.

제안하는 BICF 알고리즘은 고객의 온라인 행동 정보를 활용함으로써 기존의 추천 시스템에서 전형적

으로 나타나는 데이터 희소성의 문제와 평점 정보가 충분하지 않은 제품의 추천 문제를 해결할 수 있는 방안으로 활용될 수 있을 것으로 기대된다.

2. BICF 기반 추천 시스템 구조



<그림 1> BICF 기반 추천 시스템

<그림 1>은 제안하는 BICF 기반 추천 시스템의 구조를 보여준다. 먼저 L.point에서 제공하는 익명화된 3,196,362개의 온라인 행동 정보를 구매확정인 22,239개의 데이터로 분석 범위를 줄이고 고객을 세션 단위로 세분화하여 더 세밀하고 개인화된 분석이 가능하게 하였다.

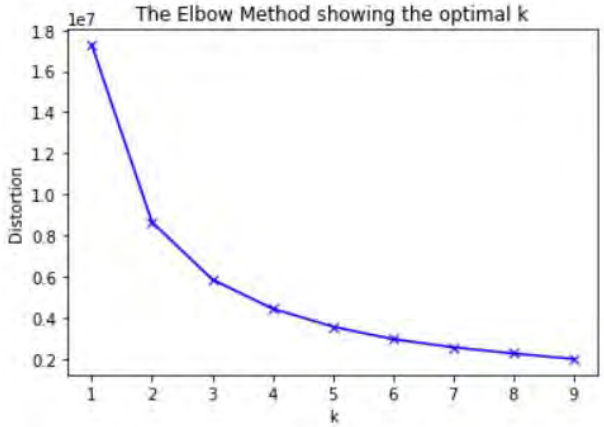
고객 세분화를 위해 K-means 클러스터링을 실시하였는데, 클러스터링 과정으로 들어가기 전 유의미한 온라인 행동 정보 변수인 세션에 머문 시간, 유입 경로, 총 페이지 방문 횟수, 유입 기기, 구매 시간대를 뽑아 전처리를 진행하였다. 명목형 변수는 숫자를 부여하여 one-hot-encoding을 적용하였고, 수치형 변수는 MinMaxScaler을 이용하여 정규화하였다. 전처리된 명목형 변수와 수치형 변수를 하나의 입력으로 처리하기 위해 pipeline을 적용하였다.

K-means 클러스터링을 이용하여 고객을 네 그룹의 세분화된 군집으로 나누고, 군집별 평균 구매비용을 각 제품에 대한 평점으로 생성하였다. 마지막으로 생성된 평점을 활용하여 사용자 기반 추천과 아이템 기반 추천에 적용하였으며, 이후 성능평가를 진행하였다.

3. BICF 알고리즘

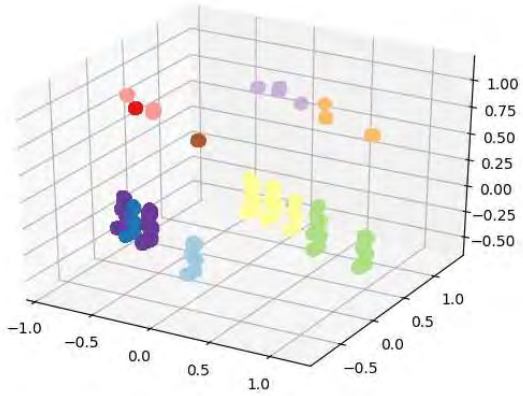
3.1 온라인 행동 정보를 이용한 스코어 생성

여러 개의 변수를 하나의 평점 데이터로 바꾸기 위해 행동 정보를 kmeans로 군집화하여 비슷한 고객 그룹으로 나누었다. 클러스터의 개수를 지정하기 위해 엘보우 기법을 사용하여 최적의 군집 개수를 정하였다. 엘보우 기법이란 클러스터 개수에 따라 SSE(오차 제곱합) 값을 그려주는 함수로 다음과 같은 엘보우 그래프를 얻을 수 있었다. <그림 2> 에서 볼 수 있는 바와 같이 군집의 개수 3과 4에서 좋은 군집을 찾을 수 있다.



