개인화된 상품추천을 위한 협동적 필터링에서의 데이터 선정과 추천 성과간의 관계

Relationship between Data Selection and Prediction Performance in Collaborative Filtering

이홍주*, 김종우**, 박성주*
* 한국과학기술원, 테크노경영대학원
** 한양대학교 경영학부

Abstract

전자상거래와 고객관계관리에서 고객의 개인화를 위해 사용되는 협동적 필터링 방안은 고객이 상품에 대해 표시한 선호도에 기반을 두어 선호도가 유사한 사용자를 찾고, 유사한 사용자의 선호도를 활용하여 추천할 상품을 선정하는 방안이다. 고객간의 유사도 계산과 상품에 대한 선호도 계산을 위한 다양한 방안들의 계산식에 대해서는 명확하게 정의되어 있으나, 이에 활용되는 데이터의 선정에 대해서는 명확한 규정이나 가이드라인이 존재하지 않는다. 즉, 몇 번 이상의 선호도를 표시한 사용자를 대상으로 추천을 수행할 것인지, 혹은 몇 번 이상 선호도가 표시된 상품을 추천에 활용할 것인지와 같은 데이터 선정에 활용되는 계수와 협동적 필터링의 추천 성과간의 관계에 대한 연구는 아직 부족하다.

본 연구에서는 협동적 필터링의 연구에 많이 활용되는 EachMovie 데이터를 가지고 협동적 필터링의 계수와 추천 성과간의 관계에 대해 실험적으로 연구하였다. 첫 번째는 몇 번 이상 선호도를 표시한 사용자를 협동적 필터링에 활용하는 것이 추천 성과를 높일 수 있는지에 대해 연구하였으며, 두 번째는 몇 번 이상 선호도가 표시된 상품을 고객에게 추천하는 것이 협동적 필터링의 추천 성과를 높일 수 있는가에 대한 연구를 수행하였다. 계수와 추천 성과간의 관계에 대한 두 가지 실험에서 선호도 표시의 한계가치(marginal value)가 점진적으로 감소하는 것을 볼 수 있었다. 본 연구의 결과는 협동적 필터링의 수행을 위한 효과적인 데이터의 선정에 도움을 줄 수 있을 것이다.

1. 서론

전자상거래에서 고객에게 적합한 혹은 관심 있어 할 만 한 상품을 개인화하여 제공하는 것은 고객의 상품 검색노력을 줄여 줄 뿐만이 아니라, 적합한 상품 추천으로 인해 상거래 사이트에 대한 고객의 충성도 제고에도 도움을 주기 때문에 고객관계관리 측면에서도 중요하게 인식되고 있다 (Kim et al., 2002; Mild and Natter, 2002). 현재 상용화된 추천 시스템들에서 가장 많이 활용되고 있는 추천 기법은 협동적 필터링 (Collaborative Filtering, CF)이며, 해당 고객과 제품에 대한 선호도가 유사한 고객들의 제품에 대한 평가점수를 활용하여 고객에게 적합한 상품 정보를 제공하는 기법이다 (Breese et al., 1998; Resnick et al., 1994).

협동적 필터링은 사용자가 상품에 대해 표시한 선호도나 구매 데이터를 기반으로 상관계수(correlation coefficient)나 코사인(cosine) 척도를 통해 사용자간의 유사도를 계산한 후에, 상품별로 예측 선호도를 생성하여 가장 높은 예측 선호도를 가진 상품들을 추천하는 절차를 거친다 (Sarwar et al., 2000). 고객간의 유사도 계산과 상품에 대한 선호도 계산을 위한 다양한 방안들의 계산식에 대해서는 명확하게 정의되어 있으나, 이에 활용되는 선호 데이터의 선정에 대해서는 명확한 규정이나 가이드라인이 존재하지 않는다. 즉, 몇 번 이상의 선호도를 표시한 사용자를 대상으로 추천을 수행할 것인지 혹은 몇 번 이상 선호도가 표시된 상품을 추천에 활용할 것인지와 같은 데이터 선정에 활용되는 계수와 협동적 필터링의 추천 성과 간의 관계에 대한 연구는 수행되지 않았다.

