Journal of Korean
Data & Information Science Society
2008, Vol. 19, No. 3, pp. 819 ~ 830

Improving the MAE by Removing Lower Rated Items in Recommender System

Sun Ok Kim[1)] · Seok Jun Lee[2)] · Young Seo Park[3)]

Abstract

Web recommender system was suggested in order to solve the problem which is cause by overflow of information. Collaborative filtering is the technique which predicts and recommends the suitable goods to the user with collection of preference information based on the history which user was interested in. However, there is a difficulty of recommendation by lack of information of goods which have less popularity. In this paper, it has been researched the way to select the sparsity of goods and the preference in order to solve the problem of recommender system's sparsity which is occurred by lack of information, as well as it has been described the solution which develops the quality of recommender system by selection of customers who were interested in.

Keywords: 예측정확도; 추천시스템; 협력적 여과기법; 희소상품.

1. 서론

인터넷을 이용한 전자상거래의 활성화로 고객이 접하는 정보의 양이 늘어나고 있으며, 이에 따라 정보의 홍수 속에서 고객들은 원하는 서비스와 상품을 선택하기 위해 많은 시간과 노력이 필요하게 되었다. 추천시스템은 고객의 특성에 따라 적절한 상품을 추천함으로써 많은 정보로 인해 결정이 어려운 문제를 해결하기 위한 방안으로 제시되었다(이희춘과 이석준, 2006; Kim and Lee, 2007).

추천시스템은 분류 알고리즘, 군집분석 그리고 비정칙치 분해 기법 등 여러 가지 방법으로 구현될 수 있는데, 웹 추천시스템에 사용되는 중요한 기법 중의 하나는 다

1)Professor, Department of Communication Engineering, Halla University, Wonju, Gangwon 220-712, Korea. E-mail : sokim@halla.ac.kr

2)Professor, Department of Business Administration, Sangji University, Wonju, Gangwon 220-702, Korea. E-mail : crco909@yahoo.co.kr

3)(Corresponding Author) Professor, Dept. of Mathematics, Sunmoon University, Asan, Chungnam 336-708, Korea. E-mail : yseo@sunmoon.ac.kr

양한 정보 속에서 고객에게 적절한 정보를 찾아내는 정보 여과기법이다(강현철 등, 2004; 이희춘 등, 2006). 정보 여과기법은 크게 내용기반 여과기법과 협력적 여과기법으로 나눌 수 있다. 내용기반 여과기법은 추천대상 고객이 선호했던 과거의 정보를 바탕으로 관심을 갖게 될 상품을 추천한다. 협력적 여과기법은 내용기반 여과기법과 달리 추천대상 고객의 기본 정보와 유사성을 지닌 이웃 고객들을 선정하며 이들의 정보를 이용하여 추천대상 고객에게 적합한 상품을 추천한다. 본 논문은 협력적 여과기법을 이용한 추천시스템의 예측 선호도 개선을 위한 방법에 대한 연구이다.

2. 관련연구

협력적 여과기법을 이용한 추천시스템은 가장 성공적인 기법으로 현재 상용시스템에 많이 이용되고 있다. 이 시스템은 추천대상 고객과 유사성을 지닌 이웃 고객들의 선호 정보를 다른 고객의 선호도에 대한 정보와 함께 사용하여 고객이 좋아할만한 상품을 추천함으로 상품에 대한 적절한 선호 정보가 있어야 한다. 이러한 선호 정보는 일반적으로 상품에 대한 선호도 평가치의 형태로 수치 척도로 얻어진다. 적은 양의 선호도 평가치를 사용한 협력적 여과기법으로 이루어진 추천은 추천시스템의 추천 정확도와 신뢰성에 심각한 문제를 일으킬 수 있다. 이것을 추천시스템의 희소성 문제라 하며 이를 해결하기 위한 연구가 계속 진행되고 있다.

