경영정보학연구
제14권 제2호
2004년 6월

협업 필터링 기반 개인화 추천에서의 평가자료의 희소 정도의 영향*

김 종 우**, 배 세 진***, 이 홍 주****

Sparsity Effect on Collaborative Filtering-based Personalized Recommendation

Jong Woo Kim, Se Jin Bae, Hong Joo Lee

Collaborative filtering is one of popular techniques for personalized recommendation in e-commerce sites. An advantage of collaborative filtering is that the technique can work with sparse evaluation data to predict preference scores of new alternative contents or advertisements. There is, however, no in-depth study about the sparsity effect of customer's evaluation data to the performance of recommendation. In this study, we investigate the sparsity effect and hybrid usages of customers' evaluation data and purchase data using an experiment result. The result of the analysis shows that the performance of recommendation decreases monotonically as the sparsity increases, and also the hybrid usage of two different types of data; customers'evaluation data and purchase data helps to increase the performance of recommendation in sparsity situation.

Keywords : personalization, recommendation techniques, collaborative filtering, sparsity

* 본 연구는 2003년 한양대학교 교내 연구비 지원으로 연구되었음.
** 한양대학교 경영학부
*** 충남대학교 자연대학 통계학과
**** 한국과학기술원 테크노경영대학원

Ⅰ. 서 론

고객의 취향이나 관심에 맞추어 상품이나 컨텐츠를 제공하는 고객 맞춤(customization) 또는 개인화(personalization) 서비스는 인터넷 상점이나 정보 서비스 제공자의 주요한 성공 요인의 하나로 인식되고 있다[이재규 외, 2002; Allen et al., 1998; Schafer et al., 2001; Turban et al., 2003]. 이러한 개인화 서비스를 구현하는 기술 중의 하나인 상품 추천 기술(recommendation techniques)은 고객의 인구통계학적인 정보, 유사 고객의 선호도, 과거의 구매 행위 등의 정보들을 바탕으로 고객에게 개인화된 상품을 추천해주는 기술이다[Ansari et al., 2000; Cho et al., 2002; Dragon, 1997; Kim, J.K. et al., 2001; Lee et al., 2002; Schafer et al., 2001]. 실제로 추천 기술들은 Amazon, Yahoo, HP Shopping Village, Wal-Mart, Half.com, Musician's Friend 등의 유수의 전사상거래 사이트들에서 적용되고 있다[Allen at al., 1998; Ansari et al., 2000; BroadVision, 2004; NetPerception, 2004]. 이러한 추천 기술은 고객 입장에서는 원하는 상품을 찾기 위한 노력과 시간을 줄여주고, 인터넷 상점 입장에서는 고객의 충성도(loyalty) 증대, 판매 증대, 광고 수익의 증대, 타겟 홍보의 이익을 가져다 준다[Ansari et al., 2000; Resnick et al., 1994].

현재 상용화된 추천 시스템들에서 가장 많이 활용되고 있는 추천 기법은 협업 필터링(collaborative filtering)이다[김재경 외, 2002; 황병연, 2000; Balabanovic and Shoham, 1997; Breese et al., 1998; Konstan et al., 1997; Resnick et al., 1994]. 협업 필터링은 해당 고객과 선호도가 유사한 고객들의 상품에 대한 평가점수를 활용하여 고객에게 적합한 상품 정보를 제공한다. 협업 필터링의 주요한 장점 중에 하나는 고객들의 상품에 대한 평가점수가 결측치를 갖는 경우(즉, 각 고객이 대상 상품들을 모두 평가하지 않고 일부만 평가한 경우)에도 적용할 수 있다는 것이다. 비록 평가점수가 결측지를 갖고 있어도 협업 필터링을 사용할 수 있지만, 평가 데이터의 희소 정도(sparsity)는 협업 필터링의 성능에 중요한 영향을 미친다. 하지만 이러한 희소 정도가 협업 필터링의 성능에 미치는 영향에 대한 체계적인 분석은 이루어지고 있지 못하다. 본 논문에서는 평가 데이터의 희소 정도가 협업 필터링에 미치는 영향에 대하여 온라인 서점에 대한 실험 데이터를 토대로 분석하도록 한다. 또한 협업 필터링은 고객의 상품들에 대한 평가점수뿐만 아니라 구매 데이터를 활용하여 적용할 수 있는데, 희소 정도에 따라서 이 두 가지 형태의 데이터를 함께 활용함으로써 얻을 수 있는 추천 성능 향상 정도에 대하여 분석하도록 한다. 즉, 평가점수 또는 구매 데이터의 희소 정도가 커지면 협업 필터링 추천의 성능 감소가 어떤 형태로 일어나는가와, 평가점수와 구매 데이터의 희소 정도를 고려하여 추천 성능을 향상시킬 수 있는가에 대한 분석을 수행하도록 한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 전자상거래 상품 추천 기술과 협업 필터링에 대하여 살펴보도록 한다. Ⅲ장에서는 본 연구의 희소 효과 분석을 위한 실험 내용을 소개한다. Ⅳ장에서는 본 연구에서 비교한 Hybrid 협업 필터링 추천 모델을 제시하고, 희소 효과 분석 결과를 제시한다. Ⅴ장에서는 희소 효과와 관련된 논의사항들을 다루고, Ⅵ장에서는 결론과 추후 연구 과제를 제시한다.

Ⅱ. 관련 연구

2.1 상품 추천 기술

웹 사이트 개인화란 특정 고객 또는 고객 집단에게 맞춤 웹 경험을 제공함으로써, 고객의 반응성(responsibility)을 높이기 위한 마케팅 전

략이다. 특히 전자상거래 상품 추천 기술은 웹 사이트 개인화를 지원하기 위한 방법의 하나로, 고객 프로파일, 거래 데이터, 웹 로그 데이터 등을 토대로 각 고객의 구매 성향과 특성을 분석하여 각 개별 고객에게 적합한 정보를 제공하는 기술을 의미한다. 상품 추천 방법들은 고객에게 개인적인 선호도를 어느 정도 명시적으로 표시하도록 요구하느냐에 따라 구분할 수 있다. Mulvenna 등은 이러한 기준으로 추천 기술을 크게 세 가지 형태, 체크박스형, 협업 필터링형, 관찰형으로 구분하였다[Mulvenna et al., 2000]. 가장 단순한 형태인 체크박스형 추천 기술은 고객에게 명시적으로 자신의 선호도를 표시하도록 요구하는 형태이다. 고객의 참여 정도에 대한 스펙트럼에서 체크박스형의 반대 쪽에 존재하는 추천 기술이 관찰형이다. 관찰형 추천 기술의 경우는 고객의 명시적 선호도 표현 없이, 구매 데이터, 웹 로그 데이터 등 인터넷 상점에서 고객 행위를 활용하여 고객의 선호도를 추측하여 추천하는 방법이다. 일반적으로 의사결정나무추론, 연관성분석, 신경망 등의 데이터 마이닝 기법들이 관찰형 추천 기술을 위해서 사용된다[김종우 외, 2000; 이건창 외, 1999; Cho et al., 2002; Kim, J.W. et al., 2001; Mobasher et al., 2000]. 고객의 참여 정도 스펙트럼의 중간 정도에 위치하는 기술이 협업 필터링형이다. 협업 필터링 기반 추천 기술은 개인화 서비스 개시 이전에, 고객에게 몇몇 상품이나 컨텐츠들을 제공하고 이에 대한 선호 점수를 입력 받아서, 타 고객과의 선호도 유사도를 추정하고 이를 이용하여 추천하는 방법이다[Konstan et al., 1997]. 상품 추천을 위해서 실제로 가장 많이 활용되고 있는 협업 필터링 기법에 대한 주요 연구는 미네소타 대학의 GroupLens, 버클리 대학의 Jester, MIT의 음악추천시스템 RINDO 등이 있으며, 주요 상용화 제품에는 Net Perceptions 사의 GroupLens Recommendation Engine, Gustos사의 Gustos Prediction Server, BeFree사의 BSELECT, Autonomy사의 Autonomy Collaborative Filtering 등이 있다[Autonomy, 2004; BeFree, 2004; Gupta et al., 1999; Gustos, 2004; Konstan et al., 1997; NetPerception, 2004; Resnick et al., 1994; Shardanand and Maes, 1995].