본 연구에서는 협동적 필터링의 연구에 많이 활용되는 EachMovie 데이터를 가지고 협동적 필터링에 활용되는 데이터 선정을 위한 계수와 추천 성과간의 관계에 대해 실험적으로 연구하였다. 첫 번째는 몇 번 이상 선호도를 표시한 사용자를 협동적 필터링에 활용하는 것이 추천 성과를 높일 수 있는지에 대해 연구하였으며, 두 번째는 몇 번 이상 고객들에 의해서 선호도가 표시된 상품을 고객에게 추천하는 것이 협동적 필터링의 추천 성과를 높일 수 있는가에 대한 연구를 수행하였다. 2장에서 협동적 필터링의 데이터 선정에 관한 관련 문헌에 대해 살펴보며, 3장에서 협동적 필터링에 활용되는 계수와 추천성과에 과한 실험방안과 실험결과를 소개한다. 4장에서 협동적 필터링의 계수 선택과 관련된 논의사항을 다루고, 5장에서는 결론을 제시한다.

2. 관련문헌

고객간의 유사도를 상관계수 형태로 구하는 계산하는 식은 (식1) 과 같다. (식1)은 고객 i, j 간의 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 경우에는 -1에 가까운 값을 가지게 된다.

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

상품에 대한 고객의 선호도 점수 예측은 다음 (식2)를 통해서 이루어진다. (식2)는 고객 i의 상품 k에 대한 선호도 점수인 P_{ik}를 예측하는 식으로, *Rater(k)*는 상품 k를 평가한 고객의 집합을 의미한다.

$$P_{ik} = \bar{S}_i + \frac{\sum_{l \in Rater(k)} (S_{lk} - \bar{S}_l) r_{il}}{\sum_{l \in Rater(k)} | r_{il} |} \quad (식2)$$

(식1)과 (식2)에서 알 수 있듯이 유사도 계산과 예측치를 계산하는 식은 명확하지만, 활용되는 데이터의 속성이나 계수에 대한 어떠한 제한이나 선택에 대한 규정은 없다. 이로 인해 많은 연구자들이 다양한 협동적 필터링 방안에서의 계수와 추천 성과 간의 관계에 대하여 연구하였으며 이들이 연구한 계수와 추천 성과 간의 관계를 정리한 것이 <Table 1>이다.

사용자 기반의 협동적 필터링에서 가장 문제시 되고 있는 부분은 사용되는 데이터의 희소성(sparsity)과 필터링 알고리즘의 확장성(scalability)으로 보고 있기 때문에 (Sarwar et al., 2001), 이와 관련된 계수들의 선택에 대한 연구들이 많이 진행되었다. 알고리즘의 확장성 문제로 인해 모든 고객들의 데이터를 활용하여 추천을 수행하는 것보다 일정한 수의 유사한 고객만을 활용하여도 추천 성과가 많이 떨어지지 않는 최적 유사 사용자(optimal number of neighbors)에 대한 연구들이 많이 수행되었다 (Mild & Natter, 2002; Sarwar et al, 2000; Sarwar et al., 2001).

<Table 1> 협동적 필터링에서의 계수 선택

	데이터 집합 사용자 수, 상품 수	사용자의 최소 선호도 표시 수	상품의 최소 피 선호도 표시 수	유사 사용자 (Neighbor)의 수	추천 상품 수
Mild & Natter (2002)	EachMovie 61,007명, 419개	3개	50회	10-80명 (최적)	-
Sarwar et al. (2000)	MovieLens 943명, 1682개	20개	1회	80-120명 (최적)	10개
Sarwar et al. (2001)	MovieLens 943명, 1682개	20개	1회	30명 (최적)	-
Ansari et al. (2000)	EachMovie 2000명, 340개	1개	1회	-	-
Breese et al. (1998)	EachMovie 4119명, 1623개	-	-	-	-

또한, 모든 상품에 관한 데이터를 활용하여 유사도 계산과 추천을 수행할 경우에는 많은 계산 노력이 필요하기 때문에, 일정한 수의 상품에 관한 데이터만으로 추천을 수행하여도 추천 성과가 많이 하락하지 않는다는 연구도 수행되었다 (Sarwar et al., 2001).