Kim et. al. (2008)은 상품에 대한 희소성의 수를 정의하고 이에 따라 집단을 분리하여 희소성이 선호도 예측 정확도에 미치는 변화를 분석하였고, 분류된 집단에 따라 선호도 예측 정확도에 유의적인 차이가 있음을 밝혔다. Pazzani (1999)은 희소성에 따라 데이터를 우선 선별하고 선별된 데이터를 속성별로 추출하여 추천시스템의 선호도 예측 정확도를 향상시키는 연구를 하였다. Soboroff and Nocholas (1999)은 희소성 문제를 행렬의 비정칙치 분해(SVD; Singular Value Decomposition)를 이용하여 추천시스템 성능 향상을 연구하였지만 선호도 예측의 정확도는 기존의 결과와 크게 차이가 나지 않았다. Kim and Kim (2005)은 데이터 변형기법을 사용하여 희소성이 높은 데이터의 희소성 감소를 제안하였다. 그리고 상품의 추가 속성 정보에 대한 확률분포를 이용하여 희소성의 데이터를 변경하고, 변경된 선호도 데이터를 협력적 여과기법을 이용하여 추천 성능을 향상시키는 연구를 하였다. 여기서는 다양한 형태의 선호도 평가 값에 대한 데이터들의 특성을 무시하고 확률분포만을 사용하였으므로 각 데이터들에 대한 정보가 정확하게 반영되지 않았다. Melville et. al. (2002)는 희소성이 있는 사용자의 평가치를 행렬을 이용한 내용기반 여과기법을 통해 사용자 평가 행렬을 생성하고, 이를 기반으로 협력적 여과기법을 이용하여 추천에 사용하였다. 이 연구에서는 희소성의 문제는 조금 완화되었지만 추천의 정확도는 크게 향상되지 못하였다.

본 논문에서는 상품에 대한 희소성을 조사하고, 희소성이 있는 상품을 선별하는 방법으로 고객들이 상품에 대해 평가한 선호도 응답수를 조사하였다. 그리고 선호도 응답수가 적은 상품을 선별하여 희소상품이라 정하고 이들에 대해 선호도를 평가한 고객을 조사하여 추천시스템의 예측 성능을 개선하기 위한 방법을 제시하였다.

3. 협력적 여과 기법

추천시스템에서 협력적 여과기법이 널리 사용되고 있으며, 협력적 여과기법은 Usenet 뉴스 기사의 선정을 도와주는 미네소타대학의 GroupLens에서 이용되었으며, 영화 추천을 위한 도구로 MovieLens, 음악 추천을 위해서 Ringo에서 사용되는 등 여러 영역에 적용되었다.

협력적 여과기법에서 가장 일반화된 알고리즘은 이웃기반 협력적 여과기법(Neighborhood Based Collaborative Filtering)이다. 이웃기반 협력적 여과기법은 추천대상고객의 선호도와 이웃 고객의 선호도를 함께 사용하여 상품에 대한 선호도를 예측하는 알고리즘이다. 추천을 위한 선호도 예측 값을 생성하기 위하여 우선 먼저 추천 대상 고객의 이웃을 선정한다. 많은 고객 중에서 추천대상 고객의 이웃을 선정하기 위한 이웃 선정방법에 대한 다양한 연구가 진행되고 있으며, 본 논문에서 사용한 이웃선정 방법은 추천 대상 고객의 선호도를 예측하고자 하는 상품에 선호도를 평가한 고객만을 이웃으로 선정한다. 다음으로, 선정된 이웃고객과 추천대상 고객과의 상품들에 대한 선호도 유사정도를 알아야 하는데, 선호도 유사정도는 추천대상 고객과 이웃 고객이 상품들에 평가한 선호도 평가치들의 상관관계를 이용한다. 다음 식은 두 고객의 유사정도를 나타내는 상관계수 중에서 추천시스템에서 가장 많이 사용되고 있는 피어슨 상관계수이다(Resnick et. al., 1994).

$$r_{uj}=\frac{\sum_{i=1}^{m}(R_{u,i}-\overline{R_u})(R_{j,i}-\overline{R_j})}{\sqrt{\sum_{i=1}^{m}(R_{u,i}-\overline{R_u})^2 \cdot \sum_{i=1}^{m}(R_{j,i}-\overline{R_j})^2}} \quad (1)$$

여기에서, r_{uj}는 추천대상 고객 u와 이웃고객 j와의 유사정도를 나타내는 가중치이며, $R_{u,i}$는 추천 대상 고객 u가 평가한 상품i 에 대한 선호도 평가 치이고, $R_{j,i}$는 이웃고객 j가 평가한 상품i 에 대한 선호도 평가이다. $\overline{R_u}$는 추천대상 고객 u가 평가한 모든 상품들에 대한 평균이고, $\overline{R_j}$는 추천대상 고객의 이웃인 j고객의 상품 선호도평가에 대한 상품들의 선호도 평가 치들에 대한 평균값이다. 유사도 가중치를 계산하기 위해 사용되는 평가치는 추천 대상 고객 u와 이웃고객 j가 공통으로 평가한 상품의 평가치만 사용한다.

