경영정보학연구
제16권 제3호
2006년 9월

전자상거래 개인화 추천을 위한 상품 카테고리 중립적 사용자 프로파일링*

박 수 환**, 김 종 우***, 이 홍 주****, 조 남 재*****

Cross-Product Category User Profiling for E-Commerce Personalized Recommendation

Soo Hwan Park, Jong Woo Kim, Hong Joo Lee, Namjae Cho

Collaborative filtering is one of the popular techniques for personalized recommendation in e-commerce. In collaborative filtering, user profiles are usually managed per product category in order to reduce data sparsity. Product diversification of Internet storefronts and multiple product category sales of e-commerce portals require cross-product category usage of user profiles in order to overcome the cold start problem of collaborative filtering. In this paper, we study the feasibility of cross-product category usage of user profiles, and suggest a method to improve recommendation performance of cross-product category user profiling. First, we investigate whether user profiles on a product category can be used to recommend products in other product categories. Furthermore, a way of utilizing user profiles selectively is suggested to increase recommendation performance of cross-product category user profiling. The feasibility of cross-product category user profiling and the usefulness of the proposed method are tested with real click stream data of an Internet storefront which sells multiple product categories including books, music CDs, and DVDs. The experiment results show that user profiles on a product category can be used to recommend products in other product categories. Also, the selective usage of user profiles based on correlations between subcategories of two product categories provides better performance than the whole usage of user profiles.

Keywords : Personalization, Recommendation Techniques, Collaborative Filtering, Electronic Commerce

---

* 본 연구는 2004년 학술진흥재단의 지원으로 연구되었음(KRF-2004-041-B00169).
** SQ Technology
*** 교신저자, 한양대학교 경영학부
**** Sloan School of Management, Massachusetts Institute of Technology
***** 한양대학교 경영학부

Ⅰ. 서 론

1.1 연구의 배경

고객 맞춤(customization) 또는 개인화 서비스는 인터넷 상점이나 인터넷 정보 서비스 제공자의 중요한 성공요인으로 인식되고 있다[이재규 외, 2002; Allen *et al.*, 1998; Ansari *et al.*, 2000]. 추천 방안 중 가장 대표적인 기술인 협업 필터링의 경우, 해당 사이트에서 고객이 자신의 관심분야를 표시하기 위해 등록 초기에 몇몇 특정 상품들에 대하여 선호도 점수를 직접 입력하여 사용자 프로파일을 생성하거나, 방문 또는 구매 데이터가 존재하는 경우, 이 정보를 활용하여 간접적으로 사용자 프로파일을 생성한다[Herlocker *et al.*, 2004; Sarwar *et al.*, 2001]. 일반적으로 협업 필터링에서 개인화 추천을 위해서 상품 카테고리마다 별도의 개인 프로파일을 생성하는데, 이는 다수 상품 카테고리에 대하여 하나의 사용자 프로파일을 만드는 경우에 상품수의 증대로 인해서 데이터의 희소성(sparsity)이 커져 추천 성과를 떨어뜨리기 때문이다[김종우 외, 2005; Kim and Lee, 2005]. 상품 카테고리 별로 개인 프로파일을 생성하는 경우에는 상품 카테고리별로 사용자의 선호도 정보가 입력되어야 하기 때문에 사용자의 번거로움이 발생하며, 다른 카테고리에서 얻은 정보를 활용하지 못하게 된다.

전자상거래 사이트의 상품 다변화(diversification)는 일상적인 현상이다. 예를 들어, 대표적인 인터넷 상점인 Amazon.com의 경우도, 서적에서의 성공을 기초로 하여 음반, DVD, 장난감은 물론, 전자제품에 이르기까지 다양한 상품을 다루는 사이트로 발전하였다[1][Krishnamurthy, 2003]. 만일 개인화된 상품 추천을 위해서 사용자 프로파일을 관리하는 임의의 전자상거래 사이트가 새로운 상품 카테고리를 다루고자 하는 경우, 신규 상품 카테고리에 대해서도 기존의 사용자 프로파일을 사용할 수 있는지에 대한 의문이 생길 수 있다. 또한 다양한 상품 카테고리를 다루는 인터넷 쇼핑몰에서, 어떤 고객이 특정 상품 카테고리 페이지들만 방문하다가 처음으로 다른 상품 카테고리 페이지를 방문했다고 가정해보자. 이 경우, 이 인터넷 쇼핑몰이 협업 필터링을 통해서 개인화 추천을 제공하고 이를 위해서 사용자 프로파일을 생성하고 있다고 가정하면, 기존의 상품 카테고리의 상품에 대한 사용자 프로파일을 가지고, 역시 처음으로 방문한 상품 카테고리 내의 상품 추천이 가능한지도 인터넷 쇼핑몰 관리자에게는 중요한 질문 사항이 될 수 있다. 하지만, 타 상품 카테고리에 대하여 기존에 구축된 사용자 프로파일이 사용 가능한 지에 대한 실증적인 연구는 아직 미미한 형편이다.

본 논문에서는 협업 필터링에서 다른 상품 카테고리에 대한 사용자 프로파일을 가지고 타 상품 카테고리 내의 상품 추천을 하는 것의 가능성을 검토하기로 한다. 실제 인터넷 쇼핑몰의 방문 데이터를 사용하여 실증적으로 사용 가능성을 검증하였으며, 또한 타 상품 카테고리 내의 상품을 추천하기 위해 기존 사용자 프로파일을 보다 효과적으로 활용하기 위한 방안을 제시하도록 한다. 본 논문의 구성은 다음과 같다. Ⅱ장에서는 전자상거래 개인화 추천을 위한 협업 필터링에 대하여 검토한다. Ⅲ장에서는 타 상품 카테고리 내의 상품을 추천하기 위한 기존 사용자 프로파일의 사용 가능성을 검토하기 위한 실험 설계와 실험 결과를 제시하도록 한다. Ⅳ장에서는 기존 사용자 프로파일을 사용하여 타 상품 카테고리 내의 상품을 효과적으로 추천하기 위한 방안을 제시하고, 이 방안의 유용성을 실증적으로 검토하도록 한다. Ⅴ장에서는 결론과 추후 연구 이슈를 제시하도록 한다.

1) 우리 나라의 Yes24.com, Aladdin.com 같은 업체들도 확장 범위가 다르기는 하지만 서적에서 CD, DVD 등으로의 확장이 이루어 졌다.

Ⅱ. 관련 연구

2.1 개인화 추천을 위한 협업 필터링

상품 추천 기술은 이비즈니스에서 개인화, 일대일 마케팅을 구현하기 위한 유용한 도구 중의 하나이다[Ansari *et al.*, 2000; Dragon, 1997; Schafer *et al.*, 2001]. 상품 추천 기술은 고객의 인구통계학적인 정보, 방문 로그 정보, 구매 이력 등을 활용하여 특정 고객에 대해 적합한 광고를 선정하거나 상품을 추천하기 위해서 사용된다[Herlocker *et al.*, 2004; Sarwar *et al.*, 2001]. 상품 추천 기술을 통해 고객에게는 자신에게 적합한 상품을 찾기 위해서 여러 상품의 페이지를 살펴보아야 하는 정보의 과부하(information overload)를 줄여줄 수 있으며, 인터넷 상점이나 인터넷 쇼핑몰에게 있어서는 단순한 방문자를 구매 고객으로 바꿀 수 있으며, 교차 판매를 증가시키고 고객의 충성도를 높일 수 있다. 대표적인 상품 추천 기술에는 협업 필터링, 내용 기반 추천, 규칙 기반 기법 등이 있다[박상규 외, 2003; 조윤호 외, 2004; Breese *et al.*, 1998; Cho *et al.*, 2002; Kim *et al.*, 2001; Resnick *et al.*, 1994; Weng and Liu, 2004].

협업 필터링은 해당 고객과 제품에 대한 선호도가 유사한 고객들의 선호도를 활용하여 추천할 상품을 선정하는 기법이다[황병연, 2000; Breese *et al.*, 1998; Gupta *et al.*, 1999; Konstan *et al.*, 1997; Sarwar *et al.*, 2001; Shardanand and Maes, 1995; Resnick *et al.*, 1994]. 협업 필터링의 첫 번째 단계는 고객의 선호도 데이터를 가지고 고객-상품행렬을 구성하는 것이다. 전통적인 협업 필터링 알고리즘에서의 입력데이터는 n명의 고객의 m개의 상품에 대한 선호도나 방문/구매이력 자료이다. 이 고객-상품간의 $n \times m$ 행렬을 S라고 하면, S_{ik}는 고객 i의 상품 k에 대한 평가 점수 또는 구매횟수, 해당 페이지 방문 여부이고, $\overline{S_i}$는 고객 i가 평가하거나 구매 혹은 방문한 상품들에 대한 평균 선호도 점수, 평균 구매 빈도 혹은 평균 방문 빈도 값이다.

