

# 협업필터링 추천시스템에서 개인별 선호도의 표준화에 따른 예측성능의 영향

이희춘<sup>a</sup> · 김선옥<sup>b</sup> · 이석준<sup>c</sup>

<sup>a</sup>상지대학교 컴퓨터데이터정보학과  
220-702, 강원도 원주시 우산동 660번지  
Tel: +82-33-730-0406, Fax: +82-33-730-0405, E-mail: choolee@sangji.ac.kr

<sup>b</sup>한라대학교 정보통신공학부  
220-712, 강원도 원주시 한라대1길 32  
Tel: +82-33-760-1582, Fax: +82-33-760-1280, E-mail: sokim@halla.ac.kr

<sup>c</sup>상지대학교 경영학과  
220-702, 강원도 원주시 우산동 660번지  
Tel: +82-33-730-0406, Fax: +82-33-730-0405, E-mail: digitaldesign@sangji.ac.kr

## 요약

본 연구는 추천시스템에서 협업필터링 알고리즘을 이용하여 특정 상품에 대한 고객의 선호도를 예측함에 있어 고객이 상품에 대해 평가한 선호도 평가치를 고객별로 표준화시켜 예측하여 기존의 예측 정확도를 향상시키는 방법에 대하여 연구하였다. 일반적으로 상품에 대한 고객의 선호도를 평가하기 위하여 절대적 기준의 수치적 척도가 제공되지만 개인에 따라서는 상품에 대한 선호 정도가 절대적 척도에 다르게 반영되어 개인별 선호도에 차이가 발생할 수 있다. 이러한 개인적 특성이 동일한 척도의 평가치로 예측되면 예측 결과의 오차를 크게 할 가능성이 있다. 또한 개인이 평가한 선호도 평가치의 편차가 협업필터링 알고리즘을 통한 선호도 예측 정확도와 밀접한 관계를 가지고 있음을 알 수 있었으며 이러한 문제를 해결하기 위하여 개별 고객이 평가한 선호도 평가치를 표준화시켜 표준화된 선호도 평가치를 이용한 선호도 예측을 실시하였다. 분석결과 표준화된 선호도 평가치를 이용한 예측 결과가 비표준화 선호도 평가치를 이용한 예측 결과보다 예측력이 우수함을 알 수 있었으며 결과에 대한 통계적 분석을 통하여 표준화된 선호도 평가치를 이용한 선호도 예측 방법과 비표준화 선호도 평가치를 이용한 선호도 예측 방법을 혼합할 경우 선호도 예측 정확도를 더 향상시킬 수 있음을 알 수 있었다.

## Keywords:

예측 정확도; 표준화; 협업필터링

## 1 서론

인터넷의 급속한 확산과 정보기술의 발달로 인하여 손쉬운 정보의 생성과 전달이 가능하게 되었으며 이로 인한 생활방식이 변화하고 있다. 특히 상거래부문에서 전자상거래의 출현은 기존의 상거래 방식과 다른 새로운 전환점을 제시하고 있다. 전자상거래 방식은 기존의 오프라인 상거래 방식과 다른 마케팅 전략과 차별화 전략들이 필요하게 되었으며 특히 개인의 선호 취향을 고려할 수 있는 서비스의 필요성이 증대되고 있다. 추천시스템은 인터넷 기반의 전자상거래에서 개인의 선호 취향을 고려할 수 있는 개인화 서비스를 제공할 수 있는 대안으로 제시되고 있으며 실제 많은 전자상거래 사이트에 적용되어 성과를 거두고 있는 것으로 평가되고 있다. 추천시스템은 무수히 많은 정보 중 개인의 취향을 고려하여 개인에게 적합하거나 잠재적으로 요구되는 정보만을 필터링하여 고객에게 제공할 수 있으며 이를 통하여 고객의 정보과잉 문제를 해결하여 전자상거래에 대한 고객의 만족도를 증대시킬 수 있기 때문에 많은 상품들이 존재하는 전자상거래에서 특히 필요한 시스템이다. 추천시스템은 다양한 접근법이 적용되지만 협업필터링 접근법이 상업적으로 성공적인 성과를 거두고 있는 것으로 평가되고 있다. 협업필터링 기법은 특정 상품에 대한 고객의 관심 정도를 시스템의 데이터베이스에서 유사한 성향을 가진 다른 고객들의 선호 정보를 이용하여 예측한다. 특히 협업필터링의 경우 상품의 다양한 속성과 고객의 특성을 의도적으로 무시하고 상품에 대한 고객의 선호 정보만을 이용한다. 본 연구에서는 상품에 대한 고객의 선호도 평가치를 표준화시켜 표준화된 선호도 평가치를 이용한 선호도 예측을 실시하여 기존의 접근법과 선호도 예측 정확도의