추천 대상 고객에게 상품을 추천하기 위해서는 추천 상품에 대한 선호도 예측 값을 계산하여야 한다. 상품에 대한 선호도 예측은 추천대상 고객의 평균과 추천대상 고객의 이웃들이 평가한 평가 값 그리고 이들 이웃고객의 평균값을 사용하며, 식(1)에서 소개된 이웃과의 유사도 가중치를 알아야한다. 다음 식은 협력적 여과기법을 사용한 선호도 예측값을 계산하기 위한 알고리즘이다(Konstan et. al., 1997).

$$\widehat{U_x}=\overline{U}+\frac{\sum_{J\in Raters}(J_x-\overline{J})r_{uj}}{\sum_{J\in Raters}|r_{uj}|} \quad (2)$$

여기에서, $\widehat{U_x}$는 상품 x에 대한 추천 대상 고객 u의 선호도 예측 치이다. $\overline{U}$는 추천 대상 고객 u가 평가한 모든 상품에 대한 평균이다. J_x는 상품 x에 대한 이웃 고객 j의 선호도 평가 치이고, $\overline{J}$는 이웃 고객 j가 평가한 모든 상품에 대한 선호도의 평균이다. $\overline{J}$의 값은 평가치 중에서 상품 x에 대한 평가치는 제외한다. r_{uj}는 추천 대상 고객 u와 추천 대상 고객의 이웃고객인 j의 선호 유사 정도를 나타내는 유사도 가중치이며, 본 논문에서는 식 (1)의 피어슨 상관계수를 사용한다.

4. 예측 성능 개선 방안

4.1 선호도 예측의 정확도 평가척도

협력적 여과기법을 사용한 추천시스템에서 선호도 예측의 정확도를 평가하기 위해서는 추천대상 고객이 실제 평가한 평가 값과 협력적 여과기법을 이용하여 계산된 선호도 예측 값과의 절대평균오차(Mean Absolute Error)를 사용한다. 다음 식은 선호도 예측의 정확도를 판정하기 위한 계산식이다(Kim, Lee and Lee, 2008).

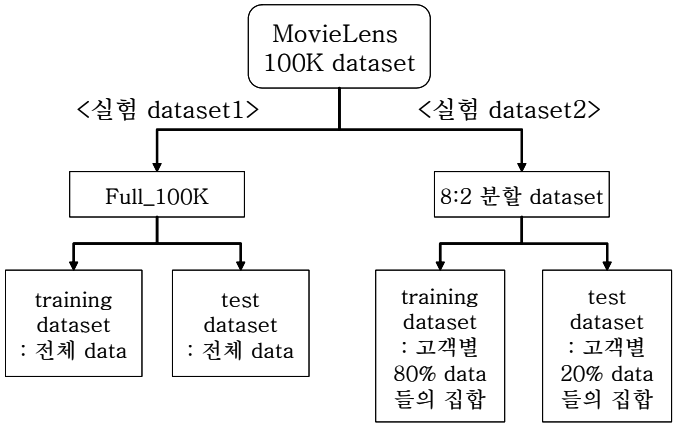
$$MAE = \frac{1}{N}\sum_{j=1}^{N}|R_{uj} - \widehat{R_{uj}}| \tag{3}$$

여기에서, R_{uj}는 상품 j에 대한 추천 대상 고객 u의 실제 선호도 평가 치이고, $\widehat{R_{uj}}$는 상품 j의 추천 대상 고객 u을 위한 선호도 예측 값이고, 예측 값은 협력적 여과기법의 알고리즘을 사용한다.

본 논문에서는 사용자 개인별로 개별 아이템에 대한 MAE를 계산하여 실험에 적용하였다.