2.2 협업 필터링과 희소 효과

협업 필터링에서 개인화 추천을 위해서는 사전에 각 고객별로 몇 개의 상품에 대한 선호도 평가점수가 필요하다. 예를 들어 <표 1>에서와 같이 4명의 고객에 대해 상품 A부터, 상품 E까지의 평가점수가 5점 척도로 주어져있고, 새로운 상품 F에 대한 고객 1의 선호도 점수를 예측하고 싶다고 하자. <표 1>에서 볼 수 있듯이 평가점수 행렬은 희소행렬(sparse matrix) 형태를 가지고 있다. 협업 필터링의 첫 단계는 각 사용자별로 다른 고객과의 유사도를 구하는 것이다. 고객간의 유사도는 식 (1)에서와 같이 상관계수(correlation coefficient) 형태로 계산된다. 식 (1)은 고객 i, j간의 상관계수 r_{ij}를 구하는 식으로, S_{ik}는 고객 i의 상품 k에 대한 평가점수이고, $\overline{S_i}$는 고객 i의 평가점수의 평균이다. 상관계수 r_{ij}는 두 고객의 선호도가 유사한 경우에는 1에 가까운 값을 가지게 되고, 상반된 선호도를 갖는 경우는 1에 가까운 값을 가지게 된다. 예를 들어 고객 1의 경우는 고객 2와 상반된 선호도를 가지고, 고객 3과는 유사한 선호도를 가지므로, 상관계수가 각각 -0.8, +1로 계산된다.

<표 1> 선호도 평가점수 예제

상품 / 고객	상품 A	상품 B	상품 C	상품 D	상품 E	상품 F
고객 1	1	5		2	4	?
고객 2	4	2		5	1	2
고객 3	2	4	3			5
고객 4	2	4		5	1	

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

고객 1의 상품 F에 대한 선호도 점수의 예측은 다음 식 (2)을 통해서 이루어진다. 식 (2)는 고객 i의 광고 k에 대한 선호도 점수를 예측하는 식으로, *Rater(k)*는 상품 k를 평가한 고객의 집합을 의미한다. 고객 1의 상품 F에 대한 선호도 점수를 예측하기 위해서, 상품 F를 평가한 고객들(고객 2, 고객 3)의 평가점수와 이들 고객들과 고객 1의 상관계수를 이용하여 계산된다. 식 (2)의 식에서 알 수 있듯이, 새로운 상품에 대해서 고객들의 선호도 점수를 예측하기 위해서는 누군가가 사전에 해당 상품에 대해서 평가를 해주어야 한다. 고객의 선호도가 파악되지 않은 신규 상품의 경우에는 협업 필터링을 적용하기 어려우며, 이를 위해서 신규 상품에 대해 선호도를 입력하여 주는 평가자 집단(Rater)을 별도로 두거나, 아니면 최소평가횟수를 정하여 그 평가횟수를 넘는 상품만을 추천 대상으로 사용할 수도 있다[Mild and Natter, 2002; Sarwar, 2000].

$$P_{ik} = \overline{S_i} + \frac{\sum_{l \in Rater(k)} (S_{lk} - \overline{S_l}) r_{ij}}{\sum_{l \in Rater(k)} |r_{ij}|} \quad (2)$$

협업 필터링의 장점 중에 하나는 <표 1>에서와 같이 고객의 상품들에 대한 평가점수 행렬이 희소행렬 형태인 경우에도 활용될 수 있다는 것이다. 하지만, 평가점수의 희소 정도가 추천의 성능에 큰 영향을 줄 것으로 예상할 수 있으므로, 협업 필터링의 실제적인 활용을 위해서는 이러한 희소 정도가 추천 성능에 어떠한 영향을 미치는 지에 대한 체계적인 분석이 필요하다. 희소한 고객 선호도 데이터 집합을 가지고 협업 필터링을 수행하게 되면 고객에게 적절한 추천을 하기가 어렵기 때문에, 기존 고객 선호도 데이터 집합을 다양한 기법을 통해 차원을 축소(dimension reduction)하여 협업 필터링에 활용하려는 연구들도 진행되고 있다[Sarwar et al., 2000]. Sarwar et al.[2000]에 따르면, 협업 필터링의 유사도 계산에 활용되는 상품의 수를 일정수준까지 줄여나가도 추천성과에 많은 영향을 미치지 않는 것으로 파악되었다. 또한 차원축소를 위해 Latent Semantic Indexing(LSI) 기법 같은 방안을 활용하여 고객-상품 행렬의 차원을 축소한 후에, 축소된 데이터 집합을 협업 필터링에 활용하여 데이터 희소로 인한 문제와 협업 필터링의 scalability 문제를 해결하는데 도움을 주고 있다[Sarwar et al., 2000]. 또한 데이터의 희소성을 해결하기 위한 대안으로 고객 클러스터링을 활용하는 방안도 제시되고 있다[Ungar and Foster, 1998]. 즉, 한 고객의 데이터는 항상 희소하기 때문에, 유사한 고객들의 군집을 생성하고, 군집에 속한 고객들을 한 고객처럼 다루는 방안도 가능하다. 하지만, 이러한 방안들의 적용에 앞서서 희소 정도와 추천 성능간의 관계에 대한 분석이 필요하다.

협업 필터링을 적용할 경우의 문제점 중에 하나는 고객들이 개인화 서비스를 받기 위해서는 일단 등록 시에 다수의 상품에 대한 선호도 평가점수를 입력해야 하는 수고가 필요하다는 것이다. 이러한 문제점을 피하기 위한 방안의 하나로는 고객에게 등록 시 선호도 평가점수를 입력 받지 않고, 구매 데이터를 사용하여, 이를 상품 선호도에 대한 이진 평가점수로 생각하여 고객간 상관계수 계산을 수행하는 방안이 있다[Breese et al., 1998]. 즉, 앞에 식 (1), 식 (2)에서의 S_{ik}를 사용하지 않고 식 (3)과 같이 고객의 해당 상품의 구매여부에 대한 변수 B_{ik}를 사용하여, 고객간 상관계수를 다음 식 (4)와 같이 계

산한다. 개별 고객의 평가점수는 존재하지 않고, 평가자 집단의 평가점수만 존재한다고 가정하면, 선호도 점수 예측은 식 (5)와 같이 계산된다. 식 (5)에서 $\overline{S_{\cdot k}}$는 상품 k에 대한 평가자 집단의 평가점수 평균이다.

$$B_{ik} = \begin{cases} 1 & \text{고객 } i\text{가 상품 } k\text{를 구매한 경우} \\ 0 & \text{구매하지 않은 경우} \end{cases} \quad (3)$$

$$r_{ij} = \frac{Cov(i,j)}{\sigma_i \sigma_j} = \frac{\sum_k (B_{ik} - \overline{B_i})(B_{jk} - \overline{B_j})}{\sqrt{\sum_k (B_{ik} - \overline{B_i})^2 (B_{jk} - \overline{B_j})^2}} \quad (4)$$

$$P_{ik} = \overline{S_{\cdot i}} + \frac{\sum_{l \in Rater(k)} (S_{lk} - \overline{S_l}) r'_{ij}}{\sum_{l \in Rater(k)} |r'_{ij}|} \quad (5)$$

고객의 평가점수를 사용하지 않고 구매 이력을 사용하는 식 (4)와 식 (5)를 사용한 추천 방법은 구매 이력이 없는 신규 고객에게는 적용할 수 없다. 또한, 평가점수와 구매 데이터가 동시에 가용한 상황이라면, 이 두 가지 정보를 함께 사용하여 추천하는 경우에 보다 좋은 추천 성능을 보일 것으로 예상할 수 있다. 이러한 경우 추천 성능이 평가점수와 구매 데이터의 희소 정도에 따라서 어떻게 달라지는 지에 대해 고찰하는 것이 필요하다. 본 연구에서 다루고자 하는 연구 질문들은 다음과 같다.

- 평가점수 또는 구매 정보의 희소 정도가 커지면 협업 필터링 추천의 성능 감소가 어떠한 형태로 발생하는가?
- 평가점수 또는 구매 정보의 희소 정도를 반영하여 두 정보를 함께 사용함으로써 추천 성능을 향상시킬 수 있을까?
- 평가점수 또는 구매 정보 중 한 쪽의 희소 정도가 적으면, 다른 정보의 사용이 불필요한가? 예를 들어, 평가점수가 희소성을 가지지 않은 경우에는 구매 정보가 추천의 성능에 추가적인 도움을 주지 않는가? 또 반대로, 구매가 몇 건 이상 이루어진 고객에 대해서는 평가점수를 사용하지 않고 구매 정보만을 사용하여 추천하여도 성능에 차이가 없는가?

