

협업 필터링 기반 개인화에서의 상품군 중립적 사용자 프로파일링 타당성 검토

Feasibility Study on Cross-Product Category User Profiling in Collaborative Filtering Based Personalization

김중우^a, 박수환^b, 이흥주^c

^a 한양대학교 경영대학 부교수
133-791, 서울시 성동구 행당동 17번지
Tel: +82-2-2290-1067, Fax: +82-2-2292-1169, E-mail:kjw@hanyang.ac.kr

^b 한양대학교 경영대학 석사과정
133-791, 서울시 성동구 행당동 17번지
Tel: +82-2-2290-1067, Fax: +82-2-2292-1169, E-mail:soohwan@digital.re.kr

^c 한국과학기술원 테크노 경영대학원 박사과정
130-722 서울시 동대문구 청량리 2동 207-43
Tel: +82-2-958-3647, Fax: +82-2-958-3604, E-mail:hjlee@kgsn.kaist.ac.kr

Abstract

초기에 하나의 상품 카테고리만을 다루던 전자상거래 사이트들이 브랜드 확립 후에 다른 상품 카테고리까지 확대해 나가는 모습을 많이 보아왔다. 고객이 아직 방문하지 않은 신규 상품 카테고리의 상품에 대하여 기존 상품 카테고리에서 만들어진 사용자 프로파일을 활용하여 개인화된 추천을 할 수 있다면, 고객이 다양한 상품 카테고리를 방문하도록 유도할 수 있을 것이다. 하지만 일반적으로 전자상거래 사이트에서는 상품 카테고리별로 사용자의 선호도를 파악하여 개인화된 추천을 수행하기 때문에, 해당 카테고리 내 상품의 구매나 방문 기록이 없다면 개인화된 추천을 수행하기가 어렵다.

본 논문에서는 협업 필터링을 통해 신규 상품 카테고리 내의 상품을 추천하기 어려운 고객들을 대상으로 기존의 사용자 선호도 데이터를 활용하여 신규 상품 카테고리 내의 상품을 추천하는 방안의 타당성을 살펴보고자 한다. 즉, 기존 사용자의 특정 상품 카테고리 선호도 데이터를 통해 사용자간 유사도를 계산하고, 이를 추천하려는 타 상품 카테고리 내의 상품들에 대한 예측 선호도 계산에 활용 타당성을 살펴본다. 이를 실증적으로 검토하기 위해서, Yes24 사이트의 서적, 음반, DVD 3개의 카테고리 내의 상품을 방문한 웹 패널 데이터를 이용하여 타당성 분석을 수행하였다.

분석 결과, 동일 상품 카테고리 내의 선호도 정보를 가지고 협업 필터링을 수행하는 것보다는

추천 성과가 낮았지만 활용할만한 추천 성과를 보였으며, 활용하는 상품 카테고리 및 예측하는 상품 카테고리별로 추천성과가 상이했다.

Keywords: 상품추천; 협업필터링; 전자상거래; CRM

제1장 서론

고객 맞춤(customization) 또는 개인화 서비스는 인터넷 상점이나 정보 서비스 제공자의 중요한 성공요인으로 인식되고 있다[6,7]. 추천 시스템 중 가장 대표적인 기술인 협업 필터링의 경우, 개인화 추천을 위해서는 해당 사이트에서 고객이 자신의 관심분야를 표시하는 초기 사용자 프로파일의 작성이 필요하고 사용자가 몇몇 특정 상품들에 대한 선호도 점수를 직접 입력해야 한다. 일반적으로 협업 필터링에서 개인화 추천을 위해서 상품 카테고리마다 별도의 개인 프로파일을 생성한다. 이것은 다수 상품 카테고리에 대하여 하나의 사용자 프로파일을 만드는 경우, 상품수의 증대로 인해서 데이터의 희소성(sparsity)이 커지고, 이것은 곧 추천의 성과를 떨어뜨리기 때문이다. 하지만, 상품 카테고리별로 개인 프로파일을 생성하는 경우, 사용자의 번거로움이 발생하고, 따라서 개인화 추천서비스의 사용을 꺼리게 하는 이유가 된다.