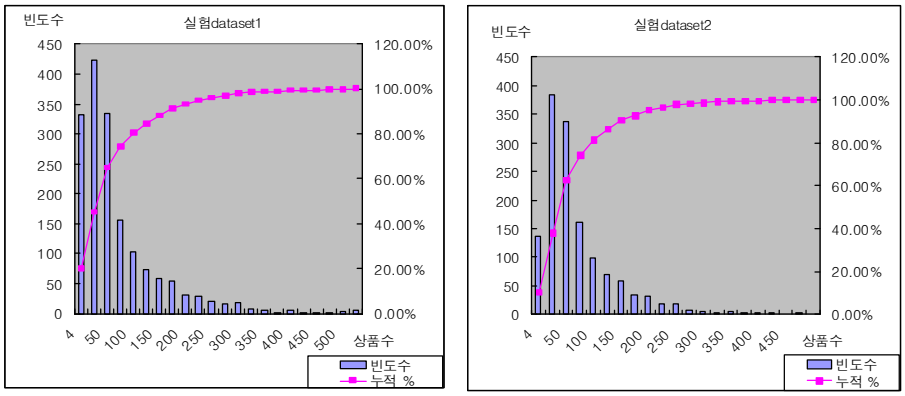
4.2 연구방법

본 논문에서 추천시스템의 성능 평가를 위해 사용된 실험 데이터는 GroupLens에서 제공된 MovieLens 100K 데이터 셋이며, 이 데이터는 영화에 관한 선호도를 나타낸 것이다. MovieLens 100K dataset은 943명의 고객이 전체 1682편의 영화 중 자신이 보았던 영화에 대한 선호도를 평가한 자료로 총 선호도 평가치의 개수는 100,000개로 구성되어 있다. 개별 고객은 20편 이상의 영화에 대해 선호도를 표시하도록 했으며, 관심의 정도에 따라 최소 1점에서 최대 5점까지 선호도를 평가할 수 있게 설계되어 있다. 본 논문에서는 실험을 위하여 GroupLens에서 제공된 MovieLens 100K dataset을 실험dataset1인 full_100K와 실험dataset2인 8:2 분할 dataset으로 나누어 연구에 사용하였다.



<그림 1> 실험 dataset의 구성

실험dataset1인 full_100K은 MovieLens 100K dataset 전체 data인 100,000개의 선호도 평가치에 대하여 선호도 예측을 실시하기 위하여 100,000개 전체 data를 이용하여 training dataset을 구성하였으며 검증용 dataset인 test dataset도 100,000개의 전체 data를 이용하여 구성하였다. 선호도 예측 알고리즘을 적용하여 test dataset의 선호도를 예측하기 위하여 알고리즘 적용 시 test dataset에서 예측을 하고자 하는 평가치를 training dataset에서 제외시키고 나머지 모든 data, 즉 99,999개를 training dataset으로 구성하였다. 실험dataset2인 8:2 분할 dataset은 알고리즘을 적용하기 위하여 MovieLens 100K dataset을 랜덤하게 80%와 20%로 나누었다. 이때, 개별 고객이 평가한 data들을 랜덤하게 80%와 20%로 나누고 개별 고객들의 80%에 해당하는 data들을 모아 training dataset을 구성하였으며 20%에 해당하는 data들을 모아 test dataset으로 구성하였다.



<그림 2> 실험 dataset들의 상품별 빈도수

실험dataset1과 실험dataset2의 상품에 대한 선호도 빈도는 <그림 2>와 같으며, 두 dataset 은 비슷한 분포를 이루고 있어 실험에 적합하다고 판단된다.

희소성을 가진 상품들은 추천시스템의 예측 성능에 영향을 미친다(Kim et. al., 2008). 따라서 본 논문에서는 빈도수가 적은 상품을 희소상품이라 정의하고 이들을 선호한 고객을 선별하여 이들이 추천시스템의 예측성능에 미치는 영향을 분석하여 예측성능을 높이고자 한다.

먼저 희소상품을 선별하기 위한 기준으로 상품에 대한 응답수를 정의하고, 고객의 응답에 따라 희소상품을 선별하기 위해 다음 식을 사용한다.

$$\sum_{j=1}^{n}\sum_{u=1}^{m}\chi_{R_{u,j}} \leqq s \tag{4}$$

여기에서, $R_{u,j}$는 추천 대상 고객 u가 평가한 상품j 에 대한 선호도 평가 값이고, j는 고객이 선호도를 표시한 상품이며 n은 전체 상품의 수이다. 그리고 m은 전체 고객의 수이며, s는 전체상품에 대한 응답수이다.

본 논문에서는 응답수가 4미만인 상품을 희소상품이라 정의하고 이들을 선호한 고객에 대한 데이터를 이용하여 추천시스템의 예측 성능을 높이기 위한 방법을 제시하고자 한다. 희소상품은 선호도 평가에 따라 식(4)에 의해 두 집단으로 나누어 사용한다.

$$C_j^-(s) = \{\, j \mid \sum_{j=1}^{k} m(j) \leqq s\},\qquad C_j^+(s) = \{\, j \mid \sum_{j=1}^{k} m(j) > s\} \tag{5}$$

여기에서, j는 고객이 선호도를 표시한 상품이고, k는 고객이 선호도를 표시할 총 상품의 개수이며 $m(j)$는 상품j에 대한 응답수이다.