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

$$r_{ij} = \frac{Cov(i,j)}{\delta_i \delta_j} = \frac{\sum_k (S_{ik} - \overline{S_i})(S_{jk} - \overline{S_j})}{\sqrt{\sum_k (S_{ik} - \overline{S_i})^2} \sqrt{\sum_k (S_{jk} - \overline{S_j})^2}} \quad (1)$$

식 (2)는 고객-상품간 행렬을 사용자 별로 벡터로 표현하여, 두 고객의 유사도를 두 고객 벡터간의 코사인 값으로 계산하는 식이다. 이 경우에는 고객 i,j를 m 차원의 벡터로 표현하게 된다.

$$similarity(\vec{i},\vec{j}) = \cos(\vec{i},\vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\| \times \|\vec{j}\|} \quad (2)$$

Herlocker *et al.*[1999]에 따르면 피어슨 상관계수나 스피어먼(Spearman) 상관계수를 활용하는 것이 코사인, 엔트로피(Entropy), mean-squared difference 알고리즘들을 활용하는 것 보다 높은 추천 성과를 보이는 것으로 조사되었다[Herlocker *et al.*, 1999].

협업 필터링의 세 번째 단계는 추천 대안 상품들에 대한 고객의 선호도 점수를 예측하는 것이다. 상품에 대한 고객의 선호도 점수 예측은 다음의 식 (3)을 통해서 이루어진다. 식 (3)은 고객 i의 상품 k에 대한 선호도 점수인 P_{ik}를 예측하는 식으로, $Rater(k)$는 상품 k를 이미 평가하거나 방문한 고객의 집합을 의미한다. 식 (3)에서 r_{ij}는 고객 i,j간의 유사도로 식 (1) 또는 식 (2)를 통해서 계산된 것을 사용한다.

$$P_{ik} = \overline{S} + \frac{\sum_{l \in Rater(k)} (S_{lk} - \overline{S_l}) r_{il}}{\sum_{l \in Rater(k)} |r_{il}|} \quad (3)$$

협업 필터링의 실제 활용에 있어서 기초 데이터 역할을 하는 고객-상품행렬은 상품 카테고리별로 생성된다. 이것은 협업 필터링의 희소성(sparsity) 이슈와 밀접한 관계를 갖는다[김종우 외, 2005; Kim and Lee, 2005]. 다시 말해서 고객-상품행렬이 다수 상품 카테고리를 통합해서 생성된다면, 상품 수의 증가로 인해서, 고객-상품행렬의 희소성이 증가하게 된다. 따라서, 이것은 협업 필터링의 추천 효과성을 저하시키게 된다. 데이터의 희소성을 감소시키기 위한 노력의 일환으로 고객-상품 행렬의 차원(dimension)을 줄이기 위한 연구들도 진행되었다[Goldberg *et al.*, 2001; Sarwar *et al.*, 2001; Ungar and Foster, 1998]. 차원축소 방안은 클러스터링[Ungar and Foster, 1998], Singular Value Decomposition [Sarwar *et al.*, 2001], 주성분분석(Principal component analysis) [Goldberg *et al.*, 2001], aggregation[Adomavicius *et al.*, 2005] 등이 활용된다. Eigenstate 알고리즘[Goldberg *et al.*, 2001]은 주성분분석에 의한 차원축소를 통해 추천성과의 감소없이 개인화 추천을 위한 계산 시간의 감소를 가져왔다. Adomavicius *et al.*[2005]는 고객-상품 행렬의 2차원적인 데이터 뿐만이 아니라 추가적인 상황정보(contextual information)을 활용하여 추천 성과를 높일 수 있는 방안을 제시하였으며, 이는 전체 상황정보들 중에서 추천 성과에 기여도가 높은 추가 상황정보들 만을 적용하는 것이다. 또한 희소성을 줄이기 위해서 다양한 차원의 정보들을 aggregation하여 추천에 적용하는 방안도 제시하였으며, 동일 분류에 속한 상품들의 평균 선호도를 해당 분류에 대한 선호도로 삼는 방안을 활용하였다. 예를 들어 개별 상품에 대한 선호도가 아니고 상품의 카테고리에 따른 선호도를 계산하여 개인화 추천에 활용하는 것이다.

또한 협업 필터링 추천에 있어서 문제점 중에 하나는 "cold start problem"이다[Rashid *et al.*, 2002]. 이것은 거의 상품에 대해서 선호도를 표시하지 않은 고객이나 처음 방문한 고객에게는 협업 필터링을 통해 추천할 수 없다는 것을 의미한다. 이러한 문제점은 상품에 대해서도 성립한다. 즉, 모든 고객이 선호도를 표시하지 않은 상품은 추천 대상 상품으로 활용할 수 없게 된다. "cold start problem"을 해결하기 위해서, 많은 전자상거래 사이트들은 인구통계학적인 정보를 사용한 고객 군집화를 사용하거나 인기 순서(popularity)에 근거한 추천을 제공하고 있다[Rashid *et al.*, 2002]. 만일 고객-상품행렬이 상품 카테고리 별로 생성되어 있고, 한 고객이 해당 카테고리에 처음으로 방문한 경우에는 다른 상품 카테고리에 대한 해당 고객의 선호도 정보가 있음에도 불구하고 "cold start problem"이 발생해서 협업 필터링을 활용한 추천을 할 수 없다. 따라서 이러한 문제를 해결하기 위해서, 본 연구에서는 다른 상품 카테고리에 대한 고객 프로파일의 활용을 통해 타 상품카테고리에 대해 추천하는 것이 가능한지에 대해 살펴보기로 한다.

2.2 연구 질문

타 상품 카테고리의 고객 프로파일을 활용은 인터넷 상점이 신규 상품을 다루려고 하는 경우와 전자상거래 포털 등의 다수의 상품 카테고리를 다루는 인터넷 사이트에서 한 고객이 처음으로 특정 상품 카테고리를 방문한 경우에 필요하게 된다. 구체적인 본 논문의 연구 질문은 다음과 같다.

[연구 질문 1]

한 상품 카테고리의 고객 프로파일은 다른 상품 카테고리의 상품을 추천하는데 활용할 수 있는가?

[연구 질문 2]

만일 그것이 가능하다면, 추천 성과를 향상시키는 방안은 무엇인가?

첫 번째 연구 질문은 타 상품 카테고리에 대한 상품추천에 기존 카테고리의 고객 프로파일이

활용 가능한 지(feasibility)에 대한 것이다. 이 첫 번째 연구 질문을 협업 필터링 내에서 정의하면 다음과 같다.

[연구 질문 1']

만일 한 상품 카테고리에 대한 고객-상품행렬을 가지고 있는 경우, 이 고객-상품행렬을 통해서 계산되어지는 고객간 선호도의 유사도는 다른 상품 카테고리의 상품을 추천하는데 활용할 수 있는가?

Ⅲ. 활용가능성에 대한 실험(실험-1)

3.1 데이터 집합과 성과 척도

본 연구에서는 협업 필터링 기반 개인화에서 기존 상품 카테고리의 사용자 프로파일이 타 상품 카테고리의 사용자 프로파일 대신 사용하는 것이 가능한 지를 알아보기 위하여 (주)코리안클릭으로부터 2004년 11월부터 2005년 3월달까지 5개월 동안 Yes24 사이트를 방문한 패널들의 클릭스트림 데이터를 제공받아 활용하였다. 회사의 분석목적에 따라 수집되는 초기데이터는 다양한 항목을 가지고 있다. 기본적으로는 <표 1>의 데이터 예제와 같이 패널들의 아이디, 방문사이트명, 방문페이지주소와 방문한 시각이 포함되어 있다. 이외에 방문페이지를 요청한 페이지의 URL, CACHE의 활용여부, 인터넷 서비스 프로바이더 등과 같이 분석목적에 활용할 수 있는 정보들을 함께 축적하고 있다. <표 1>과 같은 클릭스트림 예제에서 상품 페이지에 해당하는 주소를 해당사이트의 구조에서 파악하여, 방문페이지 주소에서 상품페이지를 방문한 경우를 찾았으며, 방문한 상품의 ID를 추출하였다. <표 1>에서 3, 4행에 있는 방문주소가 상품 페이지를 방문한 주소이며, goodsNo 뒤에 번호(3행의 경우, '204300')가 책의 고유한 상품 ID이며 CategoryNumber 뒤의 번호가 책이 속한 카테고리 정보이다. 실험을 위해서, 해당 상품페이지 주소를 찾아내어 상품 페이지 방문데이터로 표시하며 goodsNo이후의 상품 번호를 잘라내어 방문한 상품의 ID로 기록하고 CategoryNumber이후의 정보를 책의 카테고리 정보로 기록하였다. 실험에 활용된 정제된 데이터의 예제는 <표 2>와 같다.