전자상거래 사이트의 상품 다변화(diversification)는 일상적인 현상이다. 예를 들어, 대표적인 인터넷 상점인 Amazon의 경우도, 서적에서의 성공을 기초로 하여 음반, DVD, 장난감은

물론, 전자제품에 이르기까지 다양한 상품을 다루는 사이트로 발전하였다[10]. 만일 개인화 상품 추천을 위해서 사용자 프로파일을 관리하는 임의의 전자상거래 사이트가 새로운 상품 카테고리를 다루고자 하는 경우, 기존의 사용자 프로파일이 대상으로 하는 상품 카테고리나 다른 상품 카테고리에 대해서도 기존의 사용자 프로파일을 사용할 수 있는지에 대한 의문이 생긴다. 하지만, 이러한 타 상품 카테고리에 대하여 기존에 구축된 사용자 프로파일을 재사용할 수 있는지에 대한 실증적인 연구는 아직 미미한 형편이다. 따라서 본 논문에서는 하나의 상품 카테고리의 초기 사용자 프로파일이 타 상품 카테고리에 대한 상품 추천에 활용가능한 지를 검토하고자 한다. 구체적으로 다음과 같은 연구 질문들을 고찰해 보고자 한다. 첫째, 한 상품 카테고리에 대한 사용자 프로파일을 타 상품 카테고리의 사용자 프로파일 대신에 사용이 가능한가? 둘째, 동일 상품 카테고리 내의 사용자 프로파일이 타 상품 카테고리의 사용자 프로파일을 사용하는 것보다 추천성고가 더 좋은가? 셋째, 다수의 상품 카테고리의 사용자 프로파일을 통합하여 사용하는 것이 한 상품 카테고리의 사용자 프로파일을 사용하는 것보다 추천성고가 좋은가?

본 연구에서는 Yes24 인터넷 사이트의 3가지 상품 카테고리(서적, 음반, DVD)에 대한 고객 방문 패턴 데이터를 활용하여 분석을 수행하였다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 연구문헌에 대한 검토로 기존의 상품 추천기술과 협업 필터링에 대해서 살펴본다. 3장에서는 실험 설계에 대하여 살펴보고, 4장에서 실험에 대한 결과 및 분석을 제시한다. 마지막 5장에서는 결론을 제시한다.

제2장 상품 추천 기술과 협업 필터링

웹 사이트 개인화란 특정 고객 또는 고객 집단에게 맞춤 웹 경험을 제공함으로써, 고객의 반응성(responsibility)을 높이기 위한 마케팅 전략이다. 특히 상품 추천 기술은 웹 사이트 개인화를 지원하기 위한 방법의 하나로, 고객 프로파일, 거래 데이터, 웹 로그 데이터 등을 토대로 각 고객의 구매성향과 특성을 분석하여 각 개별 고객에게 적합한 정보를 제공하는 기술을 의미한다.

협업 필터링은 대표적인 상품 추천 기술이다. 협업 필터링은 해당 고객과 제품에 대한 선호도가 유사한 고객들의 선호도를 활용하여 추천할 상품을 선정하는데 활용하는 기법이다[8,11,12]. 협업 필터링의 첫 번째 단계는 사용자의 선호도 데이터를 가지고 사용자-상품행렬을 구성하는 것이다. 전통적인 협업 필터링 알고리즘에서의 입력데이터는 사용자 n 명의 상품 m 개에 대한 선호도나 구매이력 자료이다. 이 사용자-상품간의 $n \times m$ 행렬을 S 라고 하면, S_{ik} 는 고객 i 의 상품 k 에 대한 평가 점수이고

\bar{s}_i 는 고객 i 의 평가 점수 평균이다.