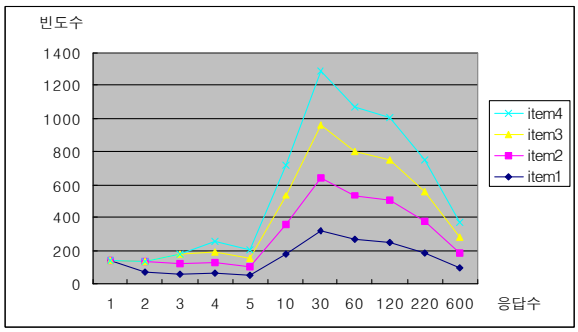
s값이 작을수록 $C_j^-(s)$은 선호도가 적은 상품들로 구성되며, s값이 증가함에 따라 선정된 상품들의 비율이 감소하는 $C_j^+(s)$에서는 선호도가 적은 상품들은 제외된다. 따라서 $C_j^+(s)$는 $C_j^-(s)$보다 희소상품을 덜 포함한다.

<표 1>은 실험dataset에서 희소상품을 응답수에 따라 분류한 집단 간 상품에 대해 선호도를 조사한 것이다. 이들에 대해 응답수에 따른 빈도를 조사하면 아래의 표와 같으며, 실험dataset1의 응답수가 1인 경우에는 $C_j^-(s)$의 선호도 상품이 하나도 없으며 $C_j^+(s)$에는 1682개의 상품수가 있음을 알 수 있다. 실험dataset2의 경우 응답수가 1인 것은 $C_j^-(s)$에서는 36개이며 $C_j^+(s)$는 1646개이다. 이것은 $C_j^-(s)$에서 36편의 영화에 1명만이 선호도를 표시했음을 의미한다. 실험dataset1의 응답수가 2인 경우에 141편의 영화가 있으며, 응답수가 3인 영화는 209편, 응답수가 4인 영화는 269편의 영화가 있음을 알 수 있다. 본 논문에 사용된 실험dataset1과 실험dataset2의 모든 경우에 응답수가 4에 가까울수록 $C_j^-(s)$의 빈도수는 커지며, 실험dataset2에서는 응답수가 4인 경우 323편의 영화에 4명의 고객이 선호도를 표시했음을 알 수 있다.

<표 1> 응답수에 따라 분류된 희소상품의 dataset들에 대한 집단 간 선호도 빈도수

응답수(s)	실험 dataset1		실험 dataset2	
	$C_j^-(s)$	$C_j^+(s)$	$C_j^-(s)$	$C_j^+(s)$
1	0	1682	36	1646
2	141	1541	174	1508
3	209	1473	245	1437
4	269	1413	323	1359

본 논문에서는 희소상품을 포함하는 $C_j^-(s)$을 제외시킨 $C_j^+(s)$를 응답수에 따라 4개의 dataset으로 구분하였다. 식(5)의 $C_j^+(1)$을 item1로, $C_j^+(2)$을 item2로, $C_j^+(3)$을 item3로 그리고 $C_j^+(4)$을 item4로 정하였다.



<그림 3> 응답수에 따라 분류된 실험dataset1의 누적빈도분포

분류된 실험dataset1에 대한 누적빈도분포를 응답수에 따라 살펴보면 위의 그림과 같으며, 응답수가 1 일 때, item1에는 141명이 상품에 대한 선호도를 평가하였고, 나머지 집단에서는 그러한 상품에 대한 선호도를 표시한 평가치가 없음이 조사되었다. item1은 선호도가 적은 상품을 가장 많이 가지고 있으며 희소성 데이터를 가장 많이 포함하고 있음을 알 수 있다.

먼저, $C_j^+(1)$의 상품에 선호도를 표시한 고객 즉, item1의 상품을 선호한 고객들의 집합을 step1이라 하고, $C_j^+(2)$, $C_j^+(3)$ 그리고 $C_j^+(4)$의 상품을 선호한 고객들의 집합을 step2, step3 그리고 step4 정하였다. 그리고 이들 dataset간에 예측 정확도를 알아보기 위하여 test dataset을 이용한 협력적 알고리즘을 통한 선호도 예측 값을 조사하였다. 계산된 선호도 예측의 평균은 <표 2>와 같으며, 응답수가 1이하인 희소상품을 제외시킨 상품을 선호한 고객의 정보를 user-based로 배열한 step1은 MAE 평균값이 실험dataset1은 0.6227이고 실험dataset2은 0.7522임을 알 수 있다. 희소성 데이터를 2개 제거한 고객들의 MAE 평균값은 0.6219와 0.7516으로 나타났으며, step1보다 MAE 평균값이 작아졌다. 실험 dataset1에서 step3과 step4의 MAE 평균값은 0.6216과 0.6211로 조사되었다. 두 실험dataset의 MAE는 모두 step1인 경우보다 step4인 경우 MAE 평균값이 작아졌으며, step4인 경우 가장 작은 값을 갖는다. 이것은 희소 상품을 적게 포함하는 실험dataset들의 MAE가 작은 값을 가지며 예측의 정확도가 개선됨

을 의미한다.