<그림 2> 군집 개수를 정하기 위한 엘보우 그래프

최적의 군집 개수를 정하기 위해 군집의 결과를 시각화해 보았다. <그림 3>에서 볼 수 있는 바와 같이 3개의 군집일 경우보다 4개의 군집일 경우가 보다 명확하게 군집이 분리되므로 고객 그룹을 4개의 그룹으로 세분화하였다.



<그림 3> 고객 군집 결과

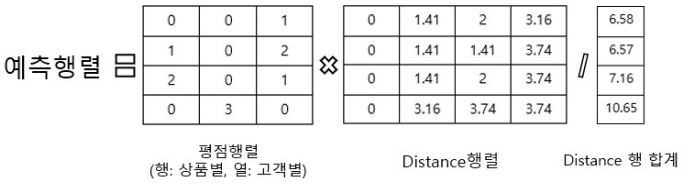
행동 정보로 생성한 클러스터를 협업 필터링에 적용하기 위해 명목형이었던 클러스터를 buy_am(총구매비용)의 평균값으로 오름차순으로 정렬하여 순서형 데이터인 평점 정보로 바꾸었다.

3.2 온라인 행동 정보 기반 협업 필터링 알고리즘

평점 정보를 기반으로 특정 고객과 유사한 성향을 지닌 다른 고객이 산 아이템을 추천하는 사용자 기반(User-based) 협업 필터링과 고객이 관심을 가진 아이템과 비슷한 아이템을 추천하는 아이템 기반(item-based) 협업 필터링을 구현하였다

사용자 기반 협업 필터링의 구현은 다음과 같이 진행된다. 앞서 변형한 평점 정보 데이터를 활용하여 다중 차원 배열의 평가 행렬로 표현하였다. 사용자 기반 협업 필터링의 경우, 아이템에 대한 사용자 평가를 해당 아이템의 다른 모든 사용자 평가에 대한 가중치의 합으로 예측하였다. 알려지지 않은 사용자의 평가를 예측하기 위해서 사용자 기반 협업 필터링은 두 단계를 거친다. 먼저 고객 간의 유사도 행렬을 생성한다. 유사도는 sklearn에서 제공되는 코사인 유사도를 활용했다. 이렇게 형성된 고객 간의 유사도 행렬과 분할된 평점 행렬의 내적을 구하고 평가 수의 데이터를 정규화하여 알려지지 않은 평가를 예측하였다.

아이템 기반 협업 필터링은 앞서 설명한 사용자 기반 협업 필터링과 유사하지만 유사도 계산 부분에 약간의 변형을 주어 두 단계로 진행하였다. 먼저 K-NearestNeighbors 알고리즘을 활용하여 아이템 간의 코사인 유사도를 계산한 후 아이템 유사도 행렬을 생성하였다. 그 다음 평점이 없는 아이템을 예측하기 위해서 사용자 기반 협업 필터링과 같은 방법으로 아이템의 평점을 예측하였다.

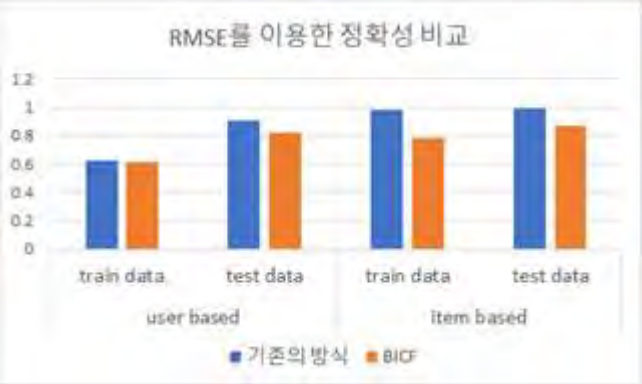


<그림 4> 예측 행렬 계산

4. 성능평가

본 연구의 모델의 정확성을 평가하기 위해 앞의 과정에서 생성된 평가 데이터를 BICF를 사용한 방식을 기존의 방식과 비교하였다. 우선 평가 데이터를 MinMaxScaler를 이용하여 정규화하였다. MinMaxScaler는 각각의 값에 최대값과 최소값의 차이를 나눠주는 방법이다. 이 방법을 사용하면 모든 feature들이 0과 1 사이에 재조정된다.