협업 필터링의 두 번째 단계는 고객간의 유사도를 계산하는 것이다. 고객간의 유사도를 상관 계수 형태로 구하는 계산 식은 (식1)과 같다. (식1)은 고객 i, j 간의 상관계수 r_{ij} 를 구하는 식으로, 상관계수 r_{ij} 는 두 고객의 선호도가 유사한 경우에는 1에 가까운 값을 가지게 되고, 상반된 선호도를 갖는 경우에는 -1에 가까운 값을 가지게 된다.

$$r_{ij} = \frac{Cov(i,j)}{\delta_i \delta_j} = \frac{\sum_k (S_{ik} - \bar{s}_i)(S_{jk} - \bar{s}_j)}{\sqrt{\sum_k (S_{ik} - \bar{s}_i)^2 \sum_k (S_{jk} - \bar{s}_j)^2}} \quad (\text{식1})$$

(식2)는 사용자-상품간 행렬을 사용자별로 벡터로 표현하여, 두 사용자의 유사도를 두 사용자 벡터간의 코사인 값으로 계산하는 식이다. 이 경우에는 사용자 a, b 를 m 차원의 벡터로 표현하게 된다.

$$\text{similarity}(a,b) = \cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|} \quad (\text{식2})$$

(식1)이나 (식2)를 통해 계산된 사용자들간의 유사도를 통해 유사 사용자(neighbor)를 선정하게 된다. 협업 필터링의 마지막 단계는 상품에 대한 고객의 선호도 점수를 예측하는 것이다. 상품에 대한 고객의 선호도 점수 예측은 다음의 (식3)을 통해서 이루어진다. (식3)은 고객 i 의 상품 k 에 대한 선호도 점수인 P_{ik} 를 예측하는 식으로, $Rater(k)$ 는 상품 k 를 평가한 고객의 집합을 의미한다.

$$P_{ik} = \bar{S} + \frac{\sum_{l \in Rater(k)} (S_{lk} - \bar{s}_l) r_{il}}{\sum_{l \in Rater(k)} |r_{il}|} \quad (\text{식3})$$

협업 필터링의 주요한 이슈로는 (1) 희소성(sparsity), (2) 확장성(scalability), (3) cold start problem이 있다 [1,3]. 협업 필터링의 실제 활용에서 다수 상품 카테고리에 대한 프로파일을 사용하기 보다는 단일 상품 카테고리에 대한 프로파일을 사용하는 것이 더 일반적이다. 이것은 다수 상품 카테고리에 대하여 사용자 프로파일을 구성하는 경우, 사용자 프로파일의 상품 차원의 숫자가 증가하여 사용자 프로파일의 데이터 희소성을 증가시키고, 이것은 곧 추천 성과의 저하로 연결되기 때문이다. 하지만, 이와 같이 상품 카테고리별로 사용자 프로파일을 구성하는 경우, 고객이 평가하지 않은 상품 카테고리의 상품들에 대하여 추천하고자 하는 경우, cold start problem이 야기된다. 본

연구에서는 이러한 cold start problem을 해결하기 위해서, 기존의 타 상품 카테고리 내의 사용자 프로파일을 사용할 수 있는 지에 대한 타당성을 살펴보고자 한다.

제3장 실험방안

3.1 데이터 집합

본 연구에서는 협업 필터링 기반 개인화에서의 상품군 중립적 사용자 프로파일링 타당성 검토를 위해 (주)코리안클릭으로부터 2004년 11월부터 2005년 3월까지 5개월 동안 Yes24 사이트를 방문한 패널들의 클릭스트림 데이터를 구해 활용하였다. 한 고객이 적어도 서적, 음반, DVD 카테고리내의 상품 상세 페이지를 한 번 이상 방문한 패널들의 데이터를 추출하였다. 상품 카테고리간의 차이를 보기 위해서 <표1>과 같이 데이터 집합을 나누었다. 서적∩음반 데이터 집합은 서적과 음반 카테고리내의 상품을 각각 4번이상 방문한 패널들의 데이터이며, 총 92명의 패널들이 속했다. 상품 수는 서적이 2541권이며, 음반은 570개이다. 서적∩DVD 데이터 집합은 서적과 DVD 카테고리내의 상품을 각각 4번이상 방문한 73명의 데이터이며, 상품 수는 서적이 1709권이고 DVD는 677개이다. 음반∩DVD 데이터 집합은 음반과 DVD 카테고리내의 상품을 각각 4번이상 방문한 37명의 데이터이며, 상품 수는 음반이 319개이고 DVD는 515개이다.