Ⅲ. 온라인 서점에서의 추천 실험

3.1 구매 데이터 수집을 위한 1차 실험

협업 필터링에서 희소 정도의 영향을 분석하기 위해서 2차에 걸쳐서 실험을 수행하였다. 1차 실험은 구매 데이터를 수집하기 위한 실험으로 와우북(www.wowbook.com) 웹사이트를 이용하여 140명을 대상으로 실험을 수행하였다. 실험자들은 와우북 홈페이지에 접속하여 실제 서적을 구매하는 것과 동일한 방법으로, 자유롭게 구매하고자 하는 도서를 선택하여 장바구니에 담았다. 충분한 구매 데이터의 수집을 위해 실험자들이 실험 당시에 선호하는 서적 외에도 예전에 실험자가 구매했던 도서도 함께 장바구니에 담도록 권유하였다. 이렇게 해서 최종적으로 개인 장바구니에 담긴 서적들을 실험자의 구매 데이터로 생성하였다. 다음 <그림 1>은 구매 데이터를 수집하기 위한 장바구니 캡쳐 화면이다.

<그림 1> 와우북 장바구니 캡쳐 화면

3.2 고객 선호도 조사를 위한 2차 실험

2차 실험은 1차 실험자와 동일한 140명에 대하여 인터넷 설문 형태로 수행하였다. 이 실험을 위해서 와우북 사이트의 해당월의 신간 중에서 <표 2>와 같은 16개 대분류마다 3권의 도서를 임의로 선정하여 총 48권의 신간 서적을 선정하였다. 이들 48개 서적에 대하여 <그림 2>에서와 같이 인터넷 설문 형태로 실험자에게 5점 척도로 선호도 평가점수를 부여하도록 하였으며, 실험자들은 자신의 컴퓨터를 통해 실험 사이트에 접속하여 설문에 참여하였으며 설문에 소요되는 시간에 제한을 두지는 않았다. 1차 실험과 2차 실험 대상 서적이 중복되는 것을 피하기 위해서 2차 실험은 1차 실험 종료 후 1개월이 경과한 후에 수행하였다. 실험 후 분석에서 16개 대분류별 3권의 도서 중 1권(총 16권)에 대한 선호도 점수를 협업 필터링의 초기 입력으로 사용하였고, 나머지 2권에 대한 평가점수를 추천 방법들의 성능 비교를 위해서 사용하였다. 또한 140명의 실험자 중 30명을 임의로 선정하여 평가자 집단으로 활용하였고, 나머지 110명의 실험자들의 평가점수를 사용하여 성능 비교를 수행하였다.

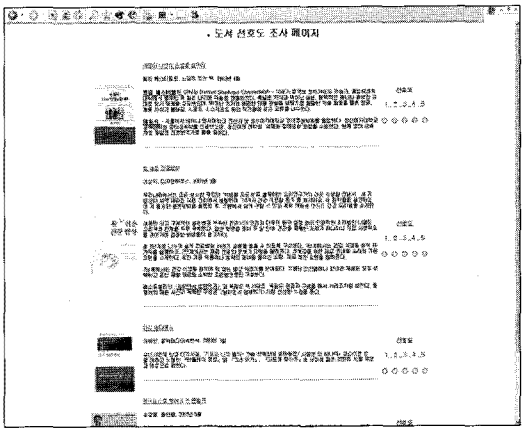
<표 2> 와우북의 서적 대분류 카테고리

컴퓨터/인터넷	경영/경제	외국어/어학	어린이
취미/건강	만화/애니메이션	소 설	시
인 문	에세이	고 전	사회과학
과 학	역 사	예 술	잡 지

Ⅳ. 희소 효과 분석

4.1 Hybrid 협업 필터링 방법과 성능 비교

본 연구에서는 Ⅱ장에서 제시한 고객 5점 척도



<그림 2> 인터넷 설문 화면

평가점수에 기반한 협업 필터링 식 (1), 식 (2)과 구매 데이터에 기반한 협업 필터링 식 (4), 식 (5) 이외에 두 가지 정보가 동시에 존재하는 경우에, 식 (1)과 식 (4)의 상관계수를 가중 평균해서 상관계수, 선호도 점수 예측값을 구하는 Hybrid 알고리즘을 함께 검토하였다(식 (6), 식 (7) 참조). 식 (6)에서는 평가점수에 기반한 상관계수에 대한 가중치로 0과 1사이의 값을 갖는다.

$$r^{\sim}_{ij} = wr_{ij} + (1-w)r'_{ij} \quad (6)$$

$$P^{\sim}_{ik} = \overline{S_i} + \frac{\sum_{l \in Rater(k)} (S_{lk} - \overline{S_l})\, r^{\sim}_{ij}}{\sum_{l \in Rater(k)} |r^{\sim}_{ij}|} \quad (7)$$

추천 방법의 성능을 비교하기 위해서, 각 추천 방법으로 평가자 집단을 제외한 110명의 각 실험자별로 32권의 서적 중에서 3권을 추천한다고 가정하였다. 즉, 각 추천 방법에 따라서 선택된 3권의 서적에 대해서 2차 실험에서 수집된 5점 척도 값들의 평균값을 활용하여 성능을 비교하였다. <표 3>은 3가지 형태의 협업 필터링과 임의로 3권을 추천한 경우, 32권의 서적 중 평균 선호도 점수가 가장 높은 3권을 추천한 경우(이하 Top 3 방법으로 부름)의 성능을 비교한 결과이다. 5가지 추천 방법간의 통계적

<표 3> 5가지 추천 방법의 성능 비교

추 천 방 법	평 균	표 준 편 차	ANOVA (F-value)	Duncan Test
(1) 임의 추천	2.7455	0.76616	16.077*	(3), (5) > (2), (4) > (1)
(2) Top 3 방법	3.2303	0.73050		
(3) 5점척도 평가점수 기반 협업 필터링	3.4545	0.79207		
(4) 구매 데이터 기반 협업 필터링	3.2272	0.85065		
(5) Hybrid 협업 필터링 (평가점수 가중치 : 0.3, 구매데이터 가중치 : 0.7)	3.5030	0.77781		

* 유의 수준 α = 0.05에서 통계적으로 유의

으로 유의한 차이가 있는지 살펴보기 위하여 분산분석(ANOVA)과 던컨 실험(Duncan Test)를 실시하였으며, 결과는 <표 3>에 포함되어 있다. 비교실험에 활용한 Hybrid 필터링 방안의 가중치는 추천 성과의 평균이 가장 높았던 평가점수 상관계수 가중치 0.3과 구매데이터 상관계수 가중치 0.7인 경우의 자료이다.

임의로 추천한 경우보다는 다른 방안들을 사용하여 추천하는 것의 추천 성과가 모두 통계적으로 유의하게 높았다. Top3 방법과 구매 데이터 기반의 협업 필터링 방안은 추천 성과가 유사하다고 볼 수 있으며, 평가점수를 활용한 방안이나 Hybrid 협업 필터링 방안보다는 성과가 낮았다. Hybrid 협업 필터링 방안이 평가점수 기반 협업 필터링 방안보다 추천 성과 평균은 약간 높았으나 통계적으로 유의한 차이는 아니었다.

4.2 희소 효과 분석

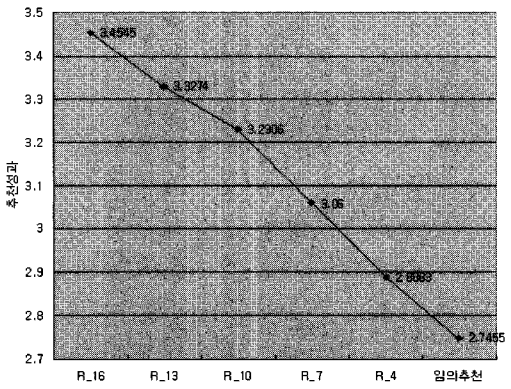
4.2.1 평가점수에 기반한 협업 필터링

먼저 5점 척도로 된 평가점수의 희소 효과를 분석하였다. 이를 위해서 협업 필터링에 입력 데이터로 사용되는 16개 서적에 대한 실험자의 평가점수에 대해서, 다음과 같이 6가지 경우로 결측치를 포함하는 평가점수 데이터를 생성하였다. 데이터의 희소도는 $1-\frac{\text{결측되지 않는 데이타의수}}{\text{모든 데이타의 수}}$ 와 같이 계산되며[Sarwar et al., 2000], 무작위 추출에 의해 결측치를 생성하였기 때문에 해당 데이타 집합의 평균적인 희소 정도를 표기하였다.