측면에서 비교하였다.

## 2 연구목적

협업필터링은 상품에 대한 고객들의 선호 정보를 이용하여 특정 상품에 대한 추천 대상 고객의 선호도를 예측한다. 특히 협업필터링 알고리즘 중 메모리기반(memory-based)의 알고리즘은 고객의 선호도 평가치를 이용하여 고객의 선호도를 예측하기 때문에 상품에 대한 고객들의 선호도 평가치가 필요하다. 이때 상품에 대한 고객의 선호도 평가치는 일반적으로 명시적 자료인 수치척도가 이용되며 5점, 7점, 10점 척도로 고객들이 평가하도록 되어 있다. 고객에게 제시된 평가척도가 절대적 기준의 수치 척도로 주어지 있지만 고객에 따라 서로 다른 기준이 적용되어 상품에 대해 개인의 기준이 적용될 수 있으며 또한 고객에 따라서는 상품에 대한 선호 정도의 차이가 큰 고객과 차이가 크지 않은 고객이 있을 수 있다. 본 연구는 이러한 동일한 기준이 주어지더라도 고객에 따라 서로 다른 평가의 영향을 줄이기 위해 고객에 따라 선호도 평가치를 표준화시켜 표준화된 선호도 평가치를 이용한 선호도 예측의 정확도를 비표준화 선호도 평가치를 이용한 예측 결과와 비교하였다. 또한 선호도 예측 결과를 분석하여 두 분석 방법의 장점을 결합한 선호도 예측 방법을 제안하여 선호도 예측 정확도를 향상시켰다.

## 3 관련연구

### 3.1 협업필터링

협업필터링 접근법은 고객과 상품간의 관계 데이터인 선호도 평가치를 이용하는 접근법이다[10]. 협업필터링 기법은 선호도 예측 알고리즘의 유형에 따라 메모리 기반(memory-based)의 선호도 예측 알고리즘과 모형기반(model-based)의 선호도 예측 알고리즘으로 구분한다[4]. 메모리기반의 알고리즘은 상품에 대한 고객들이 평가한 전체 선호도 평가치를 기반으로 특정 상품에 대한 추천 대상 고객의 선호도 평가치에 대한 예측이 이루어지며 선호도를 예측하기 위하여 고객 데이터 베이스의 모든 평가치를 이용하게 된다. 그렇기 때문에 선호도 예측을 위해 예측 대상 상품에 대해 선호도를 평가한 모든 고객들의 정보를 이용하기 보다 선호도 평가치의 패턴에 따라 가장 유사한 일정수의 고객을 선정하여 선호도를 예측하는 k-nearest neighbor 방법을 이용하기도 한다[13]. 이웃 고객간의 유사 정도를 나타내기 위한 유사도 가중치는 일반적으로 피어슨 상관계수[10]와 벡터 유사도 가중치[4]가 이용되거나 피어슨 상관계수를 이용한 예측 정확도가 더 높은 것으로 평가된다[1][3]. 모형기반의