<표 1> 데이터 집합

데이터 집합	선정기준	패널의 수	상품 수
서적∩음반	4번이상 방문	92명	서적 2541 음반 570
서적∩DVD	4번이상 방문	73명	서적 1709 DVD 677
음반∩DVD	4번이상 방문	37명	음반 319 DVD 515

3.2 실험 설계

본 연구의 실험은 상품군 중립적 사용자 프로파일링의 타당성을 검토하기 위해 서적, 음반, DVD 3가지 상품 카테고리별로 4가지 실험을 수행하여 추천 성과를 비교하였다.

첫 번째는 추천성과 비교를 위해 사용자 프로파일 없이 무작위로 상품을 추천하는 경우이며, 두 번째는 동일한 상품 카테고리 내의 방문 데이터를 가지고 추천하는 경우이다. 세 번째는 기존의 상품 카테고리에 대한 사용자 프로파일을 가지고 신규 상품 카테고리 내의 상품을 추천하는 경우이며, 네 번째는 복수의 상품 카테고리들에 속한 상품에 대한 방문 데이터를 가지고 사용자 프로파일을 구성하여 하나의 상품 카테고리내의 상품을 추천하는

경우이다. 위 4가지 실험을 서적, 음반, DVD 상품 카테고리별로 수행하여, 결과를 비교 분석해 보았다.

추천 성과 측정을 위해서는 협업 필터링의 성과 측정 방안으로 많이 사용되는 Precision, Recall, F1을 사용하였다. 평가집합에서 예측율이 가장 높은 상품 3개를 선정한 후, 아래 (식4), (식5), (식6)처럼 Precision, Recall, F1을 계산하였다. Precision은 협업 필터링에 의해서 선정된 상품들 중에서 몇 개의 상품을 고객이 실제로 방문하였는지를 나타내며, N은 추천하기 위해 선정된 상품의 개수이고 N_{rs} 은 N에 포함된 것 중에 실제 사용자가 방문한 상품의 수를 나타낸다[9].

$$\text{Precision} \quad Pr = \frac{N_{rs}}{N} \quad (\text{식4})$$

Recall은 고객이 실제로 방문한 상품들 중에서 얼마나 많은 상품이 추천하기 위해 선정 되었는지를 뜻하며, N_r 은 평가집합에 포함된 사용자가 방문한 상품의 총 수를 나타낸다[9].

$$\text{Recall} \quad Re = \frac{N_{rs}}{N_r} \quad (\text{식5})$$

F1은 Precision과 Recall을 동시에 고려한 성과지표이다[9].

$$F1 = \frac{2(Pr \times Re)}{Pr + Re} \quad (\text{식6})$$

각 실험은 다음과 같은 단계를 거쳐 수행되었다.

- 사용자 프로파일이 없는 경우, 무작위로 상품을 추천하는 경우

- ① 실험을 위해서 <표 1>에 표시된 데이터 집합마다 <표 2>와 같이 분할하였다. <표 2>에서 N은 데이터집합의 사용자마다 방문한 상품의 총 수를 나타내고, n은 전체 상품 수를 말한다. 사용자별로 방문한 상품 중에서 무작위로 70%를 선출해서 학습집합에 넣고, 방문하지 않은 상품 중에서도 70%를 무작위로 선출하여 학습집합에 넣는다. 평가집합에는 학습집합에 포함되지 않은 나머지 30%를 넣는다.

<표 2> 데이터 집합 분할

학습집합	평가집합
방문한 data의 수: Gauss(N*0.7)	방문한 data의 수: N - Gauss(N*0.7)
방문안한 data의 수: Gauss(n*0.7 - Gauss(N*0.7))	방문안한 data의 수: Gauss(n*0.3 - N - Gauss(N*0.7))

- ② 평가집합 중에서 무작위로 3개를 선출해서 Precision과 Recall 계산한다.
- ③ 위 과정을 30번 반복수행하여 Precision과 Recall, F1의 평균값을 결과값으로 채택한다. F1의 경우에, ②에서 Precision과 Recall이 모두 0이 되는 경우가 있어서 ③에서 구한 Precision과 Recall의 평균값에서 F1을 계산하였다. 다음에 나오는 실험에서도 같은 방식으로 결과값을 도출하였다.