그러나 많은 연구들이 원래의 데이터 집합에서 실험에 활용할 데이터 집합을 추출하면서 몇 번이상의 선호도를 표시한 사용자를 활용할 것인지, 몇 번 이상 선호도가 표시된 상품을 활용할 것인지에 대한 선택방안이나 이러한 선택이 협동적 필터링의 추천 성과에 어떠한 영향을 미치는지에 대한 연구는 수행되지 않았다. 또한 몇 개의 상품을 고객에게 추천하였고, 이들의 추천 성과를 분석하였는지에 대한 언급도 하지 않고 있는 실정이다. <Table1>에서 볼 수 있듯이 대부분의 연구들이 자의적으로 계수를 선택하여 데이터를 선정하거나 어떠한 기준으로 실험용 데이터 집합을 선정하였지에 대해 언급조차 하지 않고 있는 실정이다.

3. 협동적 필터링에서의 데이터 선정과 추천 성과 간의 관계

본 연구에서는 협동적 필터링에서의 데이터 선정과 추천 성과 간의 연관관계를 파악하기 위하여, 협동적 필터링 연구에서 많이 활용되는 EachMovie 데이터를 활용하였다 (Breese et al., 1998; Mild and Natter, 2002). EachMovie 데이터 집합은 72,916명의 사용자가 1,628개의 영화에 대해 2,811,983회의 선호도를 표시한 것으로 이루어져 있다. 선호도 표시는 0점부터 5점까지 6단계로 이루어져 있다.

3.1 사용자의 선호도 표시 수와 협동적 필터링 성과간의 관계에 관한 실험

실험의 수행을 위해 전체집합에서 4개 이상의 영화에 선호도를 표시한 사용자들 중 무작위로 10%를 선정하여 실험에 활용하였다. 무작위로 선정된 데이터 집합은 모두 5,748명의 사용자가 249,840의 선호도를 표시한 집합이며, 사용자의 선호도 표시회수 평균은 43.37회, 표준편차는 51.70이다. 선정된 데이터 집합을 사용자의 선호도 표시회수를 기준으로 <Table 2>와 같이 10개의 그룹으로 구분하였다. 데이터의 희소수준은 $1 - \frac{Nonzero\,Entries}{Total\,Entries}$로 정의되며 (Sarwar et al., 2001), 여기서 $Total\,Entries$는 (사용자 수X상품의 수)이며 $Nonzero\,Entries$는 사용자들의 총 투표수이다.

각 그룹의 선호도 표시 데이터의 70%를 무작위로 선정하여 사용자간의 유사도 계산에 활용하며, 나머지 30%의 데이터에 대해 선호도를 예측하였다. 협동적 필터링의 성과 측정방안으로는 연구에서 많이 활용되는 방안인 예측치와 실제 선호도간의 차이에 기반을 둔 Mean Absolute Error (MAE)와 Root Mean Squared Error (RMSE)를 활용하였으며, 두 지표는 다음과 같은 식을 통해 계산된다.

$$MAE = \frac{\sum_{i=1}^{N} | p_i - q_i |}{N} \quad (식3)$$

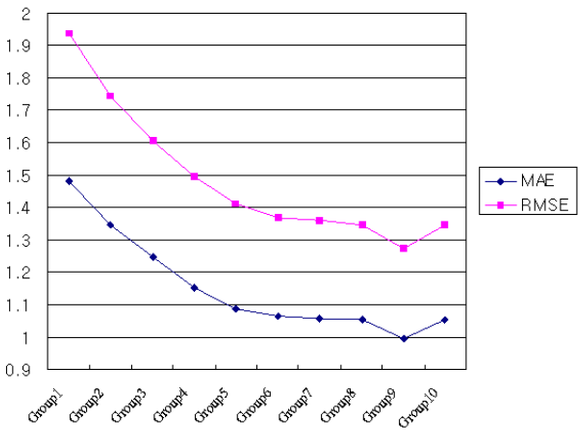
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (p_i - q_i)^2}{N}} \quad (식4)$$

N은 예측하여야하는 상품의 총 수 이며, p_i 는 i상품에 대해 사용자가 입력한 실제 선호도 값이며 q_i는 i상품에 대한 협동적 필터링에 의한 선호도 예측치이다.