알고리즘은 선호도 예측을 위하여 고객들의 선호도 평가치의 패턴을 기반으로 소규모 계층으로 고객들을 구분한다. 추천 대상 고객의 선호도 평가치를 예측하기 위하여 추천 대상 고객을 하나 혹은 다수의 소규모 계층에 분류하여 추천 상품에 대해 각 계층에 의해 예측된 평가치를 추천 대상 고객의 선호도 예측치로 사용한다. 모형기반 알고리즘의 유형에는 베이지안 네트워크, 군집화 등의 방법이 있다. 메모리 기반의 알고리즘은 데이터의 희소성에 따라 예측 오차가 커지며 고객과 상품의 수가 증가하면 이웃과의 관계를 계산하는 계산량이 증가하는 단점이 있으며 모형기반의 알고리즘은 계층의 모형 설정에 소요되는 시간이 매우 길고 추가 데이터에 따라 모형을 재 설정하여야 하는 단점을 가지고 있으며 메모리 기반의 예측 결과와 비교하여 다양한 예측 결과를 얻을 수 없다는 단점을 가지고 있다. 본 연구에서는 메모리 기반의 알고리즘을 중심으로 연구하였다. 메모리기반의 알고리즘을 이용한 선호도 예측 정확도 개선에 관한 많은 연구가 이루어지고 있다. 고객간의 관계를 나타내는 유사도 가중치에 관한 연구[12]와 이웃 고객의 선정에 관한 연구를 통하여 선호도 예측 정확도의 향상에 대한 연구가 진행되었으며 선호도 예측 알고리즘의 개선에 관한 연구도 진행되었다. 최근에는 추천시스템의 보안에 대한 연구도 진행되고 있으며 또한 시스템에 악영향을 줄 수 있는 노이즈의 선별 방법에 대한 연구도 진행되고 있다. 이석준 등[2][6][7]의 연구에서는 고객이 평가한 선호도 평가치의 패턴을 통하여 선호도 예측 이전에 고객들의 예측 오차를 평가할 수 있는 가능성이 제시되었으며 연구 결과에서 고객이 평가한 선호도 평가치의 특정 빈도에 따라 예측 오차가 큰 고객들을 예측 이전에 선별할 수 있는 방법이 제안되었다. 또한 고객들의 평가패턴 중 시간에 따른 평가치의 Run과 예측 오차와의 관계에 대한 연구도 진행되었다. 그리고 고객이 평가한 선호도 평가치의 통계적 특성이 선호도 예측 결과와 밀접한 관련이 있음을 실험을 통하여 제시하였다[2][6][7].

### 3.2 메모리기반 선호도 예측 알고리즘

최초의 자동화된 협력적 필터링 알고리즘은 GroupLens에서 제시되었으며 추천 대상 고객과 이웃 고객들의 선호도 평가치의 평균과 선호 경향의 유사 정도를 나타내는 유사도 가중치를 이용하여 선호도를 예측하게 된다[4][10]. 다음은 식(1)은 GroupLens에서 제안한 이웃 기반의 협력적 필터링 알고리즘(Neighborhood Based Collaborative Filtering Algorithm)에 대한 정의이다.

$$\hat{U}_x = \bar{U} + \frac{\sum_{J \in \text{Raters}} (J_x - \bar{J}) r_{uj}}{\sum_{J \in \text{Raters}} |r_{uj}|}, \text{ where } \bar{J} = \frac{\sum_i J_i}{n}, i \neq x \quad (1)$$

- $\hat{U}_x$ : 상품  $x$ 에 대한 추천 대상 고객  $U$ 의 선호도 예측치
- $\bar{U}$ : 추천 대상 고객  $U$ 가 평가한 모든 상품에 대한 선호도 평가치 평균
- $J_x$ : 상품  $x$ 에 대한 이웃 고객  $J$ 의 선호도 평가치
- $\bar{J}$ : 상품  $x$ 에 대한 이웃 고객  $J$ 의 선호도 평가치를 제외한 모든 상품에 대한 선호도 평가치 평균
- $r_{uj}$ : 추천 대상 고객  $U$ 와 이웃 고객  $J$ 의 선호도 유사도 가중치