● 동일한 상품 카테고리 내의 방문 데이터를 가지고 추천하는 경우

- ① <표 2>와 같이 학습집합과 평가집합으로 분할한다.
- ② (식1)을 활용하여 사용자간 유사도를 계산한다.
- ③ ②에서 계산된 사용자간 유사도와 (식 3)을 활용하여 평가집합에 속해 있는 상품에 대한 예측 선호도를 계산한다.
- ④ ③에서 계산된 예측 선호도 중에서 예측치가 높은 3개(Top-3)를 구해서 Precision과 Recall을 계산한다.
- ⑤ 30번 반복수행을 통해 Precision, Recall, F1의 평균값을 결과값으로 채택한다.

● 기존의 상품 카테고리에 대한 사용자 프로파일을 가지고 신규 상품 카테고리 내의 상품을 추천하는 경우

- ① <표 2>와 같이 추천하려는 신규 상품의 사용자 데이터를 학습집합과 평가집합으로 나눈다.
- ② 기존의 상품 카테고리에 대한 사용자 프로파일을 가지고 (식 1)을 통해 사용자간 유사도를 계산한다.
- ③ ②에서 계산된 유사도를 가지고, (식 3)을 통해 평가집합에 속한 상품의 예측 선호도를 계산한다.
- ④ ③에서 계산된 예측 선호도 중에서 예측치가 높은 3개(Top-3)를 구해서 Precision과 Recall을 계산한다.
- ⑤ 30번 반복수행을 통해 Precision, Recall, F1의 평균값을 결과값으로 채택한다.

● 복수의 상품 카테고리들에 속한 상품에 대한 방문 데이터를 가지고 사용자 프로파일을 구성하여 하나의 상품 카테고리 내의 상품을 추천하는 경우

- ① <표 2>와 같이 추천 하려는 상품 카테고리의 방문 데이터를 학습집합과 평가집합으로 분할한다. 그리고 학습집합에 추천하려는 상품 카테고리 이외의 타 상품 카테고리 방문 데이터 전체를 넣는다.

- ② (식 1)을 활용하여 학습집합으로부터 사용자간 유사도를 구한다.
- ③ ②에서 계산된 사용자간 유사도와 (식 3)을 활용하여 평가집합에 속해 있는 상품에 대한 예측 선호도를 계산한다.
- ④ ③에서 계산된 예측 선호도 중에서 예측치가 높은 3개(Top-3)를 구해서 Precision과 Recall을 계산한다.
- ⑤ 30번 반복수행을 통해 Precision, Recall, F1의 평균값을 결과값으로 채택한다.

제4장 실험 결과 및 분석

<표 3>은 서적, 음반, DVD의 3가지 상품 카테고리를 가지고 3.2절에서 언급한 4가지 실험을 수행한 결과이다. 추천 성과는 Precision과 Recall, F1으로 나타내었으며, 각 집합에서 나온 F1 값을 ANOVA 분석과 Duncan Test를 수행하여 각 집합간의 차이를 분석하였다.

<표 3>의 서적∩음반의 데이터 집합에서 4가지 실험의 추천 성과 차이는 one-way ANOVA로 측정하였다. 귀무가설은 4가지 실험의 F1 평균값이 같다는 것이다. 측정결과가 F값이 63.9869이고 P값이 0.000이어서 귀무가설은 기각되었다. 이는 4가지 실험의 추천성과가 차이가 난다는 것으로 볼 수 있다. Duncan Test 결과에서 알 수 있듯이 실험 2가 추천 성과가 가장 좋으며 실험 3과 실험 4는 유사한 값을 나타내며 실험1은 가장 낮은 F1값을 갖는다고 볼 수 있다. <표 3>에 나오는 다른 데이터 집합들 역시 F값과 P값을 볼 때 귀무가설이 기각된 것을 볼 수 있고 추천의 성과 차이가 나는 것을 볼 수 있다.