<표 2> 응답수에 따라 분류된 dataset들의 집단 간 MAE 평균

구분	MAE	
	실험 dataset1	실험 dataset2
step1	0.6227	0.7522
step2	0.6219	0.7516
step3	0.6216	0.7510
step4	0.6211	0.7500

따라서 희소성 데이터에 대한 관찰이 추천시스템의 성능을 향상시킬 수 있을 것이라 예상된다. 그러므로 선호도가 적은 데이터를 많이 포함하는 $C_j^-(s)$을 응답수에 따라 다음과 같이 나누었다.

응답수가 1 이하인 상품에 선호도를 평가한 고객들의 집단을 무조건1, 응답수가 2 이하인 상품에 선호도를 표시한 고객집단을 무조건2, 그리고 응답수가 3과 4 이하인 상품에 선호도를 표시한 고객 집단을 무조건3, 무조건4로 정하여 이들 상품에 선호도를 표시한 고객들을 분류하였다. 이렇게 분류된 고객들의 집단인 무조건1에서 응답수가 1인 희소상품을 제거한 집단을 user1이라 하고, 무조건2, 무조건3 그리고 무조건4에서 응답수가 2, 3 그리고 4 이하인 희소상품을 제거한 집단을 user2, user3, 그리고 user4로 정하였다. 그리고 이와 같이 분류된 고객집단에 대하여 희소상품을 포함했을 경우와 희소상품을 제거하였을 경우의 차이를 알아보기 위하여 test dataset을 이용한 집단 간 MAE를 대응평균 단측검정으로 조사하였다.

<표 3> 응답수에 따른 실험dataset1의 희소성제거에 의한 집단 간 MAE 대응평균검정

응답수	대응	빈도수	MAE	t값	유의확률
1	무조건1	58	0.6672	1.83	0.03*
	user1	58	0.6665		
2	무조건2	101	0.6614	3.07	0.00**
	user2	101	0.6580		
3	무조건3	159	0.6500	4.34	0.00**
	user3	159	0.6461		
4	무조건4	234	0.6479	4.82	0.00**
	user4	234	0.6437		

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

<표 3>의 결과에 따르면 응답수가 1인 희소상품을 선호한 집단인 무조건1인 경우 58개의 데이터가 있으며 이들의 MAE 평균값은 0.6672이고, 이 집단에서 희소상품을 제거한 user1도 마찬가지로 58개의 데이터를 가지고 있으며 MAE 평균값은 0.6665으로 무조건1인 경우보다 좋아졌다. user1은 무조건1인 집단보다 MAE가 개선되었으며,

user2, user3 그리고 user4도 무조건2, 무조건3 그리고 무조건4인 집단보다 MAE가 적은 값을 갖는다. 이는 희소성 데이터를 제거한 집단이 희소성 데이터를 제거하지 않은 집단보다 MAE가 향상되었음을 의미한다. user4의 MAE 평균값은 0.6437로 가장 작은 값을 가지며, user4가 user1 보다 희소성 데이터를 더 많이 제거하여 예측의 정확도가 좋아짐을 보여준다. 그리고 이들 집단 간에는 통계적으로 유의적인 차이가 있으며 모든 집단에서 응답수에 따라 희소성 데이터를 제거한 집단의 MAE 평균값이 작아짐이 조사되었다.

4.3 연구결과

본 논문에서는 실험dataset1과 실험dataset2에서 선호도가 적은 희소상품을 응답수에 따라 4집단으로 나누어 $data_1$, $data_2$, $data_3$ 그리고 $data_4$로 정하였다.

$$data_x = \begin{cases} user_x & if \quad x \in C_j^-(i) \\ step_x & if\ not \end{cases} \qquad i = 1, 2, 3, 4 \tag{6}$$

여기에서, x는 응답수이다.

$user_1$은 응답수가 1이하인 상품인 $C_j^-(1)$에 선호도를 표시한 고객 중에서 이들 고객에게 응답수가 1인 상품을 제외시킨 데이터의 집합이다. 마찬가지 방법으로 응답수가 2, 3 그리고 4이하인 상품인 $C_j^-(2)$, $C_j^-(3)$ 그리고 $C_j^-(4)$에 선호한 고객을 추출한 후에 이들 상품을 제외시킨 고객 데이터의 집합을 $user_2$, $user_3$ 그리고 $user_4$라 한다.