$r_{uj}$ 는 추천 대상 고객  $U$ 와 이웃 고객  $J$ 의 선호도의 유사 정도를 나타내는 유사도 가중치로 본 연구에서는 피어슨 상관계수를 이용한다. Lee(2006)의 연구에서는 GroupLens에서 제안한 NBCFA 알고리즘을 개선한 대응평균 알고리즘 (Correspondence Mean Algorithm)을 제안하여 선호도 예측 대상 고객의 선호도 예측 정확도를 향상하였다[8][9]. CMA는 다음 식(2)와 같이 정의한다.

$$U_x = \bar{U}_{match} + \frac{\sum_{J \in \text{Raters}} (J_x - \bar{J}_{match}) r_{uj}}{\sum_{J \in \text{Raters}} |r_{uj}|} \quad (2)$$

여기서,

- $\bar{U}_{match}$ : 선호도 예측 대상 고객  $u$ 와 각 이웃 고객  $j$ 가 공통으로 평가한 상품들의 평가치의 평균들의 평균
- $\bar{J}_{match}$ : 선호도 예측 대상 고객  $u$ 와 이웃 고객  $j$ 가 공통으로 평가한 상품들에 대한 선호도 평가치의 평균

$\bar{U}_{match}$ 는 선호도 예측 대상 고객  $u$ 의 선호도를 이웃 고객들과 공통으로 평가한 평가치의 평균을 다시 평균하여 사용하기 때문에 NBCFA에서 과도하게 반영된 고객  $u$ 의 선호도를 조정하는 역할을 한다. NBCFA에서 사용되는  $\bar{J}$ 는 이웃 고객  $j$ 의 전체 선호도 평가치의 평균이 되는 반면 CMA의  $\bar{J}_{match}$ 는 선호도 예측 대상 고객  $u$ 와 이웃 고객  $j$ 가 공통으로 평가한 상품들에 대한 선호도 평가치만을 이용한 평균으로 NBCFA에서 이웃 고객  $j$ 의 선호도가 과도하게 적용되는 것을

조정하는 역할을 한다. 본 연구에서 사용된 유사도 가중치는 벡터 유사도 가중치보다 성능이 우수한 피어슨 상관계수를 이용하였다.

### 3.3 평가척도

협력적 필터링 알고리즘에 의한 선호도 예측 정확도 성능은 실제 선호도 평가치와 예측 선호도의 절대평균오차(Mean Absolute Error)를 이용하며 MAE가 크면 전체 시스템의 예측 정확도가 낮아지는 것이고 MAE가 작으면 예측 정확도가 높아진다[4][5][11]. MAE는 다음 식(3)과 같이 정의한다.

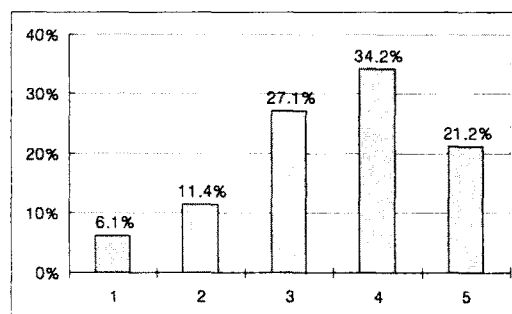
$$MAE = \frac{1}{N} \sum_j |R_{uj} - \hat{R}_{uj}| \quad (3)$$

- $R_{uj}$ : 상품  $j$ 에 대한 고객  $u$ 의 실제 선호도 평가치
- $\hat{R}_{uj}$ : 상품  $j$ 에 대한 고객  $u$ 의 선호도 평가치의 예측치