<표 2> 정제된 고객의 클릭스트림 데이터 예제

고객ID	사이트명	상품번호	카테고리번호
100034	Yes24	204300	001001005004
100034	Yes24	186149	002001004001004
100035	Yes24	2130342	001001010005001
100035	Yes24	2126467	001001025008006
100035	Yes24	1461474	001001020

주) 위 예제의 고객ID는 예시를 위해 사용된 것입니다.

<표 1> 고객의 클릭스트림 데이터 예제

고객ID	사이트명	방문 웹 페이지 주소	방문시각
100034	Yes24	http://www.yes24.com/Main/default.aspx	2005-02-12 18:05:52
100034	Yes24	http://www.yes24.com/searchCenter/searchResult.aspx?keywordAd=&qdomain=%C0%FC%C3%BC&query=%C8%A3%B9%D0%B9%E7%C0%C7+%C6%C4%BC%F6%B2%DB	2005-02-12 18:06:31
100034	Yes24	http://www.yes24.com/Goods/FTGoodsView.aspx?goodsNo=204300&CategoryNumber=001001005004	2005-02-12 18:07:04
100034	Yes24	http://www.yes24.com/Goods/FTGoodsView.aspx?goodsNo=186149&CategoryNumber=002001004001004	2005-02-12 18:07:39
100034	Yes24	http://www.yes24.com/Order/FTCartList01.aspx	2005-02-12 18:07:58

주) 위 예제의 고객ID는 예시를 위해 사용된 것입니다.

상품 카테고리간의 차이를 보기 위해서 <표 3>과 같이 3개의 데이터 집합, 서적∩음반, 서적∩DVD, 음반∩DVD를 생성하였다. 예를 들어 서적∩음반 데이터 집합은 서적과 음반 카테고리 내의 상품을 각각 4번 이상 방문한 고객들의 데이터이며, 총 92명이 이 집합에 해당한다. 서적∩음반 데이터 집합과 관련된 상품 수는 서적이 2541권이며 음반은 570개이다. 마찬가지로 서적∩DVD 데이터 집합은 서적과 DVD 카테고리 내의 상품을 각각 4번 이상 방문한 73명의 데이터이며, 상품 수는 서적이 1709권이고 DVD는 677개이다. 음반∩DVD 데이터 집합은 음반과 DVD 카테고리 내의 상품을 각각 4번 이상 방문한 37명의 데이터이며, 상품 수는 음반이 319개이고 DVD는 515개이다. 실제로 3개의 데이터 집합은 이진(binary) 행렬 형태로, 특정 고객이 특정 상품 페이지를 방문한 경우는 1, 그렇지 않은 경우는 0을 가진다.

<표 3> 실험 데이터 집합

데이터 집합	선정기준	고객 수	상품 수
서적∩음반	서적, 음반 각각 4번 이상 방문	92명	서적 2541 음반 570
서적∩DVD	서적, DVD 각각 4번 이상 방문	73명	서적 1709 DVD 677
음반∩DVD	음반, DVD 각각 4번 이상 방문	37명	음반 319 DVD 515

협업 필터링의 성과 측정 방안으로 많이 사용되는 방안들은 크게 통계적 방안과 의사결정 방안으로 나눌 수 있다[Herlocker *et al.*, 1999]. 통계적 방안은 실제 사용자 입력한 선호도와 추천 알고리즘에 의해 계산된 예측치간의 차이를 가지고 통계적인 방안으로 계산하며 mean absolute error(MAE)가 대표적이다. 의사결정 방안은 추천된 상품들이 실제로 좋은 선호도를 얻었는지, 구매 혹은 이용되었는지를 가지고 추천 성과를 계산하는 방안이며, Precision, Recall, F1, Receiver operating characteristic(ROC) curve 등이 활용된다. O'mahony *et al.*[2004]은 협업 추천 알고리즘을 정확성뿐만이 아니라 안정성(stability)도 고려한 성과지표들을 제시하고 다양한 추천 알고리즘의 성과를 평가하였다. 본 연구에서는 추천 성과 측정을 위해 Precision, Recall, F1 지표를 사용하였으며 평가집합(실험에서 평가집합의 선정은 차후에 설명하도록 한다)에 속한 상품 중 예측선호도 값이 가장 높은 상품 3개를 선정한 후, 아래 식 (4)~식 (6)처럼 Precision, Recall, F1을 계산하였다. Precision은 협업 필터링에 의해서 선택된 상품들 중에서 몇 개의 상품을 고객이 실제로 방문하였는지를 나타내며, N은 추천하기 위해 선정된 상품의 개수이고 N_{rv}은 N에 포함된 것 중에 실제 사용자가 방문한 상품의 수를 나타낸다[Herlocker *et al.*, 2004].

$$\text{Precision} \qquad \Pr = \frac{N_{rv}}{N} \tag{4}$$

Recall은 고객이 실제로 방문한 상품들 중에서 얼마나 많은 상품이 추천에 포함 되었는지를 나타낸다. 식 (5)에서 N_v은 평가집합에 포함된 고객이 방문한 상품의 총 수를 나타낸다.

$$\text{Recall} \qquad \text{Re} = \frac{N_{rv}}{N_v} \tag{5}$$

일반적으로 Precision과 Recall은 trade-off 관계를 가지고 있다. 식 (6)의 F1은 Precision과 Recall을 동시에 고려한 성과지표이다.

$$\text{F1} = \frac{2(\Pr \times \text{Re})}{\Pr + \text{Re}} \tag{6}$$

3.2 실험 방법

첫 번째 실험은 기존의 상품 카테고리에 대한 사용자 프로파일을 타 상품 카테고리에 대한 사용자 프로파일 대신 사용이 가능한 지를 살펴보기 위한 실험으로 다음 네 가지 경우를 비교하였다.

(1) 사용자 프로파일이 없이 무작위로 상품을 추천하는 경우
(2) 사용자 프로파일이 없이 인기도에 기반하여 상품을 추천하는 경우
(3) 동일한 상품 카테고리에 대한 사용자 프로파일을 가지고 추천하는 경우
(4) 기존의 상품 카테고리에 대한 사용자 프로파일을 가지고 타 상품 카테고리 내의 상품을 추천하는 경우

(1) 사용자 프로파일이 없이 무작위로 상품을 추천하는 경우

상품 카테고리 A에 대하여 추천하려고 하는 경우, 무작위 추천의 상세한 절차는 다음과 같다.

① 데이터 집합(서적∩음반, 서적∩DVD, 또는 음반∩DVD)으로부터, 상품 카테고리 A의 방문여부 데이터만을 추출한다.
② 선택한 데이터를 임의로 학습 데이터 집합과 테스트 데이터 집합으로 구분한다. 즉, 각 고객별로 방문 데이터(고객-상품행렬에서 1 값을 가지는 원소) 중 70%를 학습 데이터 집합에 할당하고, 비방문 데이터(고객-상품 행렬에서 0값을 가지는 원소) 중 70%를 학습 데이터 집합에 할당한다. 나머지 30%에 해당하는 방문 데이터와 비방문 데이터는 테스트 데이터 집합에 할당한다.
③ 각 고객별로 테스트 데이터 집합에 해당하는 상품 중에 3개를 임의로 선택한다.
④ 선택된 상품에 대한 예측값과 실제값을 기준으로 각 고객별로 precision, recall, F1값을 계산한다.
⑤ 고객들의 precision, recall, F1값들의 평균을 계산한다.
⑥ 앞에 제시된 ②단계부터 ⑤단계까지를 30번 반복한다.
⑦ 30번 반복해서 얻어진 precision, recall, F1값의 평균값을 계산하고, 이것을 무작위 선정의 추천 성과로 한다.

