경영정보학연구
제17권 제4호
2007년 12월

협업필터링에서 고객의 평가치를 이용한 선호도 예측의 사전평가에 관한 연구

이 석 준*, 김 선 옥**

Pre-Evaluation for Prediction Accuracy by Using the Customer's Ratings in Collaborative Filtering

Seok Jun Lee, Sun Ok Kim

The development of computer and information technology has been combined with the information superhighway internet infrastructure, so information widely spreads not only in special fields but also in the daily lives of people. Information ubiquity influences the traditional way of transaction, and leads a new E-commerce which distinguishes from the existing E-commerce. Not only goods as physical but also service as non-physical come into E-commerce. As the scale of E-Commerce is being enlarged as well. It keeps people from finding information they want. Recommender systems are now becoming the main tools for E-Commerce to mitigate the information overload.

Recommender systems can be defined as systems for suggesting some Items (goods or service) considering customers' interests or tastes. They are being used by E-commerce web sites to suggest products to their customers who want to find something for them and to provide them with information to help them decide which to purchase. There are several approaches of recommending goods to customer in recommender system but in this study, the main subject is focused on collaborative filtering technique.

This study presents a possibility of pre-evaluation for the prediction performance of customer's preference in collaborative filtering before the process of customer's preference prediction. Pre-evaluation for the prediction performance of each customer having low performance is classified by using the statistical features of ratings rated by each customer is conducted before the prediction process.

In this study, MovieLens 100K dataset is used to analyze the accuracy of classification. The classification

* 상지대학교 경상대학 경영학과 겸임교수
** 한라대학교 정보통신공학부 교수

criteria are set by using the training sets divided 80% from the 100K dataset. In the process of classification, the customers are divided into two groups, classified group and non classified group. To compare the prediction performance of classified group and non classified group, the prediction process runs the 20% test set through the Neighborhood Based Collaborative Filtering Algorithm and Correspondence Mean Algorithm. The prediction errors from those prediction algorithm are allocated to each customer and compared with each user's error.

Research hypothesis

Two research hypotheses are formulated in this study to test the accuracy of the classification criterion as follows.

Hypothesis 1: The estimation accuracy of groups classified according to the standard deviation of each user's ratings has significant difference.

To test the Hypothesis 1, the standard deviation is calculated for each user in training set which is divided 80% from MovieLens 100K dataset. Four groups are classified according to the quartile of the each user's standard deviations. It is compared to test the estimation errors of each group which results from test set are significantly different.

Hypothesis 2: The estimation accuracy of groups that are classified according to the distribution of each user's ratings have significant differences.

To test the Hypothesis 2, the distributions of each user's ratings are compared with the distribution of ratings of all customers in training set which is divided 80% from MovieLens 100K dataset. It assumes that the customers whose ratings' distribution are different from that of all customers would have low performance, so six types of different distributions are set to be compared. The test groups are classified into fit group or non-fit group according to the each type of different distribution assumed. The degrees in accordance with each type of distribution and each customer's distributions are tested by the test of χ^2 goodness-of-fit and classified two groups for testing the difference of the mean of errors. Also, the degree of goodness-of-fit with the distribution of each user's ratings and the average distribution of the ratings in the training set are closely related to the prediction errors from those prediction algorithms.

Through this study, the customers who have lower performance of prediction than the rest in the system are classified by those two criteria, which are set by statistical features of customers ratings in the training set, before the prediction process.

Keywords : Recommender System, Collaborative Filtering, Pre-evaluation for prediction, Pre-information

Ⅰ. 서 론

초고속 인터넷 인프라와 결합한 컴퓨터의 발달과 정보기술의 발전은 특정 분야에서의 정보화뿐만 아니라 일상생활에서도 정보화를 이루고 있다. 정보의 보편화는 전통적인 상거래 방식에도 영향을 주어 기존의 상거래방식과 구분되는 전자상거래를 활성화시키고 있다. 전자상거래는 유형의 상품뿐만 아니라 무형의 서비스까지 다양한 형태의 상품들이 거래되고 있다. 전자상거래 시장의 규모가 급속히 확대됨에 따라 거래되는 상품 또한 종류와 수가 더불어 방대해지고 있다. 통계청의 조사 결과에 따르면, 2006년 4/4분기 사이버 쇼핑몰 거래액은 3조 6251억 원으로 3/4분기에 비해서 1711억 원(5.0%)이 증가하였고 2005년 4/4분기에 비해서는 5408억 원(17.5%)으로 거래액이 증가한 것으로 나타나고 있다[한국인터넷진흥원, 2007]. 전자상거래 이용자들은 규모가 확대된 시장 환경에 적합한 검색환경을 선호하게 되었으며 또한 자신의 취향이나 성향을 고려하여 상품을 선택할 수 있는 시스템을 더욱 요구하고 있다. 추천시스템은 급속히 규모가 확대되고 있는 전자상거래 환경에서 방대한 양의 상품정보 중 개별 고객의 취향을 고려한 개인화 서비스를 제공할 수 있으며 상거래 상품정보 과잉의 문제를 완화시켜 이용자의 편의와 만족도를 높일 수 있는 시스템으로 많은 전자상거래 기업들이 도입하여 상용화하고 있다. 추천시스템 중 상업적으로 성공한 접근법이 협업필터링 접근법이며 amazon.com 등 유수의 전자상거래 기업에서 성공적인 성과를 거두고 있는 것으로 알려져 있으며 학문적으로도 많은 연구가 진행되고 있다. 협업필터링을 통한 선호도 예측의 정확도를 향상시키기 위해 이웃선정 기법, 유의성 가중치의 적용 등과 같은 다양한 방법이 제시되어 협업필터링의 예측 정확도를 높이는 효과를 얻을 수 있었다. 본 연구에서는 협업필터링의 예측 정확도를 높이기 위한 방법과 달리 선호도 예측의 오차가 클 것으로 예상되는 고객들을 선호도 예측 이전에 선별할 수 있는 방법에 대하여 연구하였다. 선호도 예측 이전에 선호도 예측 오차가 클 것으로 예상되는 고객의 선별은 협업필터링의 성능을 향상시킬 수 있는 정보로 이용할 수 있을 뿐만 아니라 선별된 고객들의 특성 연구에도 중요한 단서를 제공할 수 있을 것이다.