- 완전활용(R_16): 16개 서적에 대한 실험자의 모든 평가점수 데이터 활용 (희소도 = 0)
- Random_13(R_13): 실험자별로 정규분포 N(3,2)에 따라 난수를 발생시키고 해당 개수만큼 임의로 평가점수를 삭제. 즉, 실험자가 평가한 점수의 개수가 평균적으로 13개이고 표준편차가 2가 되도록 임의로 평가점수를 삭제하여 새로운 평가점수 데이터 생성 (희소도 = 0.1875)
- Random_10(R_10): 평가점수에서 정규분포 N(6,2)을 따라 임의의 평가점수를 삭제한 데이터 (희소도 = 0.375)
- Random_7(R_7): 평가점수에서 정규분포 N(9,2)을 따라 임의의 평가점수를 삭제한 데이터 (희소도 = 0.5625)

- Random_4(R_4): 평가점수에서 정규분포 N(12,2)을 따라 임의의 평가점수를 삭제한 데이터 (희소도 = 0.75)
- 임의추천: 성능비교를 위해 사용자들에게 임의로 책을 추천하여 주는 경우(평가점수가 전혀 없는 경우)

앞에서 제시된 6가지 평가점수 데이터의 생성과 추천을 30회 반복하여 수행하였으며, 수행결과는 <그림 3>과 같다. <그림 3>에서 볼 수 있듯이 평가점수의 희소 정도가 커질수록 평가점수에 기반한 협업 필터링의 성능이 떨어지는 것을 확인할 수 있으며, 데이터가 희소해 지더라도 임의로 추천하는 경우보다는 성과가 좋았다.



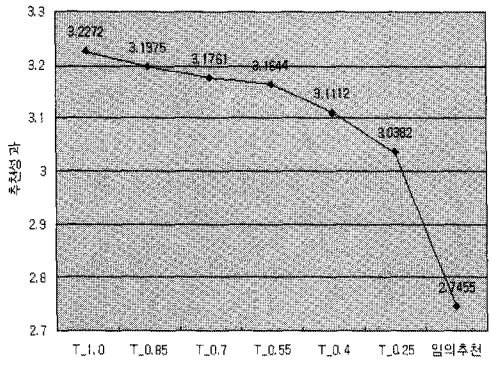
<그림 3> 평가점수에 기반한 협업 필터링에서의 희소 효과

4.2.2 구매 데이터에 기반한 협업 필터링

1차 실험에서 얻어진 구매 데이터에는 총 1078권의 책이 포함되었다. 평가자 집단의 구매 평균이 12.3권이며 표준편차는 7.0815이었다. 30명의 평가자 중의 가장 많은 구매권수는 37권이며, 최소 구매권수는 4권이었다. 평가자를제외한 나머지 실험자의 구매 평균은 10.5727권이며 표준편차 6.9486이고, 최대 구매권수는 50권이며 최소 구매권수는 3권이었다. 구매 데이터의 희소효과 분석을 위해 다음과 같이 7가지 경우로 결측치를 포함하는 구매 데이터를 생성하였다.

- 완전활용(T_1.0): 실험자들의 모든 구매 데이터 활용 (희소도 = 0)
- Random_0.85(T_0.85): 실험자들의 구매 데이터에서 평균 15%와 표준편차 2를 따르게 임의의 구매 데이터를 삭제하여 새로운 구매 데이터 생성 (희소도 = 0.15)
- Random_0.7(T_0.7): 모든 구매 데이터에서 평균 30%와 표준편차 2를 따르게 임의의 구매 데이터를 삭제한 데이터 (희소도 = 0.30)
- Random_0.55(T_0.55): 모든 구매 데이터에서 평균 45%와 표준편차 2를 따르게 임의의 구매 데이터를 삭제한 데이터 (희소도 = 0.45)
- Random_0.4(T_0.4): 모든 구매 데이터에서 평균 60%와 표준편차 2를 따르게 임의의 구매 데이터를 삭제한 데이터 (희소도 = 0.60)
- Random_0.25(T_0.25): 모든 구매 데이터에서 평균 75%와 표준편차 2를 따르게 임의의 구매 데이터를 삭제한 데이터 (희소도 = 0.75)
- 임의추천: 성능비교를 위해 사용자들에게 임의로 책을 추천하는 경우

전체 구매 데이터에서 평균 사용 정도에 근거하여 각 구매 데이터 집합을 표기하였다. 즉, T_0.85는 전체 구매 데이터 중 평균 85%를 활용하였다는 의미이고, 괄호 안의 희소도는 평가점수의 경우와 마찬가지로, 무작위 추출에 의해 결측치를 생성하였기 때문에 평균적인 희소 정도를 표기하였다. 앞에서 제시한 7가지 경우에 대해서 구매 데이터의 생성과 추천을 30회 반복하여 수행하였으며, 희소 구매 데이터에 기반한 협업 필터링의 수행결과는 <그림 4>와 같다.



<그림 4> 구매 데이터에 기반한 협업 필터링에서의 희소 효과

<그림 4>에서 볼 수 있듯이 구매 데이터의 희소 정도가 커질수록 구매 데이터에 기반한 협업 필터링의 성능이 떨어지는 것을 확인할 수 있으며, 데이터가 희소해 지더라도 임의로 추천하는 경우보다는 성과가 좋았다. <그림 3>과 <그림 4>를 비교해보면, 데이터의 희소 정도가 증가할수록 추천성과가 떨어지는 정도는 구매 데이터의 희소도가 증가하는 경우보다 평가점수의 희소도가 증가하는 경우에 더 급하게 떨어지는 것을 볼 수 있다.

4.2.3 Hybrid 협업 필터링

Hybrid 협업 필터링 방안은 식 (6)과 식 (7)에 나와 있는 것처럼 평가점수에 기반한 협업 필터링 방안과 구매 데이터에 기반한 협업 필터링 방안에 의해 구해진 상관계수들을 가중평균하여 상관계수를 구하고 이를 선호도 점수 예측에 활용하는 방안이다. Hybrid 협업 필터링의 희소효과 분석을 위해, 앞 절에서 제시된 것과 같이 평가점수와 구매 데이터에 기반한 협업 필터링 방안의 희소효과 분석을 위해 사용된 데이터 집합들을 조합하여 활용하였다. 즉, 평가점수를 줄여나간 R_16, R_13, R_10, R_7, R_4와 구매 데이터를 줄여나간 T_1.0, T_0.85, T_0.7, T_0.55, T_0.4, T_0.25 데이터 집합들과, 상관계수의 가중치를 $(\omega, 1-\omega)$ = (0.0, 1.0)에서 (1.0, 0.0)까지 0.1씩 증감하여 데이터 집합 조합마다 가중치를 달리하여 각각 30회씩 반복하여 성과를 측정하였다. 각 방안에서의 추천 성과 결과는 <표 4>에 정리되어 있다. <표 4>에서 명암 표시 된 셀들은 각 평가점수/구매 데이터 조합에서 성과 최고값과 1%이내의 성과치를 갖는 경우이다.