<Table 2> EachMovie 데이터 집합의 그룹화

그룹	투표수 범위	투표수 평균	편차	사용자 수	총투표	희소 수준
1	4 - 6	4.8547	0.7662	606	2942	0.9970
2	7 - 9	7.9822	0.8199	505	4031	0.9951
3	10 - 14	11.8134	1.4691	643	7596	0.9927
4	15 - 20	17.3996	1.7331	588	10231	0.9893
5	21 - 26	23.5579	1.6763	509	11991	0.9855
6	27 - 34	30.2385	2.3126	524	15845	0.9814
7	35 - 46	40.0608	3.5981	608	24357	0.9755
8	47 - 63	54.4125	4.8393	572	31124	0.9666
9	64 - 94	77.5317	9.1004	568	44038	0.9524
10	95 - 817	156.3456	77.1566	625	97716	0.9039

각 그룹의 데이터를 실험치와 예측치로 무작위로 구분하여 협동적 필터링을 수행하였으며, 각 그룹마다 30회씩 반복 수행하였으며 결과는 <Figure 1>과 같다.



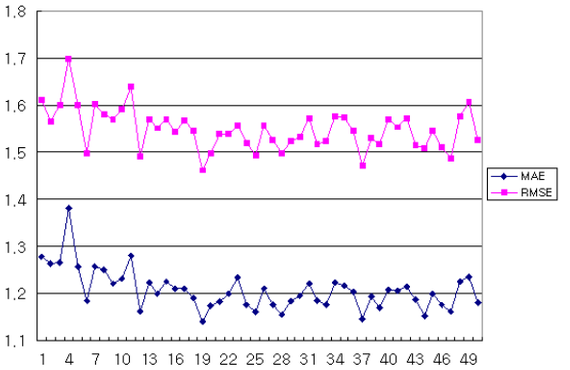
<Figure 1> 선호도 표시 수와 추천 성과 간의 관계

<Figure 1>에 표시된 MAE와 RMSE값은 30회 반복 수행을 통해 나온 MAE, RMSE 결과의 평균치이다. 그룹1에서 그룹 5, 6 까지는 MAE, RMSE 값이 모두 단조 감소하는 것을 볼 수 있으며, 그룹 6,7,8,10은 매우 유사한 값을 나타내며 그룹 9에서 가장 낮은 결과를 보였다.

3.2 사용자의 선호도 표시 수에 따른 최적 초기 추천 시점에 관한 실험

사용자의 선호도 표시 수에 따른 협동적 필터링의 최적 초기 추천 시점에 관한 실험의 수행을 위해 EachMovie 전체자료 집합에서 무작위로 사용자 2.5%를 추출하였다. 무작위로 선정된 데이터 집합은 1,437명의 사용자가 66,634회의 선호도를 표시한 집합이며(희소수준 0.9715), 사용자의 선호도 표시회수 평균은 46.37회, 표준편차는 50.64이다. 선정된 실험집합의 사용자중 선호도표시회수가 110회 이상 130회 이하인 사용자 93명을 선출하여, 선호도 예측을 위한 사용자집합으로 삼았다. 실험은 예측을 위한 사용자 집합의 사용자 선호도 자료 중에서 무작위로 i ($1 <= i <= 100$) 개의 선호도표시 정보를 선정하여 이를 다른 사용자들과의 유사도 계산과 다른 상품에 대한 선호도를 예측하는 것에 활용하였다. 한 사용자의 선호도 표시 자료 중 i개에 해당하지 않는 자료 10개를 선정하여 예측치와 실제 선호도 간의 오차를 계산하였으며, MAE와 RMSE를 성과 측정의 지표로 활용하였다. 실험은 각 i의 값마다 30회씩 반복 수행되었다.

<Figure 2>는 위 실험의 결과이며, i값이 1에서 50까지인 경우의 결과만을 나타내고 있다. 선호도를 1개에서 11개까지 활용하여 예측하였을 때에는 경우에 따라 값이 작기는 하지만 MAE가 1.2 이상의 값을 보이며, 12개에서 18개까지를 활용하여 예측하였을 때에는 경우에 따라 값이 작기는 하지만 MAE가 1.2 초반의 범위에 위치하고 있는 것을 볼 수 있다. 19개 이상을 활용하여 예측하였을 경우에는 대체로 MAE가 1.2이하이며 약간씩 1.2를 넘고 있다. 이는 50개부터 100개까지의 선호도를 가지고 예측을 수행한 경우에도 오차가 큰 경우가 있기는 하지만 오차가 위의 범위 안에 위치하였다. <Figure 2>에서 보면, 사용자의 선호도 표시 수가 12~13개를 넘으면 추천 성과의 변동폭이 안정적이 되며, 또한 선호도 표시 수가 더 늘어나도 추천 성과는 크게 개선되지 않는 것을 알 수 있다.