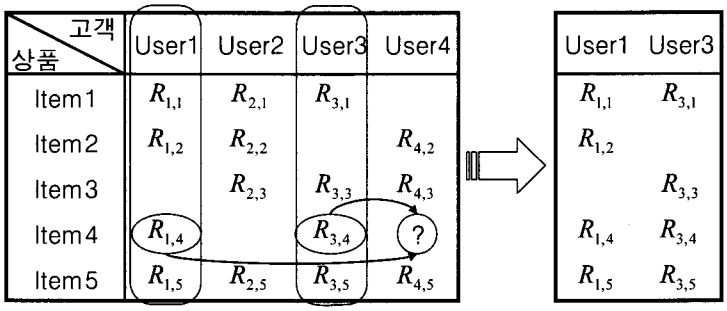
Ⅱ. 이론적 배경

2.1 추천시스템

추천시스템은 인터넷의 급속한 보급과 전자상거래가 급속하게 확대되고 있는 환경에서 전자상거래 업체들의 고객관계관리 전략을 구현하기 위한 대표적인 도구로써 적용되고 있다. 추천시스템은 개별 고객에게 자신의 성향과 취향에 적합한 상품 정보를 자동적으로 제공하여 방대한 상품 정보에서 자신들에게 필요한 정보만을 찾을 수 있도록 도와주는 시스템이다. 추천시스템은 고객의 취향이나 선호도를 사전에 예측하여 고객의 선호도에 적합한 상품을 추천함으로써 고객들의 정보탐색 비용을 절감시켜 추천시스템에 대한 고객의 충성도를 높이며 마케팅과 상품의 관리 등의 주요 자료로 활용할 수 있는 기회를 제공한다. 추천시스템은 1990년대 중반 상품에 대한 고객들의 명시적 선호도를 예측하기 위한 연구가 시작되면서 이전의 인지과학, 정보검색, 예측이론 등과 구분되는 독립적인 분야로 자리 잡았다[Adomacicius and Tuzhilin, 2005]. 고객에 대한 상품 추천 방법으로 내용기반의 접근법, 협업필터링 접근법, 혼합 접근법[Claypool *et al.*, 1999]등의 다양한 접근법이 도입되어 추천시스템에 대한 연구가 진행되었다. 초기 시스템에서는 내용기반의 접근법이 적용되기도 하였지만 일반적으로 협업필터링이 보편적으로 적용되고 있다.