$step_1$은 응답수가 1이하인 상품에 선호도를 표시하지 않은 고객의 집단이고, 응답수가 2, 3 그리고 4이하인 상품을 선호하지 않은 않은 고객 집단을 $step_2$, $step_3$ 그리고 $step_4$으로 정하였다. $data_1$, $data_2$, $data_3$ 그리고 $data_4$는 이들 두집단으로 만들어진 고객집단이다. 이 집단들에 대하여 예측의 정확도를 알아보기 위해 test dataset을 이용한 대응평균 검정 결과는 <표 4>와 같다.

<표 4> 실험dataset1에서 분류된 집단 간 test dataset을 이용한 MAE의 대응평균 검정결과

집단	MAE		N	t값	유의확률
	무조건	dataset			
$data_1$	0.5760	0.5759	942	1.68	0.02*
$data_2$	0.5760	0.5757	942	2.89	0.00**
$data_3$	0.5760	0.5753	942	4.14	0.00**
$data_4$	0.5760	0.5750	942	4.67	0.00**

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

실험dataset1의 경우는 $data_1$의 MAE 평균값은 0.5759으로 조건 없이 사용된 MAE 평균값인 0.5760 보다 약간 작으며, $data_2$와 $data_3$의 MAE 평균값은 0.5757과 0.5753

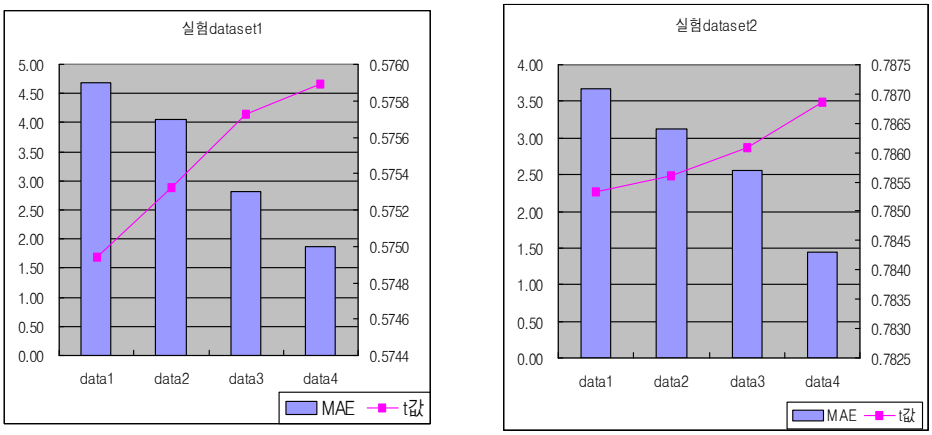
으로 두집단 모두 무조건인 경우보다 작았다. 그리고 집단 간 MAE 평균값은 단측검정결과 유의적인 차이가 있음이 조사되었다. $data_4$인 경우 0.5750으로 다른 집단들보다 가장 작은 값을 가지며, 희소성 데이터를 가장 많이 제거하여 MAE의 값이 가장 작아 추천시스템의 예측정확도가 개선됨을 알 수 있다. 따라서 본 논문에서 제시한 응답수에 따라 나누어진 집단에 대한 MAE 평균값은 응답수가 1인 경우보다는 4인 경우가 더 좋아졌으며, 기존의 방법보다 모두 개선되었음을 알 수 있다.

<표 5> 실험dataset2에서 분류된 집단 간 test dataset을 이용한 MAE의 대응평균 검정결과

집단	MAE		N	t값	유의확률
	무조건	dataset			
$data_1$	0.7879	0.7871	943	2.27	0.01*
$data_2$	0.7879	0.7864	943	2.48	0.00**
$data_3$	0.7879	0.7857	943	2.87	0.00**
$data_4$	0.7879	0.7843	943	3.49	0.00**

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

실험dataset2의 경우도 실험dataset1과 마찬가지로 $data_1$의 MAE 평균값은 0.7871로 기존의 방법으로 사용된 MAE 평균값인 0.7879보다 좋아졌다. 응답수가 1에서 4로 증가할수록 집단간은 통계적으로 유의적인 차이가 있음을 알 수 있다. 실험dataset1과 마찬가지로 $data_4$인 경우 0.7843으로 다른 집단들보다 MAE의 평균값이 가장 작은 값을 가짐을 알 수 있다. 이는 희소성 상품을 가장 많이 제거한 $data_4$의 MAE가 가장 많이 향상되었음을 의미한다. 따라서 본 논문에서 제시한 응답수에 따라 나누어진 집단에 대한 MAE 평균값은 기존의 방법보다는 모두 좋아졌음이 조사되었다. 실험dataset1과 실험dataset2에서 MAE와 t값에 대한 변화량을 구분된 집단에서 살펴보면 다음과 같다.