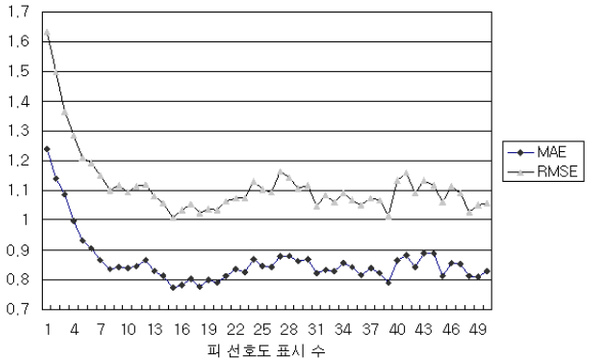


<Figure 2> 선호도 표시 수에 따른 최적 초기 추천 시점에 관한 실험

3.3 상품의 피 선호도 표시 수와 협동적 필터링 성과 간의 관계에 관한 실험

본 실험의 수행을 위해 EachMovie 전체집합에서 무작위로 5%의 사용자를 선정하여 활용하였다. 무작위로 선정된 데이터 집합은 모두 2,874명의 사용자가 134,197회의 선호도를 표시한 집합이며(희소수준 0.9713), 사용자의 선호도 표시회수 평균은 46.69회, 표준편차는 53.74이다. 영화에 대해 사용자가 선호도를 표시한 최소 회수는 1회이며, 최대 회수는 1,843회다. 실험집합을 무작위로 70%를 선정하여 사용자간의 유사도 계산과 나머지 30%에 대한 예측치 계산에 활용하였으며, 30회 반복 수행하였다. MAE와 RMSE를 성과 지표로 활용하였으며, 상품의 피 선호도 표시 수를 1개부터 최대 회수까지 늘려가며 상품에 대한 예측치와 실제 선호도차이를 비교하여 MAE와 RMSE를 계산하였다.

<Figure 3>의 실험결과를 보면 피 선호도 표시 수가 8회까지 증가하는 경우에는 추천 성과가 지속적으로 증가하였으며, 8회에서 12회까지는 유사한 성과를 보이다가 15회까지 지속적으로 성과가 증가하는 것을 확인할 수 있다. 15회 이후에는 추천 성과가 MAE 측면에서 0.8에서 0.9사이의 구간에서 지속적으로 값이 변화한다. <Figure 3>은 피 선호도 표시 수가 1회부터 50회까지인 경우의 결과만을 보여주고 있으나, 50회부터 1,843회까지의 결과도 구간에 따라 특별히 범위를 벗어나는 경우도 있으나 위의 구간 내에서 성과 값이 움직이고 있다.



<Figure 3> 상품의 피 선호도 표시 수와 추천 성과 간의 관계

4. 토의

사용자의 선호도 표시 수와 협동적 필터링 성과간의 관계에 관한 실험의 결과는 그룹1부터 그룹5까지 희소수준이 0.997에서 0.985까지 감소할수록 추천 성과가 좋아진다고 볼 수 있으나, 그룹6부터 그룹10까지는 희소수준이 0.98에서 0.90까지 감소하여도 추천 성과는 별로 나아지지 않았다. 이는 협동적 필터링에서 선호도 데이터의 희소수준이 일정수준보다 낮아져도 추천 성과의 개선을 가져오지 못한다는 것으로 볼 수 있다. 따라서 사용자로부터 많은 수의 선호도 표시 데이터를 확보하기 위해 많은 비용과 자원을 투입하는 것 보다는 일정 수준의 선호도 표시 데이터만을 확보하여 사용하는 것이 상품추천에서 효율적일 수 있다. 반대로 협동적 필터링의 수행을 통해 개인화된 상품을 추천하기 위해서는 선호도 표시 데이터의 희소수준이 일정 수준이상 되어야 안정적인 추천 성과를 보장할 수 있다는 것을 뜻하기도 한다. 상품의 수와 사용자의 수가 증가할수록 상품추천에 적절한 희소수준을 유지하는 것이 어려우므로, 적절한 마케팅 활동과 비용의 투입을 통해 사용자의 선호도 표시 데이터를 적정수준으로 유지하는 것이 필요하다.