(2) 사용자 프로파일이 없이 인기도에 기반하여 상품을 추천하는 경우

상품 카테고리 A에 대하여 추천하려고 하는 경우, 인기도에 기반한 추천의 상세한 절차는 다음과 같다.

① 데이터 집합(서적∩음반, 서적∩DVD, 또는 음반∩DVD)으로부터, 상품 카테고리 A의 방문여부 데이터만을 추출한다.
② 선택된 데이터 중에서 70%의 사용자 데이터를 무작위로 선정하여 학습집합으로 넣고, 나머지 30%의 사용자 데이터를 평가집합으로 선정한다.
③ 학습집합에 속한 사용자들이 상품별로 방문한 횟수를 인기도로 정의하여, 가장 많이 방문한 상품 3개를 선택하며, 이를 평가집합에 속한 사용자들에게 추천한다.
④ 평가집합에 속한 사용자별로 ③단계에서 추천된 상품을 실제로 방문하였었는지를 기준으로 Precision, Recall과 F1 값을 식 (4)~식 (6)을 활용하여 계산한다.
⑤ 평가집합에 속한 사용자들의 평균 Precision, Recall과 F1값을 계산한다.
⑥ 위의 단계 ②부터 ⑤까지를 30회 반복수행한다.
⑦ 각 반복에서 계산된 precision, recall, 그리고 F1값의 평균을 구하고, 이 값들을 최종적인 추천 성과로 한다.

(3) 동일한 상품 카테고리에 대한 사용자 프로파일을 가지고 추천하는 경우

동일한 상품 카테고리의 고객 프로파일을 활용하여 추천하는 상세한 실험 과정은 다음과 같다. 상품 카테고리 A의 고객 프로파일을 사용하여 상품 카테고리 A에 대하여 추천을 한다고 가정하면 실험 과정은 다음과 같다.

① 데이터 집합으로부터 상품 카테고리 A의 방문여부 데이터만을 선택한다.
② 각 고객에 대하여 임의로 70%의 방문 데이터(고객-상품행렬에서 1 값을 가지는 원소)를 학습 데이터 집합에 할당하고, 70%의 비방문 데이터(고객-상품행렬에서 0 값을 가지는 원소)를 학습 데이터 집합에 할당한다. 나머지 각각 30%에 해당하는 데이터를 테스트 데이터 집합에 할당한다.
③ ②단계에서 생성된 학습 데이터 집합을 사용하여, 식 (1)을 적용하여 고객간 선호도를 계산한다.
④ 각 고객별로 테스트 데이터 집합에 해당하는 상품에 대하여 식 (3)을 이용하여 예상 선호도를 계산한다.
⑤ 각 고객별로 ④단계에서 계산된 예측 선호도가 높은 3개의 상품을 추천 대상으로 선정한다. 추천 대상인 3개 상품의 예측 선호도와 실제 방문여부를 이용하여 precision, recall, F1을 계산한다.
⑥ 고객별로 계산된 precision, recall, F1값의 평균을 구한다.
⑦ ②단계부터 ⑥단계까지를 30회 반복한다.
⑧ 각 반복에서 계산된 precision, recall, 그리고 F1값의 평균을 구하고, 이 값들을 최종적인 추천 성과로 한다.

(4) 기존의 상품 카테고리에 대한 사용자 프로파일을 가지고 타 상품 카테고리 내의 상품을 추천하는 경우

타 상품 카테고리에 대한 사용자 프로파일을 이용하여 추천하는 경우에 대한 실험 절차를 상세히 설명하도록 한다. 먼저 데이터 집합 $A \cap B$에서, 상품 카테고리 A에 대한 고객 프로파일을 활용하여 상품 카테고리 B에 속한 상품을 추천하고자 한다고 하면 상세한 실험 절차는 다음과 같다.

① 각 고객에 대하여 70%의 상품 카테고리 A의 방문 데이터를 임의로 학습 데이터 집합에 할당한다. 상품 카테고리 A의 비방문 데이터의 70%를 임의로 학습 데이터 집합에 할당한다.
② 각 고객에 대하여 70%의 상품 카테고리 B의 방문 데이터를 임의로 학습 데이터 집합에 할당한다. 상품 카테고리 B의 비방문 데이터의 70%를 임의로 학습 데이터 집합에 할당한다.
③ ②단계에서 학습 데이터 집합에 포함되지 않은 상품 카테고리 B의 방문 데이터와 비방문 데이터를 테스트 집합에 할당한다.
④ ①에 의해서 생성된 상품 카테고리 A에 대한 학습 데이터 집합을 사용하여, 식 (1)을 적용하여 고객간 유사도를 계산한다.
⑤ ④단계에서 계산된 고객간 유사도와 ②단계에 생성된 상품 카테고리 B에 대한 학습 데이터 집합을 사용하여, 각 고객별로 테스트 데이터 집합(③단계에서 생성)에 해당하는 상품에 대하여 식 (3)을 이용하여 예상 선호도를 계산하고, 이를 기초로 계산된 예측 선호도가 가장 높은 상품 3개를 추천 대상으로 결정한다.
⑥ ⑤단계에서 선정된 3개 상품의 예측 선호도와 실제 방문여부를 이용하여 precision, recall, F1을 계산한다.
⑦ 고객별로 계산된 precision, recall, F1값의 평균을 구한다.
⑧ ①단계부터 ⑦단계까지를 30회 반복한다.
⑨ 각 반복에서 계산된 precision, recall, F1 값의 평균을 구하고, 이 값들을 최종적인 추천 성과로 한다.

실험 절차에서 보면 타 상품 카테고리(실험 절차에서는 A)를 활용하여 추천을 하는 경우에도, 추천하려는 상품 카테고리(실험 절차에서는 B)에 대한 선호도 정보가 필요하다. 이것은 협업 필터링이 가지고 있는 "cold start problem"으로 인해

서 전혀 선호도 정보가 제공되지 않은 상품에 대한 추천이 불가능하기 때문이다. 서론에서 타 상품 카테고리를 위한 사용자 프로파일 활용이 필요한 경우를 2가지로 제시하였다. 첫 번째는 전자상거래 사이트가 상품 다변화를 시도하는 경우로, 신규 상품 카테고리를 다루고자 하는 경우이다. 하지만 신규 상품 카테고리의 상품들에 대한 사용자 선호도 정보가 전혀 없다면 협업 필터링을 사용할 수 없다. 물론 이 경우에는 기존 상품 카테고리에 대한 사용자 프로파일을 사용할 수 없다. 하지만, 일부 사용자들이 신규 상품 카테고리의 상품에 대한 선호도 정보를 제공한다면, 이 정보를 사용하여 신규 상품 카테고리의 상품에 대한 선호도 정보를 제공하지 않은 대다수의 고객에게는 기존 상품 카테고리의 사용자 프로파일을 기반으로, 사용자 유사도를 계산하여 신규 상품 카테고리의 상품에 대한 추천이 가능할 것이다. 즉, 상품 다변화의 경우는 기존 상품 카테고리의 사용자 프로파일을 활용하는 방안을 전체 고객에게 다 적용할 수는 없지만, 여전히 대다수의 고객에게는 유용하다는 것이다. 타 상품 카테고리를 위한 사용자 프로파일 활용이 필요한 두 번째 경우는 다수 상품 카테고리를 다루는 인터넷 쇼핑몰에서, 어떤 고객이 특정 상품 카테고리 페이지들만을 방문하다가 처음으로 다른 상품 카테고리 페이지를 방문하는 경우이다. 이 경우에는 앞에서 제시한 절차와 같이 기존 상품 카테고리의 사용자 프로파일을 활용하여 고객간 유사도를 계산하고, 이것을 바탕으로 처음 방문한 상품 카테고리에 속한 상품에 대한 추천이 가능하다.

3.3 실험 결과

첫 번째 실험 결과를 <표 4>와 <그림 1>에 표시하였다. 세 개의 데이터 집합에 대하여, 1.*.1는 무작위 추천, 1.*.2는 인기도에 기반한 추천, 1.*.3은 동일한 상품 카테고리 사용자 프로파일을 활용한 추천, 1.*.4는 타 상품 카테고리 사용자 프로파일을 활용한 추천의 결과이다. 실험방안들의 추천 성과치인 F1값이 실험방안별로 차이가 있는 지를 통계적으로 검증하기 위해서Duncan's 테스트를 수행하였다. 이 검증 결과는 <표 4>의 마지막 열에 표시되어 있다.