<표 4>에서 같은 행에 있는 값들은 구매 데이터가 같고 평가점수의 희소도가 상이한 경우인데, 같은 행에서는 평가점수의 희소도가 높을수록 추천 성과가 낮아지는 것을 거의 모든 행에서 볼 수 있다. <표 4>에서 같은 열의 경우에는 평가점수의 희소도는 같으나 구매 데이터의 희소도가 상이한 경우인데, 같은 열에서는 구매 데이터의 희소도가 증가하면 추천 성과의 최고값이나 개별값은 약간씩 하락하였다고 볼 수 있으나 추천 성과가 하락하는 정도가 매우 작으며 거의 유사한 수치를 나타낸다. 이는 평가점수의 희소도가 상이한 다른 데이터 집합의 경우에서도 반복적으로 나타난다. 정리하면, Hybrid 협업 필터링에서는 평가점수의 희소도는 추천 성과와 큰 역의 상관관계를 보였으나 구매 데이터의 희소도는 미세한 역의 상관관계 혹은 거의 영향이 없음을 알 수 있다. 이것은 4.2.1절, 4.2.2절에서 평가점수 기반의 협업 필터링에서 데이터 희소도에 대한 추천 성과의 민감도가 구매 데이터의 경우보다 큰 것으로 파악되었는데, Hybrid 협업 필터링에서도 일관된 결과를 보이고 있다. 이러한 현상은 <표 4>를 가시화한 <그림 5>, <그림 6>을 통해서 보다 쉽게 확인할 수 있다. <그림 5>는 평가점수 데이터의 희소 정도를 일정하게 했을 때, 구매 데이터의 희소도와 가중치의 추천 성과에 미치는 관계를 3차원 그래프로 표현한 것이다. 반대로 <그림 6>은 구매 데이터의 희소 정도가 동일한 경우, 평가점수 데이터의 희소도와 가중치가 추

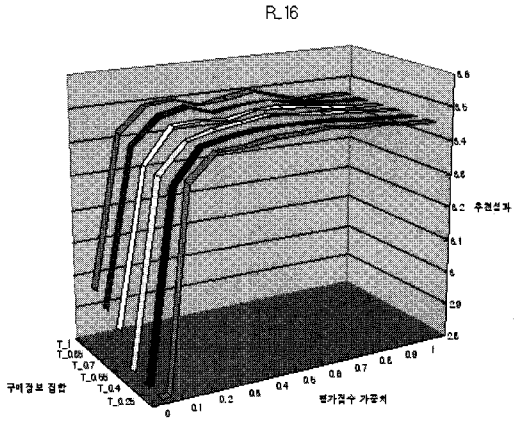
<표 4> Hybrid 협업 필터링에서의 희소 효과

평가/구매비율	R_4-T_1.0	R_7-T_1.0	R_10-T_1.0	R13-T_1.0	R16-T_1.0
0.0-1.0	2.9606	2.9606	2.9606	2.9606	2.9606
0.1-0.9	3.0481	3.1079	3.2581	3.3201	3.4212
0.2-0.8	2.9933	3.1229	3.2708	3.3528	3.4939
0.3-0.7	2.9657	3.1117	3.2703	3.3624	3.503
0.4-0.6	2.9624	3.0966	3.2628	3.3655	3.4758
0.5-0.5	2.9517	3.0907	3.232	3.3593	3.4879
0.6-0.4	2.9712	3.0764	3.1987	3.3551	3.5
0.7-0.3	2.9521	3.0874	3.2185	3.3421	3.4758
0.8-0.2	2.9321	3.0665	3.2318	3.3379	3.4576
0.9-0.1	2.9456	3.0806	3.2341	3.3394	3.4576
1.0-0.0	2.8883	3.06	3.2306	3.3274	3.4545
평가/구매비율	R_4-T_0.85	R_7-T_0.85	R_10-T_0.85	R_13-T_0.85	R_16-T_0.85
0.0-1.0	2.9306	2.953	2.951	2.9352	2.9335
0.1-0.9	3.0298	3.1286	3.2492	3.3007	3.4001
0.2-0.8	3.0054	3.122	3.2393	3.3455	3.4759
0.3-0.7	2.9768	3.1128	3.2554	3.3519	3.497
0.4-0.6	2.9661	3.1152	3.256	3.3483	3.4704
0.5-0.5	2.9524	3.0988	3.2365	3.3595	3.4832
0.6-0.4	2.9319	3.0947	3.2418	3.3501	3.4887
0.7-0.3	2.9075	3.0868	3.2125	3.3592	3.4729
0.8-0.2	2.9505	3.081	3.2398	3.3387	3.4634
0.9-0.1	2.9174	3.0638	3.2322	3.3343	3.4581
1.0-0.0	2.8954	3.0795	3.2242	3.3396	3.4545
평가/구매비율	R_4-T_0.7	R_7-T_0.7	R_10-T_0.7	R_13-T_0.7	R_16-T_0.7
0.0-1.0	2.8844	2.8749	2.8859	2.9009	2.9012
0.1-0.9	2.9964	3.1173	3.2249	3.298	3.3655
0.2-0.8	2.9682	3.1237	3.2452	3.3644	3.4673
0.3-0.7	2.9613	3.0952	3.2608	3.3708	3.4743
0.4-0.6	2.956	3.0944	3.2609	3.3719	3.4675
0.5-0.5	2.9604	3.0923	3.2327	3.3418	3.4849
0.6-0.4	2.9493	3.1067	3.2354	3.3543	3.4841
0.7-0.3	2.9464	3.0773	3.2093	3.326	3.4759
0.8-0.2	2.9293	3.0937	3.2144	3.3338	3.464
0.9-0.1	2.9144	3.093	3.2297	3.3426	3.4579
1.0-0.0	2.8938	3.0723	3.2393	3.3398	3.4545

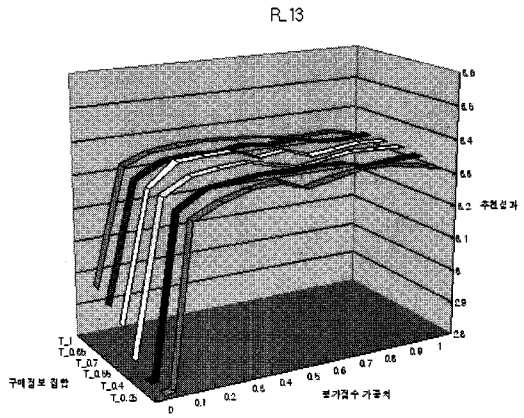
<표 4> 계속

평가/구매비율	R_4-T_0.55	R_7-T_0.55	R_10-T_0.55	R_13-T_0.55	R_16-T_0.55
0.0-1.0	2.814	2.7977	2.8112	2.828	2.8065
0.1-0.9	3.0059	3.1231	3.2291	3.2948	3.3584
0.2-0.8	2.9546	3.1127	3.2362	3.3433	3.4445
0.3-0.7	2.9609	3.0761	3.2625	3.3544	3.4657
0.4-0.6	2.9422	3.0974	3.2502	3.359	3.4766
0.5-0.5	2.9102	3.0983	3.2242	3.3398	3.49
0.6-0.4	2.9509	3.0801	3.2285	3.3222	3.479
0.7-0.3	2.9609	3.0887	3.2207	3.3353	3.4729
0.8-0.2	2.9136	3.0579	3.2386	3.3477	3.4691
0.9-0.1	2.9148	3.0631	3.2349	3.3452	3.46
1.0-0.0	2.8945	3.0703	3.2187	3.3403	3.4545
평가/구매비율	R_4-T_0.4	R_7-T_0.4	R_10-T_0.4	R_13-T_0.4	R_16-T_0.4
0.0-1.0	2.7053	2.6953	2.693	2.6764	2.7099
0.1-0.9	2.96	3.1122	3.2072	3.2775	3.3511
0.2-0.8	2.9224	3.0901	3.2257	3.3287	3.4473
0.3-0.7	2.9246	3.0838	3.2143	3.3347	3.4631
0.4-0.6	2.9341	3.0782	3.2228	3.3456	3.4744
0.5-0.5	2.9126	3.0777	3.2326	3.3427	3.4752
0.6-0.4	2.9007	3.093	3.2196	3.3471	3.4752
0.7-0.3	2.9307	3.1111	3.2362	3.3422	3.4712
0.8-0.2	2.9067	3.0808	3.2283	3.3221	3.4622
0.9-0.1	2.9106	3.0614	3.2275	3.3283	3.4576
1.0-0.0	2.9023	3.0697	3.2268	3.3323	3.4545
평가/구매비율	R_4-T_0.25	R_7-T_0.25	R_10-T_0.25	R_13-T_0.25	R_16-T_0.25
0.0-1.0	2.5361	2.5652	2.5548	2.5455	2.5288
0.1-0.9	2.9198	3.0763	3.1998	3.2745	3.3821
0.2-0.8	2.9101	3.0948	3.2201	3.3161	3.4545
0.3-0.7	2.9026	3.0562	3.2134	3.3377	3.4551
0.4-0.6	2.9153	3.0672	3.2293	3.3452	3.4703
0.5-0.5	2.9021	3.0875	3.2173	3.3151	3.4726
0.6-0.4	2.9104	3.0933	3.2448	3.3347	3.4793
0.7-0.3	2.9159	3.0572	3.2323	3.3454	3.4669
0.8-0.2	2.8923	3.0828	3.2077	3.3467	3.4663
0.9-0.1	2.8825	3.0799	3.2303	3.3332	3.4577
1.0-0.0	2.8824	3.0891	3.1982	3.3183	3.4545