<표 3>을 보면 무작위로 상품을 추천하는 경우의 추천성과가 제일 낮았고, 동일한 상품 카테고리 내의 방문 데이터를 가지고 추천하는 경우의 추천성과가 제일 높은 것으로 볼 수 있었다. 또한 기존의 상품 카테고리에 대한 사용자 프로파일을 가지고 신규 상품 카테고리 내의 상품을 추천하는 경우가 복수의 상품 카테고리들에 속한 상품에 대한 방문 데이터를 가지고 사용자 프로파일을 구성하여 하나의 상품 카테고리 내의 상품을 추천하는 경우보다 더 좋은 추천성과를 보였다. 그러나 서적∩DVD, 음반∩DVD 데이터 집합에서는 약간 다른 결과가 나타나는 것을 볼 수 있었다.

<표 3>에서 서적∩DVD의 데이터 집합을 가지고 DVD를 추천하는 경우에 동일한 상품 카테고리 내의 방문 데이터를 가지고 추천하는 경우와 복수의 상품 카테고리들에 속한 상품에 대한 방문 데이터를 가지고 사용자 프로파일을 구성하여 하나의 상품 카테고리 내의 상품을 추천하는 경우의 추천성과가 통계적으로 유사한 값이 나타났다.

<표 3>에서 음반∩DVD의 데이터 집합을 가지고

DVD를 추천하는 경우에 기존의 상품 카테고리에 대한 사용자 프로파일을 가지고 신규 상품 카테고리 내의 상품을 추천하는 경우가 동일한 상품 카테고리 내의 방문 데이터를 가지고 추천하는 경우 보다

추천성고가 높은 것을 볼 수 있다. 즉, 음반과 DVD의 사용자 프로파일을 가지고 있을 경우, 음반의 사용자 프로파일을 가지고 DVD를 추천할 경우 높은 추천성고가 나타나는 것을 볼 수 있었다.

<표 3 서적, 음반, DVD 카테고리별 추천성과 리스트>

데이터 집합				추천성과			F-value	Duncan Test
실험	학습집합	평가집합	Precision	Recall	F1			
서적∩음반	1	무작위 선출(서적)		0.01449	0.00136	0.00239	63.9869 (P-value=0.000)	2>3,4>1
	2	서적 (회소수준: 0.9901)	서적	0.44915	0.26111	0.32924		
	3	음반 (회소수준: 0.9871)	서적	0.26282	0.14365	0.18302		
	4	서적+음반 (회소수준: 0.9887)	서적	0.19915	0.12206	0.15005		
	1	무작위 선출(음반)		0.01751	0.00527	0.00799	177.972 (P-value=0.000)	2,3>4>1
	2	음반 (회소수준: 0.9871)	음반	0.27866	0.37325	0.31905		
	3	서적 (회소수준: 0.9901)	음반	0.26389	0.35900	0.30377		
	4	서적+음반 (회소수준: 0.9863)	음반	0.19323	0.31534	0.23904		
서적∩DVD	1	무작위 선출(서적)		0.01750	0.00215	0.00371	40.2968 (P-value=0.000)	2>3>4>1
	2	서적 (회소수준: 0.9877)	서적	0.44566	0.36478	0.39785		
	3	DVD (회소수준: 0.9838)	서적	0.26606	0.18271	0.2125		
	4	서적+DVD (회소수준: 0.9849)	서적	0.15038	0.08158	0.10357		
	1	무작위 선출(DVD)		0.02420	0.00524	0.00842	132.299 (P-value=0.000)	2,4>3>1
	2	DVD (회소수준: 0.9838)	DVD	0.25038	0.30646	0.27521		
	3	서적 (회소수준: 9877)	DVD	0.21202	0.25739	0.23198		
	4	서적+DVD (회소수준: 0.9831)	DVD	0.27092	0.34875	0.30446		
음반∩DVD	1	무작위 선출(음반)		0.03844	0.00999	0.01555	97.4016 (P-value=0.000)	2>3>4>1
	2	음반 (회소수준: 0.9717)	음반	0.54629	0.70745	0.61292		
	3	DVD (회소수준: 0.9730)	음반	0.36937	0.5118	0.42986		
	4	음반+DVD (회소수준: 0.9661)	음반	0.29729	0.38418	0.33412		
	1	무작위 선출(DVD)		0.03514	0.00525	0.00894	139.138 (P-value=0.000)	3>2>4>1
	2	DVD (회소수준: 0.9730)	DVD	0.48468	0.48941	0.48638		
	3	음반 (회소수준: 0.9717)	DVD	0.55976	0.58873	0.57256		
	4	음반+DVD (회소수분: 0.9688)	DVD	0.37867	0.34419	0.35929		