서적∩음반 데이터 집합에서, 서적을 추천하려 할 때 동일 상품 카테고리 사용자 프로파일을 사용하는 경우(실험 1.1.3)가 가장 좋은 성과를 보였다. 타 상품 카테고리 사용자 프로파일을 사용하는 경우(실험 1.1.4)는 실험 1.1.3보다는 낮았지만, 무작위 추천(실험 1.1.1)과 인기도에 의한 추천(실험 1.1.2) 보다는 좋은 성과를 보였다.

서적∩음반 데이터 집합에서 음반을 추천하는 경우도 역시 동일 상품 카테고리의 사용자 프로파일을 활용해서 추천한 경우가 가장 좋은 성과를 보였다. 하지만, 이 경우에는 타 상품 카테고리 사용자 프로파일을 활용한 경우와 성과 차이가 크지 않았다. Duncan's 테스트 결과 타 상품 카테고리의 사용자 프로파일을 활용한 경우와 동일 상품 카테고리의 사용자 프로파일을 활용한 경우의 F1값의 차이가 유의하게 다르지 않은 것으로 나타났다.

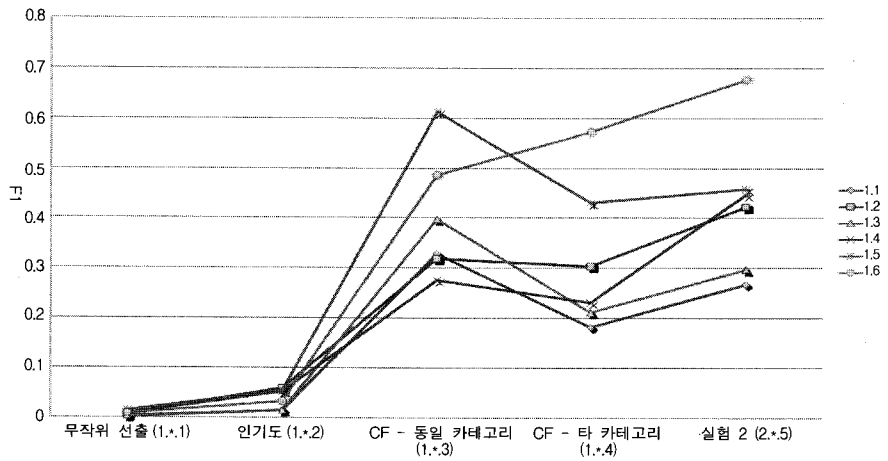
서적∩DVD 데이터 집합에서 서적을 추천했을 때와 DVD를 추천하는 두 경우 모두 타 상품 카테고리의 사용자 프로파일을 사용한 경우가 동일 상품 카테고리의 사용자 프로파일을 사용한 경우보다는 성과가 낮고, 무작위 추천과 인기도에 의한 추천보다는 성과가 높았다.

음반∩DVD 데이터 집합에서의 음반을 추천하는 경우에는 위의 실험결과들과 유사한 추천 성과의 차이를 보이고 있다. 하지만, 음반∩DVD 데이터 집합에서의 DVD 추천 성과는 위의 실험결과들과 상이함을 알 수 있다. 타 상품 카테고리의 사용자 프로파일을 사용한 경우(실험 1.6.4)가 무작위 추천의 경우(실험 1.6.1)나 인기도에 의한 추천(실험 1.6.2)는 물론이고, 동일 상품 카테고리의 사용자 프로파일을 사용한 경우(실험 1.6.3)보다도 더 좋은 추천 성과를 보이고 있다.

<표 4> 첫 번째 실험 결과

데이터 집합				추천성과			Duncan's
	실험	학습집합	평가집합	Precision	Recall	F1	Test
서적 ∩ 음반	1.1.1	Random Selection (Book)		0.01449	0.00136	0.00239	3 > 4 > 2, 1
	1.1.2	Popular (Book)		0.18636	0.07612	0.01453	
	1.1.3	Book	Book	0.44915	0.26111	0.32924	
	1.1.4	Music	Book	**0.26282**	**0.14365**	**0.18302**	
	1.2.1	Random Selection (Music)		0.01751	0.00527	0.00799	3, 4 > 2 > 1
	1.2.2	Popular (Music)		0.15050	0.03803	0.05975	
	1.2.3	Music	Music	0.27866	0.37325	0.31905	
	1.2.4	Book	Music	**0.26389**	**0.35900**	**0.30377**	
서적 ∩ DVD	1.3.1	Random Selection (Book)		0.01750	0.00215	0.00371	3 > 4 > 2, 1
	1.3.2	Popular (Book)		0.21441	0.00813	0.01550	
	1.3.3	Book	Book	0.44566	0.36478	0.39785	
	1.3.4	DVD	Book	**0.26606**	**0.18271**	**0.21250**	
	1.4.1	Random Selection (DVD)		0.02420	0.00524	0.00842	3 > 4 > 2 > 1
	1.4.2	Popular (DVD)		0.17837	0.03108	0.05205	
	1.4.3	DVD	DVD	0.25038	0.30646	0.27521	
	1.4.4	Book	DVD	**0.21202**	**0.25739**	**0.23198**	
음반 ∩ DVD	1.5.1	Random Selection (Music)		0.03844	0.00999	0.01555	3 > 4 > 2 > 1
	1.5.2	Popular (Music)		0.19017	0.03191	0.05364	
	1.5.3	Music	Music	0.54629	0.70745	0.61292	
	1.5.4	DVD	Music	**0.36937**	**0.51180**	**0.42986**	
	1.6.1	Random Selection (DVD)		0.03514	0.00525	0.00894	4 > 3 > 2, 1
	1.6.2	Popular (DVD)		0.19012	0.01904	0.03241	
	1.6.3	DVD	DVD	0.48468	0.48941	0.48638	
	1.6.4	Music	DVD	**0.55976**	**0.58873**	**0.57256**	

주) *** 유의수준 0.01



주) 1.1 = 서적∩음반 데이터 집합에서 서적을 추천할 경우, 1.2 = 서적∩음반 데이터 집합에서 음반을 추천할 경우
1.3 = 서적∩DVD 데이터 집합에서 서적을 추천할 경우, 1.4 = 서적∩DVD 데이터 집합에서 DVD를 추천할 경우
1.5 = 음반∩DVD 데이터 집합에서 음반을 추천할 경우, 1.6 = 음반∩DVD 데이터 집합에서 DVD를 추천할 경우

<그림 1> 실험방안들의 추천 성과(F1값)

Ⅳ. 타 상품 카테고리의 사용자 프로파일의 선택적 활용 대한 실험 (실험-2)

4.1 타 상품 카테고리의 사용자 프로파일의 선택적 활용

첫 번째 실험 결과를 통해서 타 상품 카테고리의 사용자 프로파일을 통해 다른 상품 카테고리의 상품을 추천하는 것의 유용성을 알 수 있었다. 두 번째 실험은 타 상품 카테고리의 사용자 프로파일을 보다 효과적으로 활용하기 위한 방안에 대한 검토이다. 두 번째 실험에 근거가 된 직관은 다음과 같다. <표 5>를 보면 실험 사이트의 24개의 서적 하위 카테고리의 목록이 보인다. 목록을 살펴보면, 전집, 대학교재, 학습참고서, 수업서/자격증 등 몇몇 하위 카테고리의 경우에는 음반이나 DVD의 선호도와는 무관한 것으로 생각된다. 이러한 직관에 근거하여 몇 개의 하위 카테고리에 해당하는 사용자 프로파일 정보를 삭제한 후, Ⅲ장의 실험을 수행해본 결과 Ⅲ장에서 제시된 타 상품 카테고리 사용자 프로파일을 모두 사용한 경우보다 추천 성과를 높일 수 있음을 확인할 수 있었다. 하지만, 이러한 방안의 가장 큰 문제점은 선호도와 관련성이 적은 하위 카테고리를 어떻게 객관적으로 결정하느냐 하는 것이다. 이를 위해서 본 연구에서는 데이터마이닝 기법 중 하나인 의사결정나무 추론에서 의사결정나무의 가지 분기 속성을 결정할 수 있는 방법 중에 하나인 Chi-Square 값을 응용하여 하위 카테고리들을 평가하는 식을 고안하였다[Berry and Linoff, 2004]. 의사결정나무 추론 알고리즘에서는 궁극적으로 관심이 있는 변수(일반적으로 목표 변수라고 부름)의 값을 가장 잘 구분할 수 있는 속성(설명 변수라고 부름)들을 차례로 결정하여 의사결정나무를 생성해간다. 예를 들어, 어떤 테스트에 합격 여부를 설명하는 변수로 성별과 출신지역을 고려한다고 하자. 만일 전체 테스트 대상의 수가 100명이고, 이중 남녀가 60명, 40명, 출신지역 A, B 소속이 각각 70명, 30명이고, 두 속성과 목표 변수간의 분할표(contingency table)가 각각 <표 6>의 (a), (b)와 같다고 하자. 이 경우 Chi-Square 값을 구하는 식은 다음과 같다.

