사용자 구매 패턴에 기반한 오픈마켓 상품 관련성 분석연구

천세린°, 최대진, 권태경

서울대학교 컴퓨터공학부

A Study on Relations of Goods based on Online Purchase Pattern in Open Market

Selin Chun°, Daejin Choi, Ted “Taekyoung” Kwon

Department of Computer Science and Engineering, Seoul National University

{slchun, djchoi}@mmlab.snu.ac.kr, tkkwon@snu.ac.kr

요 약

오픈마켓은 온라인 이용자가 상품을 등록, 판매 및 구매할 수 있는 플랫폼으로 본 논문에서는 오픈마켓 상에서 온라인 사용자의 상품 구매 패턴에 기반해 상품 간의 연관성을 관측하고 상품 연관성과 관련된 다양한 속성을 조사한다. 결과적으로, 상품의 구매 유사도는 상품의 대분류/소분류가 가장 큰 연관 관계가 있으며 상품명이나 가격도 약한 상관관계가 있음을 관찰하였다. 본 논문에서 분석한 상품 간의 상관관계는 기존의 추천 서비스의 정확도를 향상시키거나 타겟 마케팅 등에 활용될 수 있을 것이라 예상된다.

1. 서론

오픈마켓 (Open Market)은 온라인 상에서 개인 및 기업 사용자가 직접 웹 사이트 상에 상품을 등록하여 판매하고, 구매할 수 있는 플랫폼이다. 오프라인 마켓에서의 구매가 시간 및 공간의 동일성이 필수적인 것과는 달리, 오픈마켓 플랫폼은 온라인으로 구매가 이루어져서 시공간적 제약으로부터 사용자들을 벗어나도록 하였다. 또한, 개인 사용자들이 쉽게 제품을 판매하도록 진입장벽을 낮춤으로써, 온라인 사용자에게 편의성을 제공하게 되었다. 이러한 편의성을 기반으로 오픈마켓의 사용자가 크게 증가하였으며, 관련 시장도 지속적으로 성장하고 있다.

이러한 성장세에 따라, 오픈마켓 플랫폼 상에서 온라인 사용자들의 구매패턴을 분석하고 예측하기 위한 많은 연구가 진행되었으며, 최근 아마존 등 대표적인 오픈마켓 서비스를 중심으로 사용자간 및 상품간 유사도 (Similarity)에 기반한 협업필터링 (Collaborative Filtering) 추천 알고리즘, 타겟 (Target) 광고 등 이러한 분석 결과를 활용하고자 하는 시도가 진행되어 왔다. [1, 2]

그러나 오픈마켓에서 사용자들의 구매 패턴 및 예측에 관련된 많은 시도와 연구가 수행되었음에도 불구하고, 상품간의 구매 연관성을 모델링하거나 다양한 상품의 속성에 대한 연구는 아직 많이 수행되고 있지 못하다. 따라서 본 논문에서는 온라인 사용자들의 오픈마켓 구매 로그를 분석하여, 상품 구매 기록에 기반한 상품 간의 연관성을 모델링하고, 관련 있는 상품들에 대해 구매 금액, 분류 등 상품의 다양한 속성들을 비교함으로써, 상품간의 구매 연관성과 관련된 요소를 찾는 연구를 수행한다. 본 논문에서 조사된 요소는 상품 간의 연관도에 기반한 추천 알고리즘인 협업 필터링 기법 등에 주요 속성으로 이용될 수 있을 것이라 예상되며, 이를 토대로 기존 알고리즘의 성능을 개선할 수 있을 것으로 예상된다.

2. 데이터 설명 및 분석 방법

본 논문에서는 국내 주요 오픈마켓 플랫폼 중 하나의 서비스에 대해서 데이터 분석을 수행하며, 데이터셋은 두 가지로 구성되어있다: 1) 가격, 대/소분류등 상품의 속성, 2) 사용자들의 상품 구매 기록 로그. 상품 속성테이블은 각 상품별 상세 페이지를 크롤링 (Crawling)을 통해 수집하였으며, 사용자 구매 기록 로그는 제 3 자 (third party company)로부터 익명화 (anonymized)된 형태로 제공받아 분석을 수행하였다. 분석에 사용된 구매 기록은 약 200 만건이며, 구매된 상품의 종류는 약 7 만건이다.