2.2 협업필터링

추천시스템에서 가장 널리 이용되는 접근법이 협업필터링이다. 협업필터링은 특정 상품에 대한 추천대상 고객의 선호도를 예측하기 위하여 시스템 내의 다른 고객들이 그 상품에 대해 이전에 평가한 선호도 평가치를 기준으로 추천대상 고객의 선호도를 예측한다. 이러한 협업필터링 접근법은 내용기반 접근법의 단점인 상품 속성의 문자화, 고객 자신의 경험에 대한 편중, 타 고객들의 다양한 특성을 반영하지 못하는 문제[Balabanovic and Shoham, 1997; 김용수, 2006]를 개선하였다. 협업필터링 접근법은 일반적으로 고객의 개인적 취향, 정보 등과 같은 부가적인 정보와 상품이 가지고 있는 복잡한 정보를 의도적으로 무시하고 상품과 고객과의 관계에 대한 정보인 선호도를 이용하여 상품을 추천하며 이러한 선호도는 일반적으로 수치화된 선호도 평가치를 이용한다[Hill *et al.*, 1995; Resnick *et al.*, 1994; Shardanand and Maes, 1995]. 협업필터링 기법은 상품과 고객의 관계인 선호도 평가치를 행렬 자료의 형태로 구성하여 분석하는데 고객들 간의 관계를 이용하여 선호도 예측을 실시하는 사용자 기반(user-based)의 접근법[Claypool *et al.*, 1999; Sarwar *et al.*, 1998]과 상품들 간의 관계를 이용하는 아이템 기반(item-based)의 접근법[Deshpande and Karypis, 2004]으로 나누어진다. 또한 협업필터링 기법을 적용한 추천시스템의 예측 정확성을 높이기 위하여 이웃 구성의 문제, 연관규칙, 베이지안 네트워크, 군집화 모형 등의 방법들이 적용되었다. 김경재와 김병국[2005]은 유전자 알고리즘을 이용하여 고객 정보에서 성향을 추출할 수 있는 추천시스템의 추천엔진 개발에 대하여 연구하였다. 김재경 등[2003]은 거래기록에서 상품과 고객 간의 연관규칙의 패턴을 끌어내기 위하여 연관규칙 마이닝 기법을 추천시스템에 적용하였으며 또한 심장섭[2005]은 K-means 군집화 알고리즘을 이용하여 거리개념의 군집을 생성한 후 군집간의 순차적 패턴을 발견하여 이를 추천에 적용시키는 연구를 하였다. 손재봉, 서용무[2006]의 연구에서는 두 고객의 관계에서 공통으로 선호도를 평가한 평가치의 개수를 이용하여 DOM(degree of match)를 정의하고 이에 따른 가중치를 적용하여 협업필터링의 성능을 개선하였다. 이석준, 이희춘[2007]의 연구에서는 근접이웃 알고리즘을 개선한 대응평균 알고리즘을 적용하여 협업 필터링의 성능을 개선하였고 또 이석준 등[2007a]의 연구에서는 고객의 선호도 평가 패턴 중 평가시간에 따른 Run의 유무가 선호도 예측에 미치는 영향에 대하여 연구하였다. 최근 협업필터링의 연구에서 O'Mahony *et al.*[2006]은 추천시스템의 예측 성능을 저하시키는 요인으로 노이즈(noise)를 정의하고 자연적 노이즈와 악의적 노이즈로 구분하여 임계치 설정을 통한 노이즈의 발견에 대하여 연구하였다. Burke *et al.*[2006]은 시스템의 신뢰도를 떨어뜨리는 요인으로 인위



<그림 1> 추천대상 고객의 이웃 선정과정

적인 공격(attack)의 유형을 정의하였으며 그 가능성에 대해 공격모형을 설정하고 그에 따른 임의의 프로파일을 삽입하여 선호도 예측에 미치는 효과에 대하여 연구하였다.

협업필터링에서 특정 상품에 대한 추천대상 고객의 선호도 예측은 그 상품에 대해 선호도를 평가한 시스템내의 고객들인 이웃고객을 선정하고 이웃고객이 평가한 선호도 평가치와 추천대상 고객의 선호도 평가치를 이용하여 특정 상품에 대한 선호도를 예측한다[Herlocker *et al.*, 2002]. <그림 1>은 선호도 예측 대상 고객의 이웃 선정과정을 나타내며 고객과 상품의 관계는 일반적으로 행렬 구조로 나타내어진다.

<그림 1>의 이웃 선정과정은 특정 상품에 대한 추천대상 고객의 선호도를 예측하기 위하여 그 상품에 대하여 이미 선호도를 평가한 이웃 고객들을 선정하는 과정이다. <그림 1>에서 상품 Item4에 대한 추천대상 고객 User4의 선호도를 예측하기 위하여 해당 상품인 Item4에 대하여 선호도를 평가한 User1과 User3이 추천대상 고객인 User4의 이웃으로 선정된다. 선정된 이웃과 추천대상 고객의 선호도나 취향의 유사 정도를 계산하기 위하여 다양한 형태의 가중치가 설정될 수 있으며 일반적으로 벡터 유사도와 상관계수를 유사도 가중치로 많이 사용한다[Breese *et al.*, 1998; 이희춘과 이석준, 2006]. 본 연구에서는 이전 연구에서 예측 성과가 우수한 피어슨 상관계수만을 이용하여 두 고객의 유사도 가중치로 정의하였다. 두 고객의 유사도 가중치인 r_{uj}는 다음 식 (1)과 같이 정의한다.

$$r_{uj}=\frac{\sum_{i=1}^{m}\left(R_{ui}-\overline{R_u}\right)\left(R_{ji}-\overline{R_j}\right)}{\sqrt{\sum_{i=1}^{m}\left(R_{ui}-\overline{R_u}\right)^2 \cdot \sum_{i=1}^{m}\left(R_{ji}-\overline{R_j}\right)^2}}, -1 \le r_{uj} \le 1 \quad (1)$$

2.3 선호도 예측 알고리즘

협업필터링에서 최초의 자동화된 선호도 예측 알고리즘인 이웃 기반의 협업필터링 알고리즘(Neighbor Based Collaborative Filtering Algorithm)은 GroupLens에서 유즈넷 뉴스(UseNet News) 그룹의 기사를 추천하기 위해 제안하였다[Resnick *et al.*, 1994]. NBCFA는 유즈넷 뉴스의 기사에 대한 선호도 평가치를 이용하여 이웃 고객과의 선호도 유사정도를 피어슨 상관계수로 정의하여 선호도 평가치를 예측하였다. NBCFA는 다음 식 (2)와 같다.