<그림 4>실험dataset1과 실험dataset2의 MAE와 t값의 분포

본 논문에서 사용된 집단은 희소성 데이터를 제거하는 방법에 따라 MAE가 차이가 있으며 희소성 데이터를 보다 많이 제거한 집단이 그렇지 않은 집단보다 MAE가 향상됨을 알 수 있다.

5. 결론

웹 추천시스템에서 사용되고 있는 협력적 여과기법은 고객과 이웃고객간의 선호도 평가치를 사용하여 예측 값을 생성하므로 선호도 평가치가 추천시스템의 성능에 영향을 미친다. 본 논문에서는 평가치가 적어 추천시스템의 신뢰를 떨어뜨리는 data들을 희소상품이라 정한 후, 응답수에 따라 4개의 집단으로 나누어 희소성 데이터를 제거한 집단들이 추천시스템의 예측 정확도를 높일 수 있음을 실험을 통해 살펴보았다. 분석결과 조건에 따라 선별된 4개의 dataset 모두 기존의 방법보다 예측의 정확도가 높아짐이 밝혀졌고, 그중에 가장 많이 희소성 데이터를 제거한 응답수가 4인 경우가 응답수가 1인 경우보다 예측의 정확도가 더 많이 개선되었음을 알 수 있었다.

따라서 본 연구는 추천 시스템의 예측 정확도를 높이기 위한 하나의 방법으로 선호도가 작은 상품을 선별하여 이들에 대한 기준을 정한 후 이들 데이터를 희소성이 제거된 선별된 집단에서 추천시스템을 사용하면 MAE가 개선됨을 알 수 있다.

참고문헌

1. 강현철, 한상태, 정병주, 신연주 (2004). 개인화를 위한 추천시스템 알고리즘에 관한 연구, *Journal of the Korean Data Analysis Society,* Vol. 6, No. 4,pp.1043-1049.
2. 이석준, 김선옥 (2007). 협업필터링에서 고객의 평가치를 이용한 선호도 예측의 사전평가에 관한 연구, 경영정보학연구, Vol. 17, No. 42, pp. 187-206.
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. 이희춘 (2006). On the Effect of Significance of Correlation Coefficient for Recommender System, *Journal of the Korean Data & Information Science Society*, Vol. 17, No. 4, pp. 1129-1139.
5. 이석준, 김선옥, 이희춘 (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.
6. 이희춘, 이석준, 정영준 (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.
7. Kim, H. G. and Kim, J. T. (2005). Modifying Sparse Date for

Collaborative Filtering, *Journal of The Korean Society of Computer Information*, Vol. 32, No. 1, pp. 610-613.
8. Kim, S. O. and Lee, S. J. (2007). The Effect of Data Sparsity on Prediction Accuracy in Recommender System, *Journal of the Korean Society for Internet Information*, Vol. 8, No. 6, pp. 9-15.
9. Kim, S. O., Lee, S. J. and Lee, H. C. (2008). A study on Improvement of Prediction Accuracy by Critical Value, *Journal of the Korean Data Analysis Society,* Vol. 10, No. 1, pp. 591-601.
10. Konstan, B., Miller, D., Maltz, J., Herlocker, L., Gordon, J. and Riedl, J. (1997). GroupLens: *Applying Collaborative Filtering to Usenet News, Communications of the ACM,* Vol. 40, No. 3, pp. 77-87.
11. Melville, P., Mooney, R. and Nagarajan, R. (2002). Content-Boosted Collaborative Filtering for Improved Recommendations, *Proceedings of the eighteenth national Conference on Artificial Intelligence,* pp. 187-192.
12. Pazzani, M. J. (1999). Framework for Collaborative, Content_Based and Demograghic Filtering, *Artificial Intelligent Review,* pp. 394-408.
13. Resnick, P. N., Iacovou, M., Suchak, P., Bergstrom, J. 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.
14. Soboroff, I. and Nocholas, C. (1999). Combining content and collaborative in text filtering, *Preceedings of the IJCAI Workshop on Machine Learning in Information Filtering,* pp. 86-92.

[접수일(2008년 6월 20일), 수정일(2008년 7월 16일), 게재확정일(2008년 7월 20일)]