## 4 실험 방법

### 4.1 실험 dataset

본 연구에서는 MovieLens 100K dataset을 이용하여 고객의 선호도 평가치를 표준화시켜 선호도 예측을 실시하여 비 표준화 선호도 평가치를 이용한 선호도 예측 결과와 비교하였다. MovieLens 100K dataset은 943명의 고객들이 1682편의 영화에 대하여 1에서 5점의 평가치로 영화에 대한 선호도를 평가한 자료이다. 각 고객들은 최소 20편의 영화에 대하여 평가하였고 최대 737편의 영화에 대하여 평가한 고객도 있으며 평균적으로 106편의 영화에 대하여 평가하였다. MovieLens dataset은 영화에 대한 선호도 평가자료 외에도 평가자의 간략한 인구통계 정보와 영화에 대한 정보로 구성되어 있으며 평가시간에 대한 정보도 포함하고 있다. 다음 [그림1]은 MovieLens 100K dataset의 선호도 평가치 분포도이다.

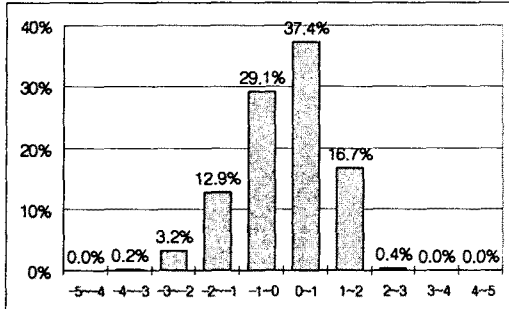


[그림1] MovieLens 100K dataset 평가치의 분포도

[그림1]에서 전체 선호도 평가치의 평균은

3.529이며 오른쪽으로 치우친 형태의 분포를 보이고 있다.

다음 [그림2]는 MovieLens 100K dataset의 선호도 평가치를 고객별로 표준화시킨 선호도 평가치의 분포도이다.

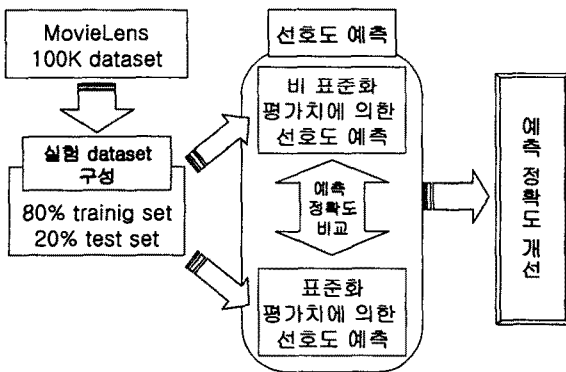


[그림2] 표준화시킨 선호도 평가치의 분포도

[그림1]과 [그림2]에서 비 표준화 선호도 평가치의 분포와 표준화시킨 선호도 평가치의 분포는 모두 우측으로 치우친 형태를 보이고 있으나 표준화시킨 선호도 평가치의 분포가 비 표준화 선호도 평가치의 분포보다 극단치를 제외한 평가치들이 가운데로 모인 것을 알 수 있다.

## 4.2 실험 구성

비 표준화 선호도 평가치를 이용한 선호도 예측 정확도와 표준화시킨 선호도 평가치를 이용한 선호도 예측 정확도를 비교하기 위하여 본 연구에서는 다음 [그림3]과 같이 실험을 구성하였다.



[그림3] 실험 구성도

본 연구는 [그림3]과 같이 실험을 구성하였으며 선호도 예측 정확도를 비교하기 위하여 MovieLens 100K dataset을 80%의 훈련집합과 20%의 실험집합으로 분할한 실험 dataset을 구성하였다. 선호도 평가치의 비 표준화와 표준화에 따른 예측 정확도를 분석하기 위하여 비 표준화 선호도 평가치를 이용한 선호도 예측을 하였다. 선호도 평가치를 표준화하기 위하여 80%의 훈련집합에서

각 고객별로 선호도 평가치의 평균과 표준편차를 이용하여 표준화 선호도 평가치를 생성하여 선호도 예측을 하였다. 최종적으로 비 표준화와 표준화 선호도 평가치를 이용한 선호도 예측 결과를 비교 분석하여 선호도 예측 정확도의 개선 방안을 제시한다.