사용자의 선호도 표시 수에 따른 최적 초기 추천 시점에 관한 실험의 결과는 데이터의 특성에 따라 유동적이기는 하지만 사용자가 12개에서 19개 이상의 상품에 선호도를 표시하였을 때가 그 이하만큼의 상품에 선호도를 표시하였을 때보다 추천 성과가 안정적인 것으로 볼 수 있다. 따라서 협동적 필터링을 통해 상품을 개인화하여 추천하는 경우에는 사용자가 20개정도의 상품에 선호도를 표시한 이후에야 안정적인 추천 성과를 나타낼 수 있다. 신규 사용자가 들어온 경우에는 사용자의 선호도 데이터가 부족하여 협동적 필터링의 성과가 만족스럽지 못하게 되는 Cold start 문제를 해결하기 위해서, 초기 사용자에게는 인구통계학적인 정보를 활용하거나 다른 전략들을 통해 추천을 수행하고 있다 (Schein et al, 2002, Rashid et al., 2002). 위의 실험결과를 이런 초기 전략들에서 협동적 필터링의 적용으로 전환하는 적절한 시점을 선택하기 위한 기준으로 활용할 수 있다.

상품의 피 선호도 표시 수와 추천 성과 간의 관계에 대한 실험의 결과는 8에서 15회 이상 선호도가 표시된 상품을 추천하는 것이 추천 성과가 안정적인 것으로 볼 수 있다. 1회에서 8회 까지 선호도 표시회수를 증가시키면서 수행한 실험에서는 추천 성과가 지속적으로 개선되지만 그 이후에서는 어느 정도의 범위 안에서 추천 성과가 움직이는 것을 볼 수 있다. 전자상거래 업체의 입장에서는 신규상품의 경우에 15명 이상의 사용자가 평가한 상품을 협동적 필터링을 통해 추천하는 것이 적절하다고 볼 수 있다. 이를 위해 신규 상품에 대해 평가하여 주는 평가자 집단을 일정 규모 확보하여 운영하는 것이 신규상품 추천에 효과적일 수 있겠다.

5. 결론

본 연구에서는 EachMovie 데이터를 활용한 분석을 통하여 협동적 필터링에 활용되는 데이터 선정을 위한 계수와 추천 성과간의 관계에 대하여 연구를 수행하였다. 실험을 통하여 희소수준이 일정 수준이상으로 감소하거나, 사용자의 선호도 표시 수와 상품의 피 선호도 표시수가 일정수준 이상으로 증가하여도 추천 성과에는 많은 영향을 미치지 않는 것으로 파악되었다. 따라서 전자상거래 업체의 입장에서는 안정적인 추천 성과를 확보하기 위해서는 사용자 선호도 데이터의 적정한 희소수준을 유지하여야 한다. 또한 사용자의 선호도 표시 수와 상품의 피 선호도 표시 수가 추천 성과에 영향을 미치므로 이를 고려한 데이터 선정과 협동적 필터링의 운영이 필요하다.

References

김재경, 서지혜, 안도현, 조윤호 (2002), 협업 필터링 기법을 활용한 개인화된 상품 추천방법론 개발에 관한 연구, *한국지능정보시스템학회논문지* 8 (2), 139-157.

김종우, 이경미 (2000), 인터넷 상점에서 개인화 광고를 위한 장바구니 분석 기법의 활용, *경영과학* 17(3), 19-30.

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

Kim, J.K., Cho, Y.H., Kim, W.J., Kim, J.R., and Suh, J.H. (2002), A personalized recommendation procedure for Internet shopping support, *Electronic Commerce Research and Applications* 1, 301-313.

Mild, A., and Natter, M. (2002), Collaborative filtering or regression models for Internet recommendation systems?, *Journal of Targeting, Measurement and Analysis of Marketing* 10 (4), 304-313.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2000), Analysis of Recommendation Algorithms for E-Commerce, Proceedings of EC'00, Minneapolis.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001), Item-Based Collaborative Filtering Recommendation Algorithms, Proceedings of WWW10, Hong Kong.

Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. (2002), Methods and Metrics for Cold-Start Recommendations, *Proceedings of the ACM SIGIR'02*, Tampere, Finland, 253-260.

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

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J. (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.