실험을 통해서 얻어진 F1 값의 통계적 분석을 통해서, 본 연구에서 검토하고자 했던 3 가지 연구 질문에 대한 검증 결과는 다음과 같다.

첫째, 한 상품 카테고리에 대한 사용자 프로파일을 타 상품 카테고리의 사용자 프로파일 대신에 사용이 가능한 것으로 나타났다. 즉, 서적 \cap 음반, 서적 \cap DVD, 음반 \cap DVD 3개의 데이터 집합에 대한 6개의 실험군 모두에서, 타 상품 카테고리의 프로파일을 사용한 경우(실험 3)의 성과가 무작위로 상품을 추천하는 경우보다 항상 통계적으로 유의하게 좋은 성과를 나타내었다.

둘째, 일반적으로 동일한 상품 카테고리 내의 사용자 프로파일을 사용하는 것이, 타 상품 카테고리의 사용자 프로파일을 사용하는 것보다 더 추천 성과가 좋을 것으로 예상할 수 있다. 4개의 실험군에서는 이와 일치하는 결과를 얻을 수 있었으나, 2개의 실험군에서는 상이한 결과를 얻을 수 있었다. 즉, 서적 \cap 음반 데이터 집합을 가지고 행한 실험 중 음반의 사용자 프로파일을 가지고 음반을 추천하는 방법(실험 2)과 서적을 가지고 음반을 추천하는 방법(실험 3)간의 성과가 통계적으로 유의하게 다르지 않은 것으로 나타났다. 또한 음반 \cap DVD 데이터 집합을 사용하여 DVD를 추천하는 실험에서는 음반 사용자 프로파일을 가지고 DVD를 추천하는 경우(실험 3)가, DVD 사용자 프로파일을 가지고 DVD를 추천하는 경우(실험 2)보다 통계적으로 유의하게 좋은 성과를 나타내었다.

셋째, 다수의 상품 카테고리의 사용자 프로파일을 통합하여 사용하는 것(실험 4)이 더 많은 정보를 사용하므로, 단일 상품 카테고리의 사용자 프로파일을 사용하는 것보다 더 좋은 성과를 보일 것으로 예상할 수 있다. 하지만 결과는 반대로 한 상품 카테고리를 사용하는 것이 다수 상품 카테고리에 대한 사용자 프로파일을 사용하는 것보다 한 경우(서적 \cap DVD 데이터 집합에서 DVD를 추천하는 경우)만 제외하고는 통계적으로 유의하게 더 좋은 성과를 보여주었다.

제5장 결론

본 연구에서는 협업 필터링을 통해 고객이 아직 방문하지 않은 상품 카테고리 내의 상품을 추천하는 방안의 타당성을 검토하였다. 고객이 방문한 다른 상품 카테고리의 방문 데이터를 활용하여 사용자간의 유사도를 계산하고, 이를 고객이 아직 방문하지 않은 카테고리에 적용하여 협업 필터링을 수행 하는 것이 가능한 지를 검토하였다. 이러한 상품중립적 사용자 프로파일의 타당성을 검토하기 위하여 서적, 음반, DVD 3가지 상품 카테고리에 대한 방문 데이터를 가지고 네 가지 유형의 실험을 수행하여 추천 성과를 비교 하였다. 첫번째는