## 5 실험 및 결과

다음 [표1]은 비 표준화 선호도 평가치와 표준화 선호도 평가치를 이용한 각 알고리즘의 예측 결과를 개별 고객별로 나누어 MAE를 구하여 대응표본 t검정을 하였다.

[표1] 비 표준화와 표준화 결과의 대응표본 t검정 결과

알고리즘	구분	평균	평균차	t값	유의확률
NBCFA	비 표준화	0.7879	0.0043	4.151	0.0000**
	표준화	0.7836			
CMA	비 표준화	0.7716	0.0039	3.863	0.0001**
	표준화	0.7677			

\*:  $p < 0.05$ , \*\*:  $p < 0.01$

분석결과에서 NBCFA와 CMA 모두에서 개인별 선호도 평가치의 표준화를 통한 선호도 예측 결과가 우수한 성능을 보이고 있음을 알 수 있다.

[표2] 집단구분을 통한 선호도 예측 성능 비교

알고리즘	구분	집단 구분	빈도	MAE 평균	F값	유의 확률	사후 검정	
NBCFA	비 표준화	1	234	0.6492	105.59	0.000**	{1}{2} {3}{4}	
		2	238	0.7075				
		3	236	0.8165				
		4	235	0.9787				
			합계	943	0.7879			
	표준화	1	234	0.6460	96.61	0.000**	{1}{2} {3}{4}	
		2	238	0.7092				
		3	236	0.8153				
4		235	0.9640					
		합계	943	0.7836				
CMA	비 표준화	1	234	0.6434	90.35	0.000*	{1}{2} {3}{4}	
		2	238	0.6964				
		3	236	0.7942				
		4	235	0.9528				
			합계	943	0.7716			
	표준화	1	234	0.6403	84.40	0.000**	{1}{2} {3}{4}	
		2	238	0.6968				
		3	236	0.7929				
4		235	0.9412					
		합계	943	0.7677				

\*:  $p < 0.05$ , \*\*:  $p < 0.01$

[표1]의 결과에 대하여 추가 분석을 위해 고객이 평가한 선호도 평가치의 표준편차의 크기에 따라 4분위수를 기준으로 4집단으로 구분하였다. [표2]는 구분된 집단간 평균에 대한 검정 결과이다. 결과에서 비 표준화 및 표준화 결과에서 모두 표준편차의 크기에 따라 구분된 집단간에 통계적으로 유의한 평균의 차이가 있음을 알 수 있으며 사후분석 결과에서도 각 집단이 잘 구분됨을 알 수 있다. 또한 표준편차가 큰 집단일수록 개별 고객들의 MAE 평균이 증가함을 알 수 있으며 이는 고객이 평가한 선호도 평가치의 표준편차와 선호도 예측 정확도가 관계가 깊음을 알 수 있다. [표2]의 결과에 대하여 비 표준화와 표준화에 의한 결과를 각 집단별로 다시 분석하였다. 다음 표[3]은 구분 집단에서 표준화와 비 표준화의 예측 정확도에 대한 분석결과이다.

[표3] 구분 집단에서 비 표준화와 표준화에 따른 선호도 예측 정확도의 비교

구분	알고리즘	구분	MAE평균	평균차	t값	유의확률
1집단 (234)	NBCFA	비 표준화	0.6492	0.0032	1.030	0.3041
		표준화	0.6460			
	CMA	비 표준화	0.6434	0.0031	1.121	0.2635
		표준화	0.6403			
2집단 (238)	NBCFA	비 표준화	0.7075	-0.0017	-1.914	0.0568
		표준화	0.7092			
	CMA	비 표준화	0.6964	-0.0004	-0.348	0.7280
		표준화	0.6968			
3집단 (236)	NBCFA	비 표준화	0.8165	0.0012	1.995	0.0472*
		표준화	0.8153			
	CMA	비 표준화	0.7942	0.0013	2.108	0.0361*
		표준화	0.7929			
4집단 (235)	NBCFA	비 표준화	0.9787	0.0148	5.784	0.0000**
		표준화	0.9640			
	CMA	비 표준화	0.9528	0.0117	4.419	0.0000**
		표준화	0.9412			