$$\widehat{U_x}=\overline{U}+\frac{\sum_{J\in Raters}\left(J_x-\overline{J}\right)r_{uj}}{\sum_{J\in Raters}|r_{uj}|},\ where\ \overline{J}=\frac{\sum_{i=1}^{n}J_i}{n}, i\neq x \quad (2)$$

여기서,

- $\widehat{U_x}$: 선호도 예측 대상 상품 x에 대한 선호도 예측 대상 고객 u의 선호도 예측
- $\overline{U}$: 선호도 예측 대상 고객 u가 평가한 모든 상품에 대한 평균
- J_x : 선호도 예측 대상 상품 x에 대한 이웃 고객 j의 선호도 평가치
- $\overline{J}$: 이웃 고객 j가 평가한 모든 상품에서 선호도 예측 대상 상품 x에 대한 평가치를 제외한 선호도의 평균
- r_{uj} : 선호도 예측 대상 고객 u와 이웃 고개 j의 선호도 유사 정도를 나타내는 유사도 가중치

이석준, 이희춘[2006]은 GroupLens에서 제안한 NBCFA를 개선한 알고리즘인 대응평균 알고리즘(Correspondence Mean Algorithm)의 예측 정확도가 더 우수함을 보였다. 다음 식 (3)은 CMA이다.

$$\hat{U}_x = \overline{U}_{match} + \frac{\sum_{J \in Raters} (J_x - \overline{J}_{match}) r_{uj}}{\sum_{J \in Raters} |r_{uj}|} \quad (3)$$

여기서,

$\overline{U}_{match}$: 선호도 예측 대상 고객 u와 각 이웃 고객 j가 공통으로 평가한 상품들의 평가치의 평균들의 평균

$\overline{J}_{match}$: 선호도 예측 대상 고객 u와 이웃 고객 j가 공통으로 평가한 상품들에 대한 선호도 평가치의 평균

NBCFA와 CMA의 근본적인 차이는 추천대상 고객의 선호도를 나타내는 추천대상 고객의 선호도 평가치 평균인 $\overline{U}$와 $\overline{U}_{match}$에 있으며 또한 이웃 고객들의 선호도 평가치 평균인 $\overline{J}$와 $\overline{J}_{match}$에 있다. NBCFA에서의 $\overline{U}$는 추천대상 고객이 평가한 모든 상품들의 평가치를 이용한 평균을 사용하지만 CMA에서의 $\overline{U}_{match}$ 선정된 개별 이웃 고객과 공통으로 평가한 상품들의 평균들을 다시 평균하여 사용한다는 점에서 차이가 있다. 즉, 선정된 이웃의 수만큼 두 고객의 상관계수를 구할 경우에 사용된 평균들의 평균을 사용하기 때문에 자신의 선호도가 너무 과대 혹은 과소 평가되는 것을 조정한다. 또한 이웃 고객의 평균도 공통으로 평가한 상품의 평가치들만 이용하기 때문에 이웃 고객의 선호도가 편향되어 평가되는 것을 조정한다.

2.4 선호도 예측 정확도 평가척도

MAE(Mean Absolute Error)는 협업필터링에 의한 예측치의 성능을 평가하기 위해 가장 일반적으로 적용되는 평가 척도이다. MAE는 계산된 선호도 예측치와 이에 대응하는 실제 선호도 평가치의 절대 편차의 평균으로 계산된다[Breese *et al.*, 1998; Shardanand and Maes, 1995].

$$MAE = \frac{1}{N}\sum_{j=1}^{N} |R_{uj} - \hat{R}_{uj}| \quad (4)$$

여기서, N은 추천을 받을 모든 고객들에 대한 예측의 총 개수를 나타내며 R_{uj}는 실제 선호도 평가치이고 $\hat{R}_{uj}$는 R_{uj}에 대응하는 선호도 예측치이다. MAE에 의한 성능평가의 결과는 MAE가 낮을수록 전체 예측 알고리즘의 정확도가 높다. MAE와 유사한 평가 척도로 MSE(Mean Squared Error), RMSE(Root Mean Squared Error), 그리고 MAE를 표준화 시킨 NMAE(Normalized Mean Absolute Error)등이 있으며 일반적으로 전체 시스템의 정확도는 MAE를 이용하여 성능을 평가한다.

Ⅲ. 가설설정 및 실험설계

3.1 실험 dataset

본 연구는 협업필터링에서 예측 알고리즘을 이용한 선호도 예측 이전의 사전자료인 선호도 평가치의 통계적 특성을 이용하여 예측 오차의 사전평가 가능성을 실험하기 위하여 GroupLens에서 공개하는 MovieLens 100K dataset을 이용하여 분석하였다. MovieLens 100K dataset은 943명의 고객이 1682편의 영화에 대하여 선호도를 평가한 자료로 총 100,000개의 평가치로 구성되어 있다. MovieLens 100K dataset은 고객에 대한 인구통계 정보와 영화에 대한 장르 및 간단한 정보로 구성되어 있으며 각 고객들은 최소 20편의 영화에 대하여 평가하였다. 본 연구에서 제안하는 사전평가 방법의 효과를 분석하기 위하여 MovieLens 100K dataset을 80%의 훈련집합(training set)과 20%의 실험집합(test set)으로 구분한 2개

의 dataset을 구성하였다. 본 연구는 사전정보를 이용한 선호도 예측 정확도의 사전평가 가능성을 검정하기 위하여 가설을 설정하고 각 dataset의 훈련집합의 선호도 평가치를 이용하여 분류기준을 설정하고 이에 따라 선별된 고객과 그렇지 않은 고객들 간의 예측 결과의 차이를 실험집합에 대한 선호도 예측 결과와 비교하여 검정하였다.