위의 데이터셋에 기반하여 분석은 다음과 같은 두 단계로 수행하였다: 1) 사용자 유사 구매 패턴에 기반한 상품 연관성 모델링, 2) 유사 상품 쌍 (Pair)에 대한 구매 금액, 상품명, 대/소분류 등의 상품 속성 비교.

본 논문에서 두 상품의 연관성은 사용자들의 유사 구매 패턴으로부터 계산된다. 즉 상품 i_a, i_b 에 대하여 각각 구매한 사용자의 집합을 U_{i_a}, U_{i_b} 라고 하면, 두 상품의 연관성은 두 집합의 자카드 계수 (Jaccard Coefficient)로 다음과 같이 표현된다.

$$\text{Similartiy}_{(i_a,i_b)} = \frac{U_{i_a} \cap U_{i_b}}{U_{i_a} \cup U_{i_b}}$$

그림 1 에는 상품의 연관성 모델의 예가 나타나 있다. 그림 1 의 상품 i_1의 구매자와 상품 i_2의 구매자는 모두 5 명이고, 이 중 u_1만이 두 상품을 구매하였으므로, 상품 i_1과 i_2의 연관성은 $1/5 = 0.2$ 로 계산된다. 결과적으로 약 51 만개의 관련 쌍이 생성되

었고, 평균 연관성은 0.035 로 계산되었다.

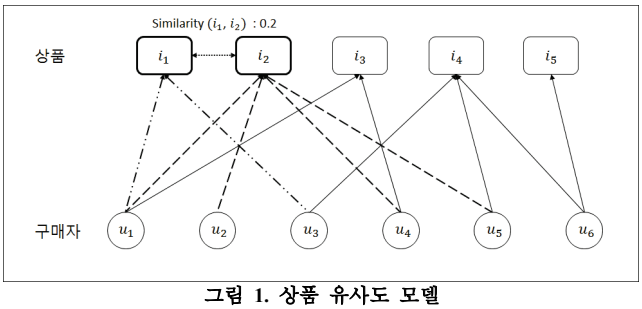


그림 1. 상품 유사도 모델

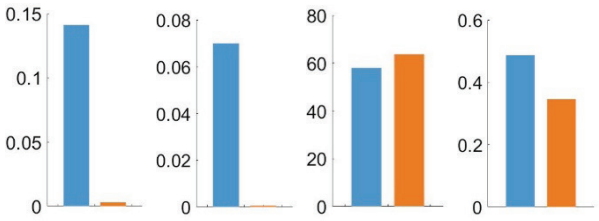
위의 연관성 모델을 기반으로, 이와 연관된 상품 속성을 분석하였다. 유효한 유사 상품 쌍을 대상으로 분석하기 위해 먼저 평균 이상의 값을 가진 쌍들을 선정하고, 상품 쌍의 상품명, 대/소분류, 가격 구간의 연관성을 측정하였다. 상품명의 비교는 레벤슈타인 거리 (Levenshtein distance)로 측정되며, 0 일 경우 완전한 일치를, 값이 클수록 낮은 유사도를 나타낸다. 대/소 분류 유사도는 일치할 경우 1 로, 일치하지 않을 경우를 0 으로 계산하여 평균값을 계산하였고, 따라서 값이 1 에 가까울수록 유사함을 나타낸다. 상품 가격은 그 값이 다양하므로, 상품 가격의 분포 그래프를 참고하여 6 개의 구간 (1 만원 이하, 5 만원 이하, 10 만원 이하, 50 만 이하, 100 만 이하, 100 만원 이상)으로 나누어 같은 구간에 속할 경우 1 을, 다른 그룹에 속할 경우는 0 의 값을 주고 평균값을 통해 유사도를 계산한다.