$$\chi = \sum_{i,j} \frac{(F_{ij} - E_{ij})^2}{E_{ij}} \qquad (7)$$

<표 5> 서적 하위 카테고리와 음반에 대한 χ^2 값

카테고리	χ^2	순위
수험서/자격증*	1.59193	1
잡지	1.20718	2
국어와 외국어*	1.10078	3
유아	1.00709	4
예술/대중문화	0.86938	5
어린이	0.86136	6
문학	0.72691	7
가정과 생활	0.70131	8
인물	0.63701	9
자연과학*	0.59749	10
비즈니스와 경제	0.53998	11
학습참고서*	0.53508	12
여행과 지리	0.53189	13
만화	0.48439	14
역사와 문화	0.48420	15
사회	0.48014	16
청소년	0.46882	17
대학교재+*	0.45308	18
인문+	0.37268	19
종교+	0.26709	20
전집+*	0.25681	21
자기관리+*	0.23569	22
건강/취미/실용+	0.15684	23
컴퓨터와 인터넷+	0.13603	24

주) * 직관에 의해 삭제된 하위 카테고리.
+ χ^2값에 의해서 삭제된 하위 카테고리.

<표 6> 테스트 합격 여부에 대한 분할표 예제

(a) 성별-합격 여부에 대한 분할표

합격여부(목표 변수) / 성별(설명변수)	합격	불합격	합계
남	40	20	60
여	10	30	40
합계	50	50	100

(b) 출신지역-합격 여부에 대한 분할표

합격여부(목표 변수) / 출신지역(설명변수)	합격	불합격	합계
A	35	35	70
B	15	15	30
합계	50	50	100

식 (7)에서 F_{ij}는 분할표의 (i,j) 셀의 실제 관측값이고, E_{ij}는 (i,j) 셀의 기대값으로, $(F_{i.}\times F_{.j})/F_{..}$으로 계산된다. 여기서, $F_{i.}$은 분할표 상의 i행 빈도수의 합을, $F_{.j}$은 분할표 상의 j열 빈도수의 합을, $F_{..}$은 분할표 상의 모두 셀의 빈도수의 합을 의미한다. 예를 들어, <표 6>(a)에서 F_{11}은 40, E_{11}은 30(=(60×50)/100)이 된다. <표 6>의 두 속성, 성별과 출신 지역에 대한 Chi-Square 값 $\chi_{\text{성별}}=16.7$, $\chi_{\text{출신지역}}=0$이 되어 성별이 출신 지역보다는 테스트의 합격여부를 예측하는데 더 유용함을 알 수 있다.

본 연구에서 이러한 Chi-Square 값의 개념을 응용하여, 서적, 음반, DVD 하위 카테고리 별로 서적, 음반, DVD 하위 카테고리 방문과의 상관관계를 파악하기 위한 척도를 설계하였다. 예를 들면, 서적 하위 카테고리들 중 어떤 하위 카테고리의 방문여부가 음반 하위 카테고리들의 방문여부를 예측하는데 더 유용한 지를 살펴보기 위하여, 특정 서적 하위 카테고리(b_i)의 방문여부를 설명변수로, 개별 음반 하위 카테고리(m_j)의 방문여부를 목표변수로 하여, 음반 카테고리 m_j의 방문평균빈도와 서적 하위 카테고리 b_i를 방문한 고객과 방문하지 않은 고객들의 음반 하위 카테고리 m_j의 방문빈도가 평균빈도와 차이가 많은지를 계산하여 방문한 고객 집합의 숫자와 방문하지 않은 고객의 숫자로 가중치를 두어, 이것을 전체 음반 카테고리 집합 M에 대하여 합하였다. 즉, 특정 서적 하위 카테고리(b_i)의 음반 카테고리 집합 M에 대한 Chi-Square 값, $\chi_{b_i\to M}{}^2$는 식 (8)과 같이 정의 된다.

$$\chi_{b_i\to M}{}^2=\sum_{m_j\in M}\left\{\frac{|C_{b_i}|}{|C|}\times(V(m_j|C_{b_i})-V(m_j))^2+\frac{|\overline{C_{b_i}}|}{|C|}\times(V(m_j|\overline{C_{b_i}})-V(m_j))^2\right\} \quad (8)$$

식 (8)에서 C는 전체 고객의 집합이고, C_{b_i}는 특정 서적 하위 카테고리 b_i를 방문한 고객집합, $\overline{C}_{b_i}$는 하위 카테고리 b_i를 방문하지 않은 고객 집합이다. 또한, $M=\{m_1,.....,m_{N_m}\}$은 음반 하위 카테고리의 집합이고, $V(m_j)$는 음반 하위 카테고리 m_j의 평균 방문 빈도, $V(m_j|C_{b_i})$는 C_{b_i}에 속한 고객의 음반 하위 카테고리 m_j의 평균 방문 빈도를 의미하고, $V(m_j|\overline{C_{b_i}})$는 $\overline{C}_{b_i}$에 속한 고객의 m_j의 평균 방문 빈도를 의미한다. 식 (8)의 $\chi_{b_i\to M}{}^2$은 b_i서적 하위 카테고리의 방문자와 비 방문자 간의 음반 하위 카테고리에 대한 방문 횟수가 차이가 나는 정도를 수치화한 것이다. 따라서 $\chi_{b_i\to M}{}^2$ 값이 큰 서적 하위 카테고리일수록 음반 하위 카테고리들과 높은 선호도 상관관계를 갖는다고 말할 수 있다. <표 5>는 음반에 대한 서적 하위 카테고리들의 Chi-square 값을 보여준다. 이 값을 기준으로 하위 7개의 서적 하위 카테고리에 대한 사용자 프로파일 정보를 삭제하여, 3장의 세 번째 실험과 동일한 절차를 거쳐서 추천 성과를 구하였다. 마찬가지 방법으로 서적∩DVD 데이터 집합에서 DVD를 추천하는 경우에도 Chi-square 값을 계산하여 7개 서적 하위 카테고리에 대한 사용자 프로파일 정보를 삭제하

고, 타 상품 카테고리 사용자 프로파일의 추천 성과를 구하였다. 음반의 경우(서적∩음반 데이터 집합에서 서적을 추천하는 경우와 음반∩DVD 데이터 집합에서 DVD를 추천하는 경우)는 음반의 하위 카테고리의 숫자가 12개이므로, 4개 하위 카테고리에 대한 사용자 프로파일을 삭제하였으며, DVD의 경우에는 14이므로 5개의 하위 카테고리의 사용자 프로파일을 삭제하였다.

4.2 실험 결과

두 번째 실험 결과는 <표 7>과 같다. 각 데이터 집합에 대하여 2.*.5로 표시된 것이 Chi-square 값을 기준으로 하위 카테고리에 대한 사용자 프로파일을 삭제한 후의 추천 성과이다. 또한 서적 하위 카테고리에 대하여 직관에 의해서 7개의 하위 카테고리에 대한 사용자 프로파일을 삭제한 후의 추천 성과(Pre.)도 <표 7>에 함께 제시되어 있다. <표 7>을 살펴보면, 2.*.5의 추천 성과가 항상 2.*.4의 추천 성과보다 높을 것을 알 수 있다. 이러한 사실들은 <그림 1>을 통해서도 확인할 수 있다. 정리하면, 본 연구에서 제시한 Chi-square 값을 기준으로 하위 카테고리에 대한 사용자 프로파일을 제거한 후의 추천 성과가 전체 사용자 프로파일을 모두 사용한 경우보다 더 추천 성과가 좋음을 알 수 있다.