\*: p<0.05, \*\*: p<0.01

[표3]에서 표준편차의 크기에 따라 구분된 각 집단에서 각 알고리즘에 따른 표준화 평가치의 예측 결과와 비 표준화 평가치의 예측 결과에 대한 대응표본 t검정 결과이다. 결과에서 표준편차의 크기가 작은 1집단과 2집단에서는 비 표준화 평가치를 이용한 선호도 예측 정확도와 표준화 평가치를 이용한 예측 정확도에는 통계적으로 유의한 평균의 차이가 나타나지 않음을 알 수 있다. 그러나 3집단과 4집단에서는 비 표준화 평가치를 이용한 예측과 표준화 평가치를 이용한 선호도 예측 정확도에서 통계적으로 유의한 차이가 나타남을 알 수 있다. 또한 표준편차가 큰 4집단의 경우 통계적 유의성이 타 집단과 비교하여 크다는 것을 알 수 있다. 분석결과를 이용하여 비 표준화 평가치를

이용한 선호도 예측 방법과 표준화 평가치를 이용한 선호도 예측 방법을 통합하여 선호도 예측을 실시하였다. [표3]의 결과를 이용하여 표준편차의 크기에 의해 분할된 집단에서 1집단과 2집단의 경우 비 표준화 선호도 평가치를 이용하여 선호도 예측을 실시하고 3집단과 4집단은 표준화 선호도 평가치를 이용하여 선호도 예측을 실시하였다. 다음 [표4]는 두 예측 방법을 통합하여 분석한 결과이다.

[표4] 통합 예측방법에 의한 선호도 예측 정확도의 비교

알고리즘	구분	평균	평균차	t값	유의확률
NBCFA	비 표준화	0.7879	0.0043	4.151	0.0000**
	표준화	0.7836			
	비 표준화	0.7879	0.0040	5.825	0.0000**
	통합	0.7839			
CMA	비 표준화	0.7716	0.0039	3.863	0.0001**
	표준화	0.7677			
	비 표준화	0.7716	0.0032	4.660	0.0000**
	통합	0.7684			

\*: p<0.05, \*\*: p<0.01

[표4]의 결과에서 비 표준화 선호도 평가치를 이용한 선호도 예측 결과와 표준화 선호도 평가치를 이용한 예측 결과와 두 방법을 통합한 방법의 결과를 이용하여 평균에 대한 분석을 하였다. NBCFA와 CMA에 의한 선호도 예측 정확도가 모두 비 표준화 선호도 평가치를 이용한 예측 정확도보다 표준화 선호도 평가치를 이용한 예측 정확도가 향상되었음을 알 수 있으며 표준화 선호도 평가치를 이용한 결과와 통합 방법을 이용한 결과는 평균에서 개선이 나타나지 않지만 통계적 유의성에서 비 표준화와 비교하여 통합 방법을 적용한 선호도 예측 결과가 개선되었음을 알 수 있다.

## 6 결론

본 연구는 고객이 상품에 대하여 평가한 선호도 평가치가 절대적 기준의 수치 척도로 제공되지만 고객에 따라서는 선호 정도의 표현이 달라질 수 있을 것으로 가정하고 각 개별 고객이 평가한 선호도 평가치를 표준화시켜 선호도 예측의 정확도를 향상시키기 위한 방법을 실험을 통하여 분석하였다. 분석결과 선호도 평가치의 표준화를 통한 선호도 예측 정확도는 비 표준화 선호도 평가치를 이용한 예측 정확도에 비하여 개선되었음을 알 수 있으며 특히 고객이 평가한 선호도 평가치의 표준편차가 큰 고객에게서 예측 정확도의 향상이 크게 나타남을 알 수 있었다. 그러나 선호도 평가치의 표준편차가 작은 고객에서는 예측 정확도의 향상이 나타나지 않았으며 분석 결과에서는 오히려 반대의 경향이