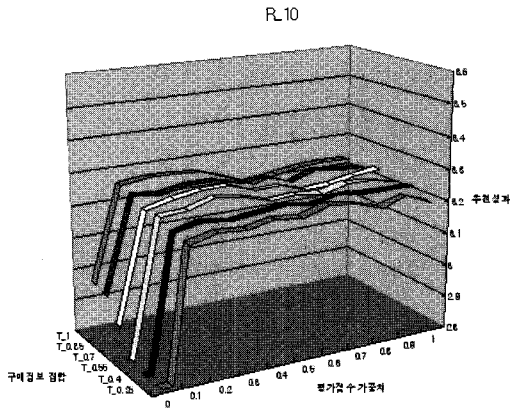
천 성과에 미치는 관계를 3차원 그래프로 표현한 것이다. <그림 5>의 (a), (b), (c), (d), (e)의 비교를 통해서 평가점수 데이터의 희소 정도가 낮아질 수로, 추천 성과가 낮아지는 것을 볼 수 있으며, 평가점수 희소 정도가 동일한 경우에는 구매 데이터의 희소도에 따른 추천 성과의 차이가 크지 않음을 볼 수 있다. 하지만 <그림 6>에서 볼 수 있듯이 구매 데이터의 희소 정도가 동일한 경우에는 평가점수 데이터의 희소도에 따라 상대적으로 추천 성과의 차이가 큰 것을 볼 수 있다.



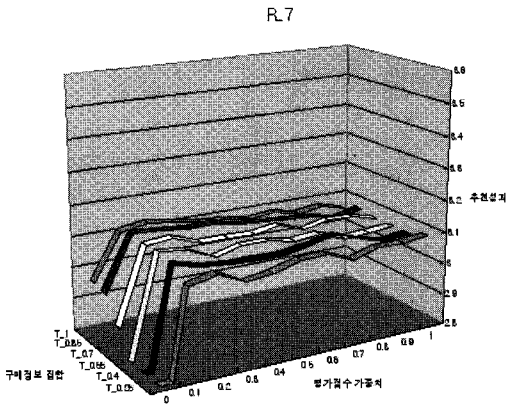
(a) 평가점수 데이터를 모두 사용한 경우(R_16)



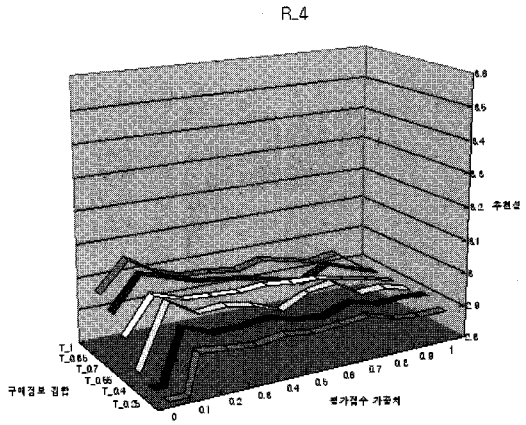
(b) 평가점수 데이터의 희소정도가 R_13인 경우



(c) 평가점수 데이터의 희소정도가 R_10인 경우

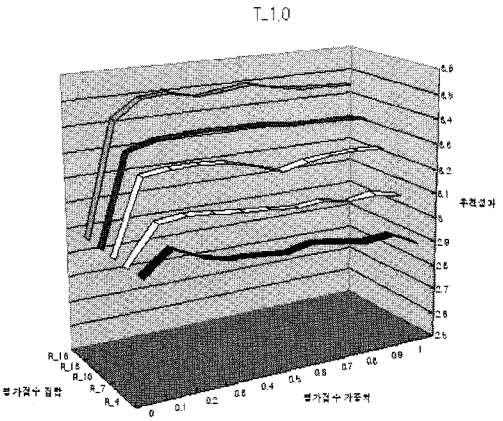


(d) 평가점수 데이터의 희소정도가 R_7인 경우

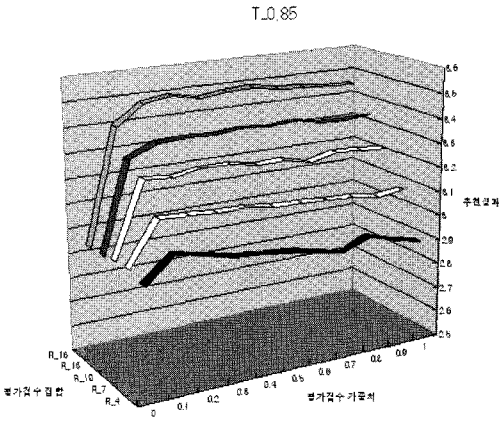


(e) 평가점수 데이터의 희소정도가 R_4인 경우

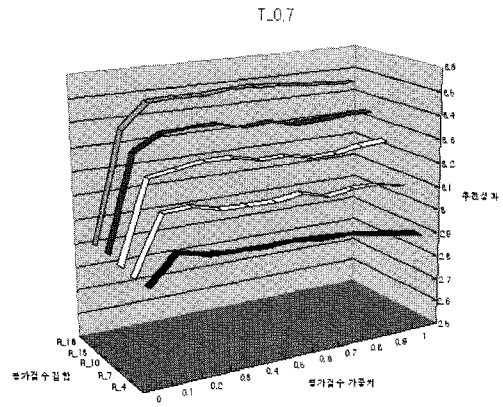
<그림 5> 평가점수 데이터의 희소도가 일정한 경우, 구매 데이터의 희소도와 가중치의 추천 성과와의 관계



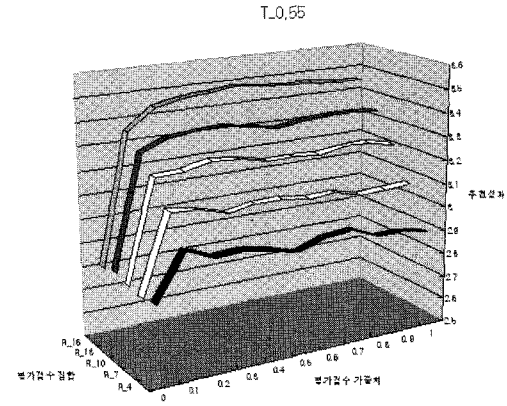
(a) 구매 데이터를 모두 사용한 경우(T_1.0)



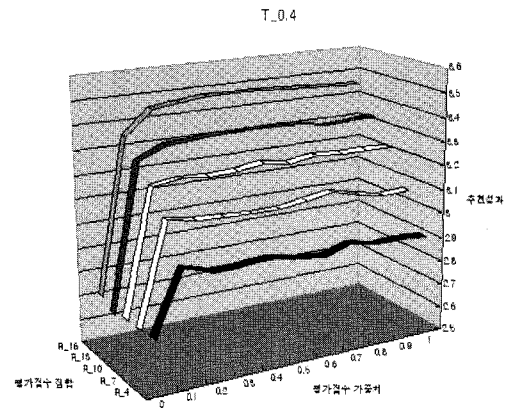
(b) 구매 데이터의 희소정도가 T_0.85인 경우



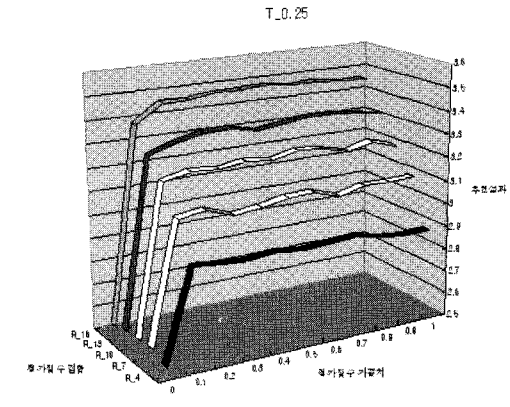
(c) 구매 데이터의 희소정도가 T_0.7인 경우



(d) 구매 데이터의 희소정도가 T_0.55인 경우



(f) 구매 데이터의 희소정도가 T_0.25인 경우



(e) 구매 데이터의 희소정도가 T_0.4인 경우

<그림 6> 구매 데이터의 희소도가 일정한 경우, 평가점수 데이터의 희소도와 가중치의 추천 성과와의 관계

두 상관계수에 대한 가중치가 추천 성과에 미치는 영향을 살펴보면, 구매 데이터에서 계산된 상관계수에 대한 가중치가 높을 때 더 좋은 성과를 보이는 경우가 많았으나, 일반적인 패턴을 보이지는 않는다. 하지만, 평가점수 희소도가 높은 경우에는 구매 데이터에서 구해진 상관계수의 가중치를 높이는 것이 더 좋은 성과를 나타내는 것을 <표 4>의 첫번째 열(R_4에 해당하는 열)의 음영 표시와 <그림 5>의 (e)에서 확인할 수 있다. 하지만, <표 4>의 마지막 6번째 블록(T_0.25 블록)과 <그림 6>의 (f)에서 볼 수 있듯이, 구매 데이터의 희소도가 높은 경우에는 추천 성과가 상관계수의 가중치에 큰 영향을 받지 않는다.