<표 7> 두 번째 실험 결과

데이터 집합				추천 성과			Test Statistics
서적∩음반	실험	학습집합	평가집합	Precision	Recall	F1	t-statistics(p-value)
	2.1.4	Music	Book	0.26282	0.14365	0.18302	-1.668 (0.104)
	2.1.5	Music	Book	0.37510	0.21056	0.26722	
	실험	학습집합	평가집합	Precision	Recall	F1	Duncan's Test
	2.2.4	Book	Music	0.26389	0.35900	0.30377	2.2.5 > Pre., 2.2.4
	Pre.	Book	Music	0.30122	0.45206	0.361	
	2.2.5	Book	Music	0.35714	0.51924	0.42299	
서적∩DVD	실험	학습집합	평가집합	Precision	Recall	F1	t-statistics(p-value)
	2.3.4	DVD	Book	0.26606	0.18271	0.21250	-2.379 (0.023)**
	2.3.5	DVD	Book	0.38033	0.24379	0.29670	
	실험	학습집합	평가집합	Precision	Recall	F1	Duncan's Test
	2.4.4	Book	DVD	0.21202	0.25739	0.23198	2.4.5, Pre. > 2.4.4
	Pre.	Book	DVD	0.38904	0.47731	0.42826	
	2.4.5	Book	DVD	0.40324	0.50253	0.44729	
음반∩DVD	실험	학습집합	평가집합	Precision	Recall	F1	t-statistics(p-value)
	2.5.4	DVD	Music	0.36937	0.51180	0.42986	-0.680 (0.505)
	2.5.5	DVD	Music	0.42110	0.50693	0.45900	
	실험	학습집합	평가집합	Precision	Recall	F1	t-statistics(p-value)
	2.6.4	Music	DVD	0.55976	0.58873	0.57256	-3.257 (0.003)***
	2.6.5	Music	DVD	0.65591	0.69979	0.67711	

주) *** 유의수준 0.01, ** 유의수준 0.05

<표 7>에서 실험 2.2, 실험 2.4는 Chi-square 값에 근거해서 제거한 추천 성과가 직관에 의한 제거 시의 추천 성과, 전체 사용자 프로파일을 모두 사용 시의 추천 성과가 차이가 있는지에 대한 Duncan's 테스트의 결과가 제시되어 있다. 직관에 의해 하위 카테고리를 제거하지 않은 음반이나 DVD의 경우(실험 2.1, 실험 2.3, 실험 2.5, 실험 2.6)는 전체 사용자 프로파일을 모두 사용한 경우인 실험 2.*.4와 Chi-square 값에 근거해서 제거한 후에 추천을 수행한 실험 2.*.5간의 평균 F1 값이 차이가 있는지가 t-테스트를 통해 검증되었다. 전체적으로 Chi-square 값에 근거해서 하위 카테고리에 대한 데이터를 제거하고 추천에 활용한 방안이 전체 데이터를 활용하여 추천하는 방안보다 더 좋은 추천 성과를 보였다. 또한 직관에 의해 하위 카테고리에 대한 데이터를 제거한 경우(실험 Pre.)보다도 Chi-square값에 근거하여 하위 카테고리에 대한 데이터를 제거하고 추천한 방안(실험 2.2.5와 실험 2.4.5)의 추천성과가 유사하거나 더 좋았다.

Ⅲ장과 Ⅳ장의 두 실험을 통해서 얻어진 내용을 바탕으로 2.2절에 제시되었던 두 개의 연구 질문에 대하여 다음과 같이 답을 정리할 수 있다.

- 첫 번째 실험의 결과에서 알 수 있듯이, 타 상품 카테고리 사용자 프로파일의 사용은 항상 무작위 추천과 인기도에 의한 추천에 비해서 나은 성과를 보여주었다. 따라서, 타 상품 카테고리 사용자 프로파일의 유용성을 확인할 수 있다. 물론 동일한 상품 카테고리의 사용자 프로파일을 사용한 경우에 비해서는 대부분 낮은 추천 성과를 보였다.
- 두 번째 실험에서 제시한 Chi-square 값에 근거한 사용자 프로파일의 선택적 사용 방안은 전체 사용자 프로파일을 모두 사용하는 것에 비해서 타 상품 카테고리 사용자 프로파일의 추천 성과를 향상시키는 것으로 확인되었다.

Ⅴ. 결 론

본 연구에서는 협업 필터링에서 타 상품 카테고리에 대한 사용자 프로파일을 사용하여 다른 상품 카테고리에 속한 상품을 추천하는 방안의 활용 가능성을 검토하였다. 타 상품 카테고리에 대한 사용자 프로파일을 활용할 필요가 있는 경우는 고객이 주로 방문하던 상품 카테고리의 웹 페이지가 아닌, 다른 상품 카테고리의 상품 페이지를 방문할 때 발생한다. 본 연구에서는 실제 특정 인터넷 상점의 패널 데이터를 활용하여 사용 가능성을 검토하였다. 실험 결과, 타 상품 카테고리에 대한 사용자 프로파일이 동일 상품 카테고리에 대한 사용자 프로파일이 존재하지 않는 경우 유용한 정보가 될 수 있음을 확인할 수 있었다. 또한 타 상품 카테고리를 효과적으로 활용하는 방안으로 상품 하위 카테고리들의 선호도 상관관계를 고려한 선택적 사용자 프로파일 활용 방안을 제시하고, 실험을 통해서 타당성을 검토하였다. 실험 결과, 제시한 선택적 활용 방법을 통해서 타 상품 카테고리 사용자 프로파일을 이용한 추천 성과를 높일 수 있음을 확인할 수 있었다.

본 연구 결과의 활용은 전자상거래 사이트에서 사용자들에게 상품 카테고리별로 개인화된 추천을 수행하는 경우에 가능하다. 상품 카테고리별로 사용자들에게 추천을 하기위해서는 일반적으로 하나의 상품 카테고리에 대한 사용자의 방문이력, 구매이력 등을 활용해서 해당 상품 카테고리에 속한 상품들을 추천하고 있다. 전자상거래 사이트의 경영자 입장에서는 다양한 상품 카테고리를 취급하는 방향으로 사이트의 확장이 이루어져 왔기 때문에 본 연구에서 제시된 기존 사용자 프로파일을 활용하여 사용자가 방문하지 않았던 상품 카테고리에 속한 상품들 중에 사용자가 선호하는 상품을 추천할 수 있으며, 이를 통해 판매증진이나 상품페이지 방문을 유도할 수 있다. 상품 카테고리들간의 사용자 유사도가 다

를 수 있기 때문에 어떤 상품 카테고리에 대한 사용자 프로파일을 추천에 활용하는 것이 적합한 지에 대한 시뮬레이션을 통해 적합한 상품 카테고리를 찾는 것이 필요하다.

본 연구에서 사용한 패널 데이터는 5개월간 특정 사이트에 방문한 전체 로그 데이터로 상당히 방대한 데이터로부터 실험을 시작하였다. 하지만 다수 상품 카테고리를 동시에 여러 번 방문한 고객만을 추출하여 실험을 하였기에 샘플의 크기가 만족스러울 만큼 크지 못했다. 따라서, 추후에는 더 방대한 데이터 집합을 활용하여 검증하는 것이 필요할 것으로 보인다.

본 연구에서는 고객 유사도 척도로 피어슨 상관계수를 사용하였다. 본 연구의 주목적이 협업 필터링 내에서의 타 상품 카테고리 고객 프로파일의 활용에 대한 것이고, 협업 필터링에서 고객 유사도 계산을 위해 활용되는 대표적인 방법이 피어슨 상관계수이므로 이를 사용하였다. 협업 필터링 연구에서 이진 데이터에 대하여 피어슨 상관계수를 적용한 연구들도 존재하지만[Breese *et al.*, 1998], 이진 데이터에 대한 피어슨 상관계수 사용에 대한 반론들도 존재한다[Mild and Reutterer, 2001; Mobasher *et al.*, 2001]. 따라서, 이진 데이터를 위해서 제시된 다른 유사도 척도들을 활용한 추가적인 연구가 추후 필요하다. 또한 본 연구에서는 협업 필터링에서의 타 상품 카테고리 프로파일의 활용 가능성을 검토하였는데, 내용 기반 추천에서의 타 상품 카테고리 프로파일의 활용 가능성을 검토하는 것도 흥미로운 주제가 될 것이다.

〈참 고 문 헌〉

[1] 김종우, 배세진, 이홍주, "협업 필터링 기반 개인화 추천에서의 평가자료의 희소 정도의 영향," *경영정보학연구*, 제14권 제2호, 2004, pp. 131-149.

[2] 박상규, 이재성, 신승은, 강유환, 오효정, 장명길, 서영훈, "문서필터링을 위한 질의어 확장과 가중치 부여기법," *정보처리학회논문지B*, 제10권 제7호, 2003, pp. 743-751.