나타날 수도 있음을 알 수 있다. 본 연구에서는 실험결과를 바탕으로 표준편차가 작은 고객은 비표준화 선호도 평가치를 이용한 선호도 예측을 실시하고 표준편차가 큰 고객은 표준화 선호도 평가치를 이용한 선호도 예측을 실시하는 통합적인 방법을 적용하여 통계적 유의성을 향상시켰다. 본 연구에서는 선호도 평가치의 표준화를 통한 방법이 협업필터링을 이용한 고객의 선호도 예측 정확도를 향상시킬 수 있음을 제안하였으며 차기 연구로는 선호도 예측 알고리즘에 대한 개선을 통하여 분석결과를 추가할 수 있는 방법과 개별 고객의 선호도 예측치와 실제 선호도 평가치의 순위적합에 대한 분석과 연구가 필요하다고 생각한다.

## 참고문헌

- [1] 이석준, 이희춘 (2007). “협업 필터링 추천에서 대응평균 알고리즘의 예측 성능에 관한 연구”, *Information Systems Review*, 제9권, 제1호, pp. 85-103.
- [2] 이석준, 김선옥, 이희춘 (2007). “협력적 필터링에서 평가치의 Run 특이자와 예측 정확도의 관계에 관한 연구”, *Journal of the Korean Data Analysis Society*, Vol.9, No.4, pp. 2043-2054.
- [3] 이희춘, 이석준 (2006). “사용자 기반 추천시스템에서 근접이웃 알고리즘과 수정알고리즘의 예측 정확도에 관한 연구”, *Journal of the Korean Data Analysis Society*, Vol.8, No.5, pp. 1893-1904.
- [4] Breese, J. S., Heckerman, D. and Kadie, C. (1998). “Empirical Analysis of Predictive Algorithms for Collaborative Filtering”, *In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp.43-52.
- [5] Herlocker, J.L., Konstan, J., Borchers, A., and Riedl, J. (1999). “An algorithmic framework for performing collaborative filtering”, *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp.230-237.
- [6] Lee, S. J., Kim, S. O., and Lee, H. C. (2007). “A Study on the Interrelationship between the Prediction Error and the Rating's Pattern in Collaborative Filtering”, *Journal of the Korean Data & Information Science Society*, Vol.18, No.3, pp.659-668.
- [7] Lee, S. J., Kim, S. O., and Lee, H. C. (2007). “Pre-Evaluation for Detecting Abnormal Users in Recommender System”, *Journal of the Korean Data & Information Science Society*, Vol.18, No.3, pp.619-628.
- [8] Lee, H. C. (2006). “An Exploratory Study for Decreasing Error of Prediction Value of Recommended System on User Based”, *Journal of the Korean Data & Information Science Society*, Vol.17, No.1, pp.77-86.
- [9] Lee, H. C., Lee, S. J., and Chung, Y. J. (2006). “The Effect of Co-rating on the Recommender System of User Base”, *Journal of the Korean Data & Information Science Society*, Vol.17, No.3, pp.775-784.
- [10] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). “GroupLens: an open architecture for collaborative filtering of netnews”, *In Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pp.175-186.
- [11] Shardanand, U. and Maes, P. (1995). “Social information filtering: algorithms for automating ‘word of mouth’”, *In Proceedings of the SIGCHI conference on Human factors in computing systems*, pp.210-217.
- [12] Symeonidis, P., Nanopoulos, A., Papadopoulos, A., and Manolopoulos, Y. (2006). “Collaborative filtering process in a whole new light”, *In Proceedings of the 10th International Database Engineering and Applications Symposium*, pp. 29-36.
- [13] Xue, G. R., Lin, C., Yang, Q., Xi, W., Zeng, H. J., Yu, Y., and Chen, Z. (2005). “Scalable collaborative filtering using cluster-based smoothing”, *In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp.114-121.