V. 논의 사항

5.1 데이터의 희소 정도와 협업 필터링의 성능

데이터의 희소 정도와 협업 필터링의 성능관계는 희소도가 높을수록 협업 필터링의 성과가 역의 상관관계를 갖는다. <그림 3>, <그림 4>, <표 4>에서 보듯이 데이터에 따라 감소하는 정도의 차이는 있지만 성과가 지속적으로 감소한다. 평가점수의 희소도가 증가할수록 추천 성과가 매우 가파르게 감소하는 것을 <그림 3>에서 볼 수 있으며, <표 4>에서 같은 행의 추천 결과를 보아도 평가점수의 희소도가 높을수록 추천 성과는 크게 감소하였다. 구매 데이터의 경우에는 총 1078권의 책 중에서 평가자 집단에 속한 실험자는 12.3권의 책을, 평가자 집단에 포함되지 않은 실험자는 10.5727권의 책을 평균적으로 구매하였기 때문에 데이터를 모두 활용하는 경우(T_1.0)에서도 이미 희소도가 매우 높은 상태로 볼 수 있다. 따라서 구매 데이터에 기반한 협업 필터링의 경우 절반 정도의 구매 데이터를 줄여가도(T_0.85, T_0.7, T_0.55) 추천 성과의 감소폭이 높지 않았으며, Hybrid 협업 필터링의 경우에도 이미 완전 구매 데이터의 희소도가 높은 상태이기 때문에 구매 데이터의 희소도를 인위적으로 증가 시켜도 추천성과에 미세한 영향만을 미치는 것으로 파악되었다.

5.2 평가점수와 구매 정보의 동시 활용의 효과

Hybrid 협업 필터링 방안을 통해 평가점수와 구매 정보를 동시에 활용하여 추천을 수행한 결과는 <표 3>처럼 구매 데이터만을 활용하는 경우보다는 추천 성과가 좋은 것으로 파악되었으며, 평가점수에 기반한 협업 필터링과 비교해서는 추천 성과는 미세하게 항상 높으나, 그 차이가 통계적으로는 유의하지는 않았다. <표 4>를 보면 평가점수의 희소도가 낮은 경우(R_16, R_13)에는 구매 정보를 활용하는 비율이 낮아도 최고의 성과를 보이는 데이터 집합이 있지만 평가점수의 희소도가 높은 경우(R_7, R_4)에는 구매 정보를 활용하는 비율이 높아야만 최고의 성과를 보여준다. 이는 평가점수의 희소도가 높을수록 구매 정보를 많이 활용하는 것이 Hybrid 협업 필터링 방안에서 더욱 유용하다는 것이며, 구매 데이터의 희소도가 높아지면(T_0.4, T_0.25) 이러한 경향이 약간 흐트러져서 중간 비율의 영역이 최고의 성과를 나타내기는 하지만 평가점수 희소도가 높을수록 구매 정보를 많이 활용하는 것이 좋은 추천 성과를 나타내는 경향을 지속적으로 보이고 있다. 평가점수의 희소도가 적을수록 구매 정보의 활용이 가져오는 추천 성과의 향상은 적어진다고 볼 수 있다.

5.3 구매 이력이 많은 경우, 평가점수의 유용성

구매 이력의 차이에 따라 평가점수의 유용성

이 달라지는지에 대해 비교하기 위하여 다음과 같은 추가 분석을 수행하였다. 구매권수에 따라 평가자 집단을 제외한 실험자를 3개의 그룹으로 나누었다. 처음 그룹은 구매권수가 13권 이상인 그룹으로 실험자중에 29명이 속한다. 두 번째 그룹은 구매권수가 8권 이상 12권 이하인 그룹으로 38명이 속하며, 마지막 그룹은 구매권수가 7권 이하인 그룹으로 실험자 중 43명이 속한다. 평가자 집단을 제외한 실험자를 위와 같이 세 그룹으로 나누어 앞에서 제시되었던 3가지 협업 필터링 방안을 통해 추천을 수행하여 그룹간의 추천 성과에 대한 차이가 있는지 알아보았다. 그룹 간의 추천 성과를 비교한 결과를 <표 5>에 정리하였다.

모든 방안들에서 구매권수에 따라 구분된 그룹들의 추천성과는 같다고 볼 수 있다. 구매이력이 작은 그룹의 추천성과가 전체적으로 작게 나타났으나 통계적으로 차이가 나지는 않았다. 같은 그룹에서 추천방안간의 추천성과 비교에 대한 분석에서는 구매이력이 적은 그룹 2와 그룹 3에서는 평가점수를 활용한 협업필터링과 Hybrid 협업필터링 방안이 좀 더 성과가 좋은 것으로 나타났으나, 구매이력이 많은 그룹 1에서는 구매데이터를 활용한 협업필터링의 성과가 낮기는 하지만 세가지 방안의 추천성과가 통계적으로 차이가 있다고 볼 수 없다.

Ⅵ. 결 론

본 연구에서는 인터넷 상점에서 개인화 추천을 위해서 가장 많이 활용되고 있는 협업 필터링 기법에 있어서 데이터의 희소 정도(sparsity)가 추천 성능에 미치는 영향을 분석하였다. 상품에 대한 평가점수만을 활용하는 경우, 고객의 구매 데이터만을 활용하는 경우, 이 두 가지 정보를 함께 활용하는 Hybrid 방안 등 3가지 형태의 협업 필터링에 대하여, 평가점수의 희소 정도, 구매 데이터의 희소 정도의 영향을 분석하였다. 인터넷 서점에 대한 실험 데이터를 토대로 분석한 결과, 평가점수에 기반한 협업 필터링에서 평가점수의 희소도와 추천 성과간에는 큰 역의 상관관계를 나타냈으며, 구매 데이터를 활용한 협업 필터링 방안에서도 구매 데

<표 5> 구매 이력에 따른 추천 성과의 비교

	구매이력에 따른 그룹분류			ANOVA (F-value)	Duncan Test
	그룹 1	그룹 2	그룹 3		
구매권수 (N=실험자수)	13권 이상 (N = 29)	8권 이상 12 이하 (N = 38)	7권 이하 (N = 43)		
(1) 구매데이터만 활용 시	3.1839 (0.9417)	3.0614 (1.11999)	2.7209 (0.91412)	2.165	-
(2) 평가점수만 활용 시	3.5172 (0.74847)	3.6404 (0.70767)	3.2481 (0.85787)	2.677	-
(3) Hybrid 활용 시 최고 값	3.5977 (0.66892) (평가 0.2, 구매 0.8)	3.7982 (0.67357) (평가 0.6, 구매 0.4)	3.3643 (0.93948) (평가 0.1, 구매 0.9)	3.076	-
ANOVA (F-value)	2.212	7.768*	6.179*		
Duncan Test	-	(1) < (2), (3)	(1) < (2), (3)		

* 유의 수준 α = 0.05에서 통계적으로 유의

이터의 희소도와 추천 성과간의 역의 상관관계를 보였으나 그 정도는 평가점수의 경우보다 적었다. 평가점수의 희소도나 구매정보의 희소도에 관계없이 언제나 평가 정보와 구매 정보를 함께 활용하여 추천하는 Hybrid 협업 필터링 방안의 추천 성과가 좋았으나, 평가점수만을 활용한 경우와 큰 차이를 보이지는 않았다. Hybrid 협업 필터링의 성능 분석에서 평가점수의 희소도가 높을수록 구매 정보를 많이 활용하여야만 좋은 추천 성과를 보이며, 평가점수의 희소도가 낮은 경우에는 구매 정보의 활용이 추천 성과에 많은 영향을 미치지 못하였다.