[3] 이재규, 권순범, 김우주, 김민용, 송용욱, 최형림, *전자상거래원론*, 법영사, 2002.

[4] 조윤호, 박수경, 안도현, 김재경, "재구성된 제품 계층도를 이용한 협업 추천 방법론 및 그 평가," *한국경영과학회지*, 제29권 제2호, 2004. 6, pp. 59-76.

[5] 황병연, "개선된 추천을 위해 클러스터링을 이용한 협업 필터링 에이전트 시스템의 성능," *정보처리논문지*, 제7권 제5S호, 2000. 5, pp. 1599-1608.

[6] Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A., "Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach," *ACM Transactions on Information Systems*, Vol. 23, No. 1, January 2005, pp. 103-145.

[7] Allen, C., Kania, D., and Beth, Y., *Internet World Guide to One-to-One Web Marketing*, John Wiley & Sons, Inc., New York, 1998.

[8] Ansari, A., Essegaier, S., and Kohli, R., "Internet Recommendation Systems," *Journal of Marketing Research*, Vol. XXXVII, August 2000, pp. 363-375.

[9] Berry M.J.A. and Linoff, G.A., *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*, Wiley, Indianapolis, Indiana, 2004.

[10] Breese, J.S., Heckerman D., and Kadie, C., "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Technical Report*, MSR-TR-98-12, Microsoft Research,

1998.

[11] Cho, Y.H., Kim, J.K., and Kim, S.H., "A Personalized Recommender System Based on Web Usage Mining and Decision Tree Induction," *Expert Systems with Applications*, Vol. 23, No. 3, 2002, pp. 329-242.

[12] Dragon, R.V., "Recommendation Systems - Advice from the Web," *PC Magazine*, 1997.

[13] Goldberg, K., Roeder, T., Gupta, D., and Perkins, C., "Eigenstate: A Constant Time Collaborative Filtering Algorithms," *Information Retrieval*, Vol. 4, 2001, pp. 131-151.

[14] Gupta, Ohruv, Mark Digiovanni, Hiro Norita, and Ken Goldberg, "Jester 2.0: Evaluation of a New Linear Time Collaborative Filtering Algorithm," *22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, August, 1999.

[15] Herlocker, J., Konstan, J., Borschers, A., and Riedl, J., "An Algorithmic Framework for Performing Collaborative Filtering," Proceedings of SIGIR'99, Berkley, CA, U.S.A, 1999, pp. 230-237.

[16] Herlocker, J., Konstan, J., Terveen, L., and Riedl, J., "Evaluating Collaborative Filtering Recommender Systems," *ACM Transactions on Information Systems*, Vol. 22, No. 1, January 2004.

[17] Kim, J.W., Lee, B.H., Shaw, M.J., Chang, H., and Nelson, M., "Application of Decision Tree Induction Techniques to Personalized Advertisements on Internet Storefront," *Informational Journal of Electronic Commerce*, Vol. 5, No. 3, 2001, pp. 45-62.

[18] Kim, J.W. and Lee, H.J., "Data Sparsity and Performance in Collaborative Filtering-based Recommendation," *International Journal of Management Science*, Vol. 10, No. 3, 2005, pp. 19-45.

[19] Krishnamurthy, Sandeep, *E-Commerce Managegment*: Text and Cases, South Western, 2003.

[20] Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., and Riedl, J., "GroupLens: Applying Collaborative Filtering to Usenet News," *Communication of the ACM*, Vol. 40, No. 3, 1997, pp. 77-87.

[21] Mild, A. and Reutterer, T., "Collaborative Filtering Methods for Binary Market Basket Data Analysis," *Lecture Notes in Computer Science*, Vol. 2252, 2001, pp. 302-313.

[22] Mobasher, B., Dai, H., Luo, T., and Nakagawa, M., "Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data," *Proceedings of the workshop Intelligent Techniques for Web Personalization, IJCAI-2001*, Seattle, Washington, 2001, pp. 53-60.

[23] O'mahony, M., Hurley, N., Kushmerick, N., and Silvestre, G., "Collaborative Recommendation: A Robustness Analysis," *ACM Transactions on Internet Technology*, Vol. 4, No. 4, November 2004, pp. 344-377.

[24] Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., McNee, S.M., Konstan, J.A., and Riedl, J., "Getting to Know You: Learning New User Preferences in Recommender Systems," *Proceedings of the ACM IUI'02*, San Francisco, 2002, pp. 127-134.

[25] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," *Proceedings of the ACM 1994 Conference on Computer Support-*

ed Cooperative Work, New York, ACM, 1994, pp.

[26] Sarwar, B., Karypis, J. Konstan, and Riedl, J., 2001, "Item-Based Collaborative Filtering Recommendation Algorithms," *Proceedings of WWW01*, Hong Kong, 285-295.

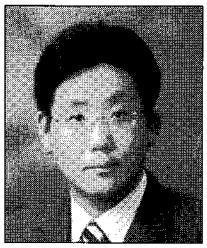
[27] Schafer, J.B., Konstan, J., and Riedl, J., "E-commerce Recommendation Applications," *Data Mining and Knowledge Discovery*, Vol. 5, No. 1-2, 2001, pp. 115-153.

[28] Shardanand, U. and Maes, P., "Social Information Filtering: Algorithms for Automating 'Word of Mouth'," *Proceedings of Conference on Human Factors in Computer Systems*, 1995, pp. 210-217.

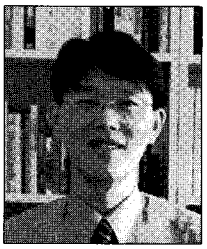
[29] Ungar, L.H. and Foster, D.P., "Clustering Methods for Collaborative Filtering," *Proceedings of Workshop on Recommendation Systems at the Fifteenth National Conference on Artificial Intelligence*, 1998.

[30] Weng, S.S. and Liu, M.j., "Feature-based Recommendations for One-to-one Marketing," *Expert Systems with Application*, Vol. 26, No. 4, 2004, pp. 493-508.

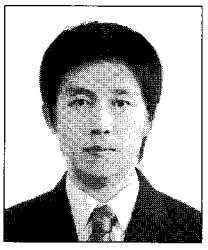
◆ 저자소개 ◆



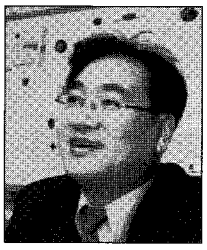
박수환 (Park, Soo Hwan)
홍익대학교에서 경영정보학과를 졸업하고 한양대학교 대학원에서 경영학 석사를 취득하였으며, 현재 SQ Technology에서 재직 중이다. 주요 관심분야는 추천시스템, 지식경영, 데이터 마이닝 등이다.



김종우 (Kim, Jong Woo)
현재 한양대학교 경영학부 부교수로 재직 중이다. 서울대 수학과에서 이학사, 한국과학기술원 경영과학과에서 공학석사를 취득하고 한국과학기술원 산업경영학과에서 공학박사를 취득하였다. 한국과학기술원 경영정보연구센터 연수연구원, University of Illinois at Urbana-Champaign 방문연구원, 충남대학교 통계학과 부교수로 근무한 경력이 있다. 주요 관심분야는 경영정보시스템, 의사결정지원시스템, 전자상거래, 추천시스템, 데이터 마이닝 응용, B2B 비즈니스 프로세스 모델링 등이다.



이홍주 (Lee, Hong Joo)
현재 MIT Sloan School of Management의 Center for Collective Intelligence에서 Post Doc.으로 재직 중이다. 한국과학기술원(KAIST) 산업경영학과에서 이학사, KAIST 테크노경영대학원에서 공학석사, 공학박사를 취득하였다. 주요 관심분야는 개인화추천, 정보검색, 시맨틱 웹, 모바일 웹, 데이터 마이닝 응용 등이다.



조남재 (Cho, Namjae)
서울대학교에서 산업공학 학사, 한국과학기술원에서 경영과학 석사, 미 보스턴대학교에서 경영정보학 박사를 취득하였다. 현재 한양대학교 경영학부 교수로 재직 중이다. 한국소프트웨어 진흥원자문 위원, 서울도시철도공사 전산자문위원, 산자부 e-biz 인덱스자문위원, 한양대디지털 경영 연구센터소장 등을 수행하고 있다. 주요 관심분야는 전자상거래와 e-비즈니스, 지식경영, 디지털 산업 전략 및 정책 등이다.

◈ 이 논문은 2006년 5월 2일 접수하여 1차 수정을 거쳐 2006년 8월 29일 게재확정되었습니다.