<그림 5>는 본 논문에서 제시한 방법이 타당한지 증명한 그래프이다. 모델성능 평가의 방법으로 RMSE(Root mean squared error)를 사용하였다. RMSE는 추천 시스템 모델을 평가하는 데에 가장 많이 사용하는 방법 중 하나로 숫자가 작을수록 좋은 모델임을 나타낸다. 그래프를 보면 user-based, item-based 두 가지 방식 모두 BICF의 RMSE가 더 작다. 그러므로 BICF 알고리즘이 기존의 알고리즘에 비해 더 향상된 알고리즘이라고 결론지을 수 있다.



<그림 5> 모델 정확성 평가를 위한 RMSE 비교

4. 결론 및 향후 연구

기존의 평점을 이용한 협업 필터링은 데이터 수집의 속도가 느리고 비용이 많이 든다는 점 등 많은 문제점이 존재한다. 본 연구에서는 이러한 문제를 해결하기 위해 온라인 고객 행동 정보를 활용하여 개선된 추천 시스템을 제안하였다.

본 연구에서 제안한 BICF 알고리즘은 기존보다 정교한 상품 추천을 위해 온라인 고객 행동 정보를 활용하여 군집화를 수행하였고, 군집화된 행동 정보를 활용하여 유사평점을 생성하였다. 그 다음 유사평점을 기반으로 고객와 아이템의 유사도를 계산하여 상품을 구매하지 않은 고객들의 평점을 예측하는 협업 필터링을 적용하였다.

온라인 고객 행동 정보 군집화를 통해 더 개인화된 추천을 가능하게 하였고, 기존의 평점 생성 방식 대신 온라인 행동 정보를 활용한 유사평점을 생성함으로써 평점이 있어야 가능했던 컨텐츠 추천을 더 다양한 분야와 데이터에 적용할 수 있게 되었다. 그 결과, 본 논문에서 제시한 BICF 알고리즘은 기존의 알고리즘에 비해 향상된 정확도를 보였다.

본 연구는 기존 상품 추천 시스템의 한계인 평점

정보의 희소성과 그로 인한 추천 시스템의 성능 하락의 문제를 보완하였다는 의의가 있으나 구매 관련 데이터 이외에 인구 통계 데이터나 검색 키워드 등 이용 가능한 변수를 모두 활용하지 못하였다는 한계점을 가진다. 향후 연구에서는 TF-IDF를 이용한 키워드 분석과 인구 통계 정보를 활용한 세대별 아이템 선호도 분석 등 고객 관련 정보를 충분히 활용한 연구가 필요하다.

참고문헌

[1] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, “Item-Based Collaborative Filtering Recommendation Algorithms”

[2] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen, “Collaborative Filtering Recommender Systems”

[3] 김병희, 장병탁. 2015. 「온라인 공군집화를 이용한 추천 및 콜드스타트 문제 해법 연구」. 『한국지능시스템학회 학술발표 논문집』, 25(1), 41-42.

[4] 김효석, 채선규, 유재성, 배석주. 2019. 「온라인 쇼핑몰 고객의 구매 유형 군집화 및 구매력 평가모형」.『대한산업공학회 춘계공동학술대회 논문집』, 2597-2616.

[5] Suresh K. Gorakala,『Building Recommendation Engines』, 에이콘 출판. 2017

[6] 김경재, 안현철, 「개선된 데이터 마이닝 기술에 의한 웹 기반 지능형 추천 시스템 구축」, 『Journal of Information Technology Applications & Management』 , 41-56 (2005)

데이터 출처: 롯데멤버스, L.pay|L.POINT, 제6회 L.POINT Big Data Competition