협업 필터링 적용 시에 고객에게 다수 상품들에 대한 평가점수를 입력 받아야 하는 수고를 배제하기 위해서 구매 데이터만을 활용한 협업 필터링의 사용이 제안되고 있지만, 본 연구 결과에서 평가점수를 활용하는 경우가 구매 데이터만 사용하는 경우보다 추천 성과에 있어서 통계적으로 유의하게 좋은 것을 확인할 수 있었다. 따라서 고객으로부터 평가점수를 획득하여 개인화 추천을 하는 것이 추천의 성과를 높이는데 바람직하며, 평가점수의 희소 정도가 추천 성과에 많은 영향을 주기 때문에 평가점수의 결측치를 최대한 줄이는 노력이 함께 필요하다.

본 연구에서 사용한 실험 데이터는 실험적목적으로 제한적으로 수집된 데이터 집합으로써, 평가 데이터나 구매 데이터의 희소도에 있어서 실제 전자상거래 데이터와는 차이점을 갖는다. 즉, MovieLens 평가 데이터의 경우 희소도는 93.69%, Fingerhut Inc. 전자상거래 사이트의 구매 데이터는 99.94%에 이른다[Sarwar et al. 2000]. 따라서, 본 연구의 결과를 일반화하는데 주의할 필요가 있으며, 추후 연구 과제 중에 하나로, 실제적인 데이터 집합에 대하여 희소 정도와 추천 성과에 대한 추가적인 분석이 필요하다. 또한 본 연구의 연장선 상에서 협업 필터링의 실제적인 활용에 도움을 주기위해서 다음과 같은 연구가 추가적으로 필요하다. 첫째, 고객으로부터 초기에 평가점수를 입력 받아야 하는 최적의 상품 개수를 결정하기 위하여, 평가점수의 개수와 추천 성능간의 관계를 파악하는 것이 필요하다. 둘째로, 새로운 상품에 대하여 평가해주어야 하는 평가자의 수를 합리적으로 결정하기 위한 최적 평가자 수에 대한 연구도 함께 필요하다. 셋째로 본 연구에서는 인터넷 상점을 중심으로 평가점수와 구매 데이터의 추천 성과와의 관계를 살펴보았는데, 인터넷 정보 제공 사이트의 경우에는 정보의 내용 기반 추천이 가능한데, 이러한 기법과 협업 필터링의 성과 차이 및 효과적인 공동 활용 방안에 대한 연구가 필요하다.

〈참 고 문 헌〉

[1] 김재경, 서지혜, 안도현, 조윤호, "협업 필터링 기법을 활용한 개인화된상품 추천 방법론 개발에 관한 연구," *한국지능정보시스템학회논문지*, 제8권 제2호, 2002년 12월, pp. 139-157.

[2] 김종우, 이경미, "인터넷 상점에서 개인화 광고를 위한 장바구니 분석 기법의 활용," *경영과학*, 제17권 제3호, 2000년 11월, pp. 19-30.

[3] 이건창, 정남호, "데이터 마이닝 기법과 지능형 에이전트 기법을 결합한 인터넷 쇼핑몰의 설계 및 구현에 관한 연구," *정보기술응용연구*, 제1권 제2호, 1999년 10월, pp. 113-137.

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

[5] 황병연, "개선된 추천을 위해 클러스터링을 이용한 협동적 필터링 에이전트 시스템의 성

능," *정보처리논문지*, 제7권, 제5S호, 2000년 5월, 2000, pp. 1599-1608.

[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. 37, August 2000, pp. 363-375.

[8] Autonomy, http://www.autonomy.com, 2004.

[9] Balabanovic, Marko and Yoav Shoham, "Content-based, Collaborative Recommendation," *Communication of ACM*, Vol. 40, No. 3, March 1997, pp. 66-72.

[10] BeFree, http://www.befree.com, 2004.

[11] Breese, John S., David Heckerman, and Carl Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Technical Report*, MSR-TR-98-12, Microsoft Research, October 1998.

[12] BroadVision, http://www.broadvision.com, 2004.

[13] Cho, Yoon Ho, Jae Kyeong Kim, and Soung Hie Kim, "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.

[14] Dragon, Richard V., "Recommendation Systems - Advice From the Web," *PC Magazine*, September 9, 1997.

[15] 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.

[16] Gustos, http://www.gustos.com, 2004.

[17] Kim, Jae Kyeong, Yoon Ho Cho, Woo Ju Kim, Je Ran Kim, and Ji Hae Suh, "A Personalized Recommendation Procedure for Internet Shopping Support," *Electronic Commerce Research and Applications*, Vol. 1, No. 3/4, 2001, pp. 301-313.

[18] Kim, Jong Woo, Byung Hun Lee, Michael J. Shaw, Hsin-Lu Chang, and Mathew Nelson, "Application of Decision Tree Induction Techniques to Personalized Advertisements on Internet Storefront," *International Journal of Electronic Commerce*, Vol. 5, No. 3, Spring 2001, pp. 45-62.

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

[20] Lee, Wei-Po, Chih-Hung Liu, and Cheng-Che Lu, "Intelligent Agent-based Systems for Personalized Recommendation in Internet Commerce," Expert Systems with Applications, Vol. 22, No. 4, 2002, pp. 275-284.

[21] Mild, Andreas and Martin Natter, "Collaborative Filtering or Regression Models for Internet Recommendation Systems," *Journal of Targeting, Measurement and Analysis for Marketing*, Vol. 10, No. 4, Jun, 2002, pp. 304-313.

[22] Mobasher, Bamshad, Robert Cooley, and Jaideep Srivastava, "Automatic Personalization Based on Web Usage Mining," *Communication of the ACM*, Vol. 43, No. 3, 2000, pp. 142-151.

[23] Mulvenna, M.D., S.S. Anand, and A.G. Buchner, "Personalization on the Net

Using Web Mining," *Communication of the ACM*, Vol. 43, No. 8, August, 2000, pp. 122-125.

[24] Net Perception, http://www.netperceptions.com, 2004.

[25] Resnick, P. N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," *Proceedings of the ACM 1994 Conference on Computer Supported Cooperative Work*, New York, ACM, 1994, pp. 175-186.

[26] Sarwar, Badrul, George Karypis, Joseph Konstan, and John Riedl, "Analysis of Recommendation Algorithms for E-Commerce," ACM SIG EC'00, 2000.

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

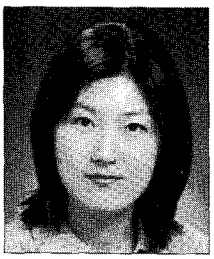
[28] Shardanand, U. and P. Maes, "Social Information Filtering: Algorithms for Automating 'Word of Mouth'," *Proceedings of Conference on Human Factors in Computer Systems*, 1995, pp. 210-217.

[29] Turban, Efraim, David King, Jae K Lee, and Dennis Viehland, Electronic Commerce 2004: A Managerial Perspective, Prentice Hall, 2003.

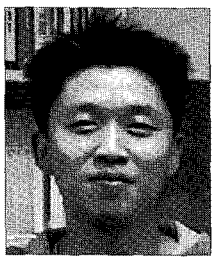
[30] Ungar, Lyle H. and Dean P. Foster, "Clustering Methods for Collaborative Filtering," AAAI-98 Workshop on Recommendation Systems, Madison, WI, July 1998, pp. 114-129.

◆ 저자소개 ◆

김종우 (Kim, Jong Woo)
현재 한양대학교 경영학부 부교수로 재직 중이다. 서울대 수학과에서 이학사(1989), 한국과학기술원 경영과학과에서 공학석사(1991)를 취득하고 한국과학기술원 산업경영학과에서 공학박사(1995)를 취득하였다. 한국과학기술원 경영정보연구센터 연수연구원, University of Illinois at Urbana-Champaign 방문연구원, 충남대학교 통계학과 부교수로 근무한 경력이 있다. 주요 관심분야는 경영정보시스템, 의사결정지원시스템, 전자상거래, 데이터 마이닝 응용, 지식경영시스템 등이다.



배세진 (Bae, Se Jin)
충남대학교 통계학과와 동대학원을 졸업하였으며, 관심분야는 CRM, 추천, 정보시스템 등이다.



이홍주 (Lee, Hong Joo)
한국과학기술원(KAIST) 산업경영학과를 졸업(1997)하고 KAIST 테크노경영대학원 경영공학과정에서 석사(1999)학위를 취득하였으며, 현재 박사과정에 재학 중이다. 주요 관심분야는 지식경영지원시스템, Semantic Web, 가상 협업시스템, CRM 등이다.

◆ 이 논문은 2004년 1월 30일 접수하여 1차 수정을 거쳐 2004년 4월 12일 게재확정되었습니다.