3. 분석 결과

위의 분석 실험 결과를 정확하게 이해하기 위해, 널 모델 (Null Model) 분석을 수행하고, 널 모델의 결과값과 실제 유사 상품 쌍 간의 값을 비교한다. 널 모델 실험은 전체 상품에 대해서 임의로 두 개의 상품을 선택해서 위와 동일하게 상품명, 대/소분류, 상품 가격에 대해 유사도를 계산하는 것을 한 번의 시행으로 하여, 2000 번 수행하여 평균을 측정함으로써 충분한 시행 횟수를 확보하였다. 이러한 널 모델 결과와의 비교를 통해, 널 모델의 결과보다 높은 유사도를 보일 경우, 상품 간 유사도와 관련있는 요소라고 판단할 수 있으며, 널 모델과 비슷할 경우에는 관련성이 없는 요소로 판단할 수 있다.

그림 2 는 유사 상품 쌍과 널 모델과의 실험결과를 나타낸다. 상품명의 경우 유사 상품 쌍의 거리가 짧은 것으로 보아 (그림 2(c)), 사용자들이 유사한 이름의 상품을 함께 구매한 것으로 나타났다. 하지만, 유사 상품 쌍의 경우 절대적인 값이 높고, 널 모델과의 차이가 작은 것으로 볼 때 (약 5.68), 약한 상관성이 존재하는 것으로 보인다.

상대적으로 대/소분류는 강한 상관관계를 보인다. (그림 2(a), (b)) 대분류의 경우, 유사 상품 쌍에 대해 약 14% 정도가 서로 같은 분류에 속한 상품으로 나타났으며, 유사 상품 쌍 중 약 7%가 동일한 소분류에 속해있는 것으로 나타났다. 이는 널 모델의 0.3%, 0.05%에 비해서 약 50~140 배 정도 더 유사한 수치로, 사용자가 특정 대분류, 소분류의 상품을 구매했을 경우, 높은 확률로 해당 대/소분류의 다른 상품을 구매한다는 것을 의미한다. 유사 상품 쌍의 소분류가 대분류의 일치도보다 낮은 것은, 대분류의 개수가 상대적으로 소분류에 비해 더 적어 유사가능성이 높기 때문인 것으로 보인다.



(a) 대분류 (b) 소분류 (c) 상품명 (d) 가격

그림 2. 모델 간 속성 비교 그래프 (좌측: 유사 쌍, 우측: 널 모델)

가격 구간도 약한 상관성을 보인다. 같은 가격대에 속할 확률이 유사 상품 쌍의 경우, 약 48% 정도에 달하였고, 임의의 쌍은 약 34% 정도로 나타났다.

4. 결론

본 논문에서는 오픈마켓 상에서 사용자의 구매 패턴에 기반해 상품의 연관성을 측정하고 구매 연관성과 관련된 요소가 무엇인지를 연구하였다. 대/소분류 속성의 경우 상품의 구매 연관성과 관련이 가장 높았으며, 상대적으로 상품명, 가격대 요소는 관련성이 존재하지만 그 영향은 약한 것으로 나타났다. 이러한 결과로 볼 때, 오픈마켓 이용자들은 상품의 구매에 있어서 상품명, 분류, 가격대가 같거나 유사한 상품을 구매할 확률이 높고, 그 중에서도 대/소분류가 일치할 확률이 높다고 볼 수 있다. 이러한 관련 요소들은 기존 추천 알고리즘의 주요 속성으로 이용되어 성능향상에 기여할 수 있을 것으로 예상된다.

5. ACKNOWLEDGMENT

이 논문은 2015 년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (B0190-15-2013, 유무선 통합 네트워크에서 접속 방식에 독립적인 차세대 네트워킹 기술 개발)

6. 참고문헌

[1] G.Linden, B. Smith, J. York, “Amazon.com recommendations: item-to-item collaborative filtering” *IEEE Internet Computing,* vol 7, pp.76-80, Jan. 2003.

[2] Keng-Chieh Yang, Chia-Hui Huang, Chen-Wei Tsai, “Applying Reinforcement Theory to Implementing a Retargeting Advertising in the Electronic Commerce Website”, *International Conference on Electronic Commerce*, Aug. 2015.