3.2 연구가설

본 연구는 추천시스템에서 선호도 예측 이전에 주어진 사전정보인 고객의 선호도 평가치를 이용하여 개별 고객의 선호도 예측 오차를 사전에 평가하여 예측 오차가 크게 나타나는 고객을 선별 할 수 있는 방법에 대하여 연구하였다. 기존의 연구에서는 선호도 예측의 정확도를 향상시키기 위한 선호도 예측 알고리즘의 연구와 이웃의 선정에 대한 연구가 많이 진행되었으며 이를 통한 추천시스템의 추천 성능 향상을 위한 다각적인 접근법이 진행되었다[이석준, 이희춘, 2007; Kim and Yang, 2005]. 또한 전자상거래에서 발생하는 거래 data의 희소성으로 인하여 선호도 예측 정확도가 낮아지기 때문에 이를 개선하기 위한 방법들이 진행되었으며 data의 차원을 감소시키기 위한 SVD등이 적용되었다[김종우 등, 2004; Huseyin and Wenliang, 2005]. 그러나 선호도 예측 정확도에 영향을 미치는 영향에 대한 연구는 협력적 필터링 알고리즘의 보안성 취약이라는 주제로 최근에 연구가 진행되고 있으며 특히 Burke *et al.*[2005]의 연구에서는 선호도 예측에 영향을 줄 수 있는 악의적 고객이 전개할 수 있는 공격의 유형을 정의하고 각 유형에 대한 영향을 분석하였으며 이를 해결하기 위한 선호도 예측 알고리즘 차원에서의 접근법을 연구하고 있다[Burke *et al.*, 2005]. O'Mahony *et al.* [2006]의 연구에서는 이러한 악의적 영향을 필터링하기 위한 방법을 제시하고 있지만 선호도 예측 결과를 이용하여 필터링하는 접근법을 제시하고 있기 때문에 선호도 예측 이전에 이를 찾아내는 방법에 대하여는 제안하지 못하였다. 그러나 이석준 등[2007b]의 연구에서 개별 고객이 평가한 선호도 평가치들의 기초 통계량과 선호도 예측 결과의 상관성에 대한 연구를 통하여 고객이 평가한 선호도 평가치의 표준편차가 선호도 예측의 정확도와 관련성이 높음을 보이고 있다. 이석준 등[2007b]의 연구를 바탕으로 고객이 평가한 선호도 평가치의 편차 정도가 선호도 예측의 정확도에 영향을 미치고 있음을 가정할 수 있다. 또한 Lee *et al.*[2007a]의 연구에서는 고객이 평가한 평가치의 특정 비율이 선호도 예측 정확도와 매우 밀접한 관련성이 있음을 연구하였으며 이 기준을 예측 정확도가 낮은 고객 분류의 기준으로 선정할 수 있음을 제안하였다. 그러나 Lee *et al.*[2007a]의 연구의 기준에 따라 선정된 고객은 특정 선호도 평가치 비율로 선정되었기 때문에 일반화시키기 어렵다. 결과에서 제시된 기준에 따라 선정된 고객들의 평균 평가치 분포는 전체 dataset의 분포와 상이한 유형을 나타내고 있으며 1과 5의 극단치의 비율이 높음을 보이고 있다. 또한 통계적 접근법은 무작위로 추출된 표본은 모집단의 성격을 나타내고 있음을 전제로 진행되기 때문에 본 연구에서 전체 MovieLens 100K dataset에서 개인별로 무작위 추출된 80%의 훈련집합과 20%의 실험집합은 전체 dataset의 성격을 가지고 있을 것으로 가정할 수 있다. 결국 개별 고객이 평가한 선호도 평가치들의 표준편차가 크면서 전체 고객들의 평가치 분포와 상이한 형태를 가진 고객들이 선호도 예측 결과가 나빠질 수 있음을 가정할 수 있다. 관련 연구를 통하여 선호도 예측 이전에 예측 정확도가 낮을 것으로 기대되는 고객을 선정하기 위하여 다음과 같이 가설을 설정하였다.