무작위로 상품을 추천하는 경우이고, 두번째는 동일한 상품 카테고리 내의 방문 데이터를 가지고 추천하는 경우이고, 세번째는 기존의 상품 카테고리에 대한 사용자 프로파일을 가지고 신규 상품 카테고리 내의 상품을 추천하는 경우를 말한다. 마지막으로 복수의 상품 카테고리들에 속한 상품에 대한 방문 데이터를 가지고 사용자 프로파일을 구성하여, 하나의 상품 카테고리 내의 상품을 추천하는 방안이다. 실험 분석 결과, 전반적으로 동일한 상품 카테고리 내의 방문 데이터를 가지고 협업 필터링을 수행하는 것이 가장 추천 성과가 좋은 것으로 나타났고, 무작위로 상품을 추천하는 경우의 추천 성과가 가장 낮게 나타났다. 또한 일부 경우에는 타 상품 카테고리의 사용자 프로파일을 가지고 추천하는 것이 동일 상품 카테고리를 사용하는 경우보다 우수한 경우도 나타났다. 또한 다수 상품 카테고리의 대한 사용자 프로파일을 함께 사용한 경우의 추천 성과가 단일 상품 카테고리의 사용자 프로파일을 사용한 경우보다 도리어 추천 성과가 떨어지는 것을 알 수 있었다.

종합적으로, 한 상품 카테고리의 사용자 프로파일은 타 상품 카테고리의 사용자 프로파일을 대신하여 사용할 때 동일 상품 카테고리의 사용자 프로파일보다는 추천성과가 좀 떨어지기는 하지만, 유용하게 사용할 수 있음을 확인할 수 있었다. 또한, 다수의 상품 카테고리의 사용자 프로파일을 통합하여 사용하는 것보다 단일 상품 카테고리의 사용자 프로파일을 사용하는 것이 추천성과를 높일 수 있는 것으로 확인되었다.

Reference

- [1] 김종우, 배세진, 이흥주, “협업 필터링 기반 개인화 추천에서의 평가자료의 희소 정도의 영향,” 경영정보학연구, 제14권, 제2호, 2004.
- [2] 이재규, 최형림, 김현수, 이경전, 전자상거래원론, 법영사, 1999.
- [3] 조운호, 박수경, 안도현, 김재경, “재구성된 제품 계층도를 이용한 협업 추천 방법론 및 그 평가,” 한국경영과학회지, 제29권, 제2호, 6월, 2004년.
- [4] 황병연, “개선된 추천을 위해 클러스터링을 이용한 협업 필터링 에이전트 시스템의 성능,” 정보처리논문지, 제7권, 제55호, 5월, 2000, pp. 1599-1608.
- [5] Andreas Mild, Thomas Reutterer, “An improved collaborative filtering approach for predicting cross-category purchase based on binary market basket data,” Journal of Retailing and Consumer Services, Vol.10(2003), pp. 123-133.

[6] Allen, Cliff, Deborah Kania, and Yaeckel Beth, Internet World Guide to One-to-One Web Marketing, John Wiley & Sons, Inc., New York, 1998.

[7] Ansari, Asim, Skander Essegaier, and Rajeev Kohli, "Internet Recommendation Systems," Journal of Marketing Research, Vol. XXXVII, August, 2000, pp. 363-375.

[8] Breese, J.S., D. Heckerman and C. Kadie. (1998). "Empirical analysis of predictive algorithms for collaborative filtering," Technical Report, MSR-TR-98-12, Microsoft Research.

[9] Jonathan L. Herlocker and Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, "Evaluating Collaborative Filtering Recommender Systems," ACM Transactions on Information Systems, Vol. 22, No. 1, January 2004.

[10] Krishnamurthy, Sandeep, E-Commerce Management: Text and Cases, South Western, 2003.

[11] Sarwar, B., Karypis, J. Konstan, and J. Riedl. 2001, "Item-Based Collaborative Filtering Recommendation Algorithms," Proceedings of WWW10, Hong Kong, 285-295.

[12] Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. (1994). "GroupLens: An open architecture for collaborative filtering of netnews," proceedings of the ACM 1994 Conference on Computer Supported Cooperative Work, New York, 175-186.