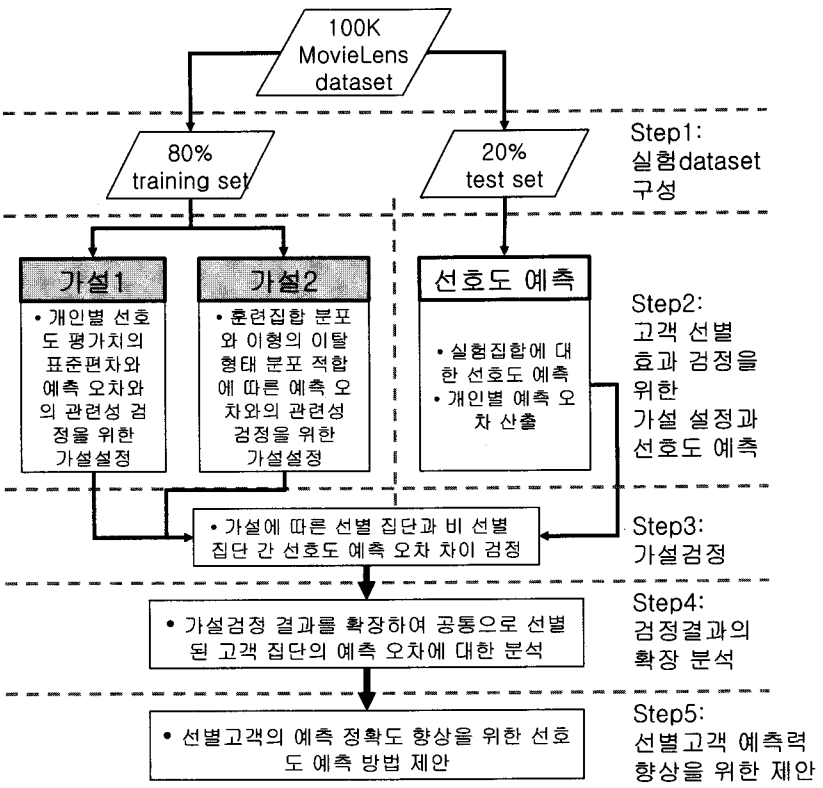
가설 1: 고객 선호도 평가치의 표준편차에 따라 집단들의 선호도 예측 정확도는 유의적

인 차이가 있을 것이다.

가설 2: 전체 고객의 선호도 평가치 분포 유형과 다른 유형의 분포 유형을 갖는 고객들의 예측 정확도는 유의적인 차이가 있을 것이다.

가설 1을 검정하기 위하여 훈련집합의 선호도 평가치를 고객별로 나누어 표준편차를 계산하였다. 계산된 표준편차에 따라 표준편차의 크기에 따른 예측 정확도의 영향을 분석하기 위하여 집단을 구분하였다. 집단 구분 기준의 선정은 다양한 방법을 적용시킬 수 있다. 고객별 표준편차들의 분포에서 정규분포의 확률밀도에 따라 표준편차의 배수로 구분하는 방법을 사용할 수 있으나 이 방법은 선별 고객 집단의 비율이 달라지므로 본 연구에서는 표준편차들의 100분위수를 기준으로 동일 비율로 25%의 4개 집단으로 구분하여 연구 결과를 검정하였다. 가설검정의 진행은 <그림 2>의 실험설계의 진행과정에 따라 이루어진다.

MovieLens 100K dataset에서 고객별로 무작위로 추출된 80%의 훈련집합은 100K dataset의 분포 유형과 유사한 형태를 나타내고 있으며 20%의 실본집합의 분포 유형도 전체 100K dataset의 분포 유형과 유사할 것이기 때문에 가설 2를 검정하기 위하여 훈련집합의 선호도 평가치 분포와 상이한 형태의 분포를 가정하였다. 본 연구에서는 훈련집합의 선호도 평가치 분포와 다른 형태의 분포를 이탈형 분포로 정의하고 6개의 유형으로 구분하였다. 6개의 이탈유형의 분포와 개별 고객의 선호도 평가치의 분포와의 적합정도를



<그림 2> 실험 설계

이용하여 각 유형별로 고객을 분류하였다. 이탈유형과의 적합정도를 통계적으로 구분하기 위하여 본 연구에서는 χ^2 분포적합도 검정을 실시하였으며 유의수준 0.05를 기준으로 적합여부를 결정하여 선별하였다. 분류된 고객 집단 간 예측 정확도의 차이를 검정하기 위하여 독립2표본 t검정을 실시하여 가설을 검정한다.

3.3 가설검정을 위한 실험설계

전술한 가설을 검정하기 위하여 다음 <그림 2>와 같이 실험을 진행하였다. 먼저 단계 1에서는 MovieLens 100K dataset을 고객별로 평가치의 80%를 훈련집합으로, 20%를 실험집합으로 분할한 dataset을 구성하였으며 실험의 신뢰성을 높이기 위하여 2개의 dataset을 구성하여 실험을 진행하였다. 단계 2에서는 가설 1의 검정을 위하여 80%의 훈련집합에서 고객별 선호도 평가치들의 표준편차를 구하고 각 고객별 표준편차의 크기에 따라 고객들을 4집단으로 분류하였다. 가설 2의 검정을 위하여 훈련집합의 선호도 평가치의 분포를 확인하고 이 분포에 이탈되는 유형의 분포를 6개 가정하여 고객별 선호도 평가치의 분포와 이탈유형의 분포가 잘 적합 되는지를 χ^2 분포적합도 검정을 통하여 통계적으로 유의한 고객들 집단과 그렇지 아니한 집단으로 분류하였다. 단계 3에서는 각 가설에 따라 선별된 고객집단의 개인별 MAE가 선별되지 않은 고객집단의 개인별 MAE와 차이가 있는지를 통계적으로 분석하였다. 단계 4에서는 가설검정 결과를 확장하여 공통으로 선별된 고객 집단의 MAE가 그렇지 않은 집단의 MAE와 차이가 있는지를 분석하였다.

Ⅳ. 실험을 통한 가설검정

4.1 가설검정을 위한 선호도 예측

가설검정을 위하여 dataset1과 dataset2의 훈련집합을 NBCFA와 CMA를 이용하여 20%의 실험집합에 대하여 선호도 예측을 실시하여 943명에 대한 고객의 MAE를 계산하였다. 다음 <표 1>은 dataset 1과 dataset 2에서 선호도 예측 알고리즘별 개인 MAE의 분포이다.

<표 1> 개인 MAE 분포표

개인별 MAE	dataset1		dataset2	
	NBCFA	CMA	NBCFA	CMA
0.3 이하	0.6%	0.6%	1.0%	0.7%
0.3~0.5	8.7%	10.1%	8.8%	10.1%
0.5~0.8	49.9%	52.1%	51.0%	51.9%
0.8~1.1	29.1%	27.0%	28.0%	27.9%
1.1~1.4	9.5%	8.2%	9.0%	6.6%
1.4~1.7	1.6%	1.6%	2.0%	2.5%
1.7~2.0	0.5%	0.4%	0.2%	0.2%
평균	0.788	0.772	0.777	0.762

<표 1>에서 MAE의 분포는 0.5~0.8 범위에서 가장 높은 비율을 차지하고 있으며 시스템의 정확도를 평가하는 실험집합 전체 data의 MAE는 dataset1의 경우 NBCFA에서 0.753, CMA에서 0.736이고 dataset2의 경우 NBCFA에서 0.750, CMA에서 0.732로 분석되었다. 분석을 위한 개인별 MAE들의 평균은 <표 1>에서와 같이 분석되었으며 실험집합 전체 data에 대한 MAE 보다 높게 나타났다.

4.2 가설검정

4.2.1 가설 1의 검정

가설 1: 고객 선호도 평가치의 표준편차에 따라 집단들의 선호도 예측 정확도는 유의적인 차이가 있을 것이다.

가설 1을 검정하기 위하여 훈련집합에서 고객의 선호도 평가치를 이용하여 고객의 표준편차를 계산하였다. 1682편의 영화에 대한 고객의 평

<표 2> 표준편차에 대한 집단 간 평균검정 결과

dataset	알고리즘	집단구분	빈도	MAE 평균	F값	유의확률	Duncan
dataset1	NBCFA	집단1	234	0.6492	105.597	0.000**	{1}{2}{3}{4}
		집단2	238	0.7075			
		집단3	236	0.8165			
		집단4	235	0.9787			
		합계	943	0.7879			
	CMA	집단1	234	0.6434	90.355	0.000**	{1}{2}{3}{4}
		집단2	238	0.6964			
		집단3	236	0.7942			
		집단4	235	0.9528			
		합계	943	0.7716			
dataset2	NBCFA	집단1	235	0.6439	97.228	0.000**	{1}{2}{3}{4}
		집단2	236	0.7021			
		집단3	236	0.7971			
		집단4	235	0.9653			
		합계	942	0.7770			
	CMA	집단1	235	0.6395	77.705	0.000**	{1}{2}{3}{4}
		집단2	236	0.6865			
		집단3	236	0.7926			
		집단4	235	0.9311			
		합계	942	0.7624			

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

가치 개수는 최대 591개이고 최소 16개로 구성되어있다. 고객들의 표준편차를 4분위수를 기준으로 4개의 집단으로 분류하여 구분된 집단 간의 평균에 대하여 일원분산분석을 실시하고 Duncan의 다중비교 검정을 실시하였다. 다음 <표 2>는 개인별 선호도 평가치의 표준편차에 의해 분류된 4개 집단의 평균에 대한 일원분산분석과 사후검정 결과이다.

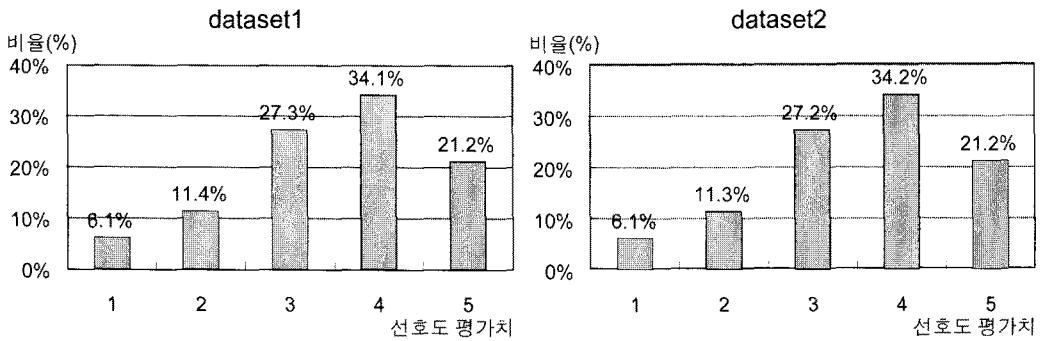
<표 2>에서 dataset1과 dataset2, 모두 표준편차의 크기에 따른 집단 간에 통계적으로 유의한 평균 차가 있음을 알 수 있으며 Duncan의 다중비교 검정 결과에서도 4개의 집단으로 잘 분류되어 있음을 알 수 있다. 또한 표준편차가 큰 4집단의 MAE의 평균이 높음을 알 수 있어 예측 이전의 사전정보인 고객 선호도 평가치의 표준편차에 따라 구분한 집단의 선호도 예측 정확도는 유의적인 차이가 있는 것으로 나타났다. 실험결과를 통하여 가설 1을 채택할 수 있다.

4.2.2 가설 2의 검정

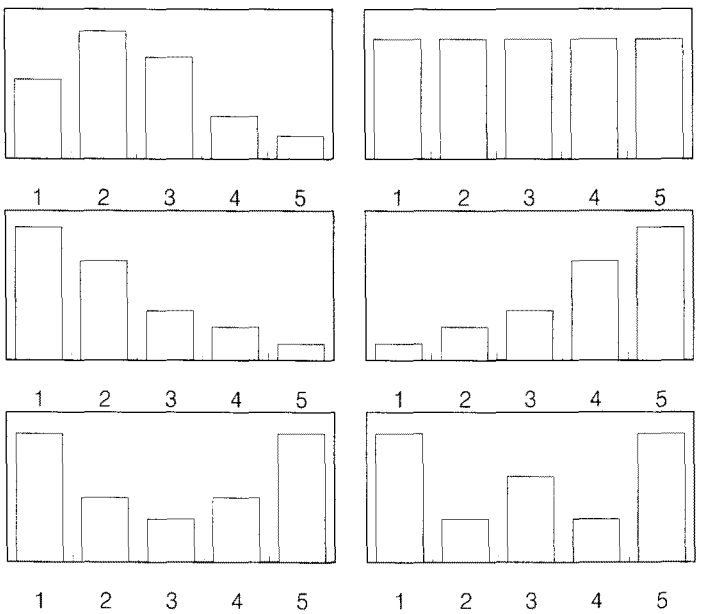
가설 2: 선호도 평가치의 분포유형에 따라 예측 선호도 평가치는 유의적인 차이가 있을 것이다.

가설 2를 검정하기 위하여 dataset 1과 dataset 2에서 훈련집합의 선호도 평가치의 분포를 살펴보았다. 다음 <그림 3>은 dataset 1과 dataset 2의 훈련집합의 선호도 평가치의 분포이다.

<그림 3>에서 dataset 1과 dataset 2의 훈련집



<그림 3> dataset1과 dataset2의 훈련집합의 선호도 평가치 분포도



<그림 4> 6가지 이탈 분포 유형

합의 선호도 평가치의 분포는 매우 유사한 유형을 나타내고 있음을 알 수 있다. 가설에서 훈련집합의 선호도 평가치 분포와 유사한 유형의 분포를 가지는 고객과 그렇지 않은 고객을 분류하기 위하여 다음 <그림 4>와 같이 훈련집합의 분포에 이탈하는 형태의 분포를 6가지로 가정하였다.

<그림 4>의 분포들은 각 dataset에서 훈련집합의 분포 유형에서 이탈할 것으로 가정한 분포 형태로 훈련집합의 분포와 대칭형의 분포를 이탈 제 1형으로 정의하고 균등분포의 유형을 이탈 제 2형, 감소형태의 분포 유형을 이탈 제 3형, 증가형태의 분포 유형을 이탈 제 4형, "V"형태의 분포 유형을 이탈 제 5형, "W"형태의 분포 유형을 이탈 제 6형으로 정의하였다. 훈련집합에서 정의된 이탈 유형의 분포와 적합한 선호도 평가치의 분포를 가진 고객들을 분류하기 위하여 χ^2 분포 적합도 검정을 실시하여 유의수준 0.05를 기준으로 이탈 분포에 적합한 유형의 집단과 그렇지 않은 집단으로 분류하였다. 각 유형별로 집단의 개인별 MAE의 평균 차를 검정하기 위하여 독립2

<표 3> dataset1과 dataset2에서 χ^2 분포적합도 검정에 따라 분류된 집단 간의 독립 2표본 t검정 결과

유형	알고리즘	집단 구분	dataset1				dataset2			
			빈도	MAE 평균	t값	유의확률	빈도	MAE 평균	t값	유의확률
이탈 제1형	NBCFA	비적합	908	0.7807	-4.538	0.000**	908	0.7709	-3.161	0.003**
		적합	35	0.9738			34	0.9398		
	CMA	비적합	908	0.7649	-4.300	0.000**	908	0.7569	-3.489	0.001**
		적합	35	0.9472			34	0.9091		
이탈 제2형	NBCFA	비적합	802	0.7425	-11.177	0.000**	801	0.7452	-8.216	0.000**
		적합	141	1.0461			141	0.9577		
	CMA	비적합	802	0.7279	-10.456	0.000**	801	0.7332	-7.169	0.000**
		적합	141	1.0203			141	0.9280		
이탈 제3형	NBCFA	비적합	931	0.7860	-2.039	0.042*	930	0.7750	-2.156	0.031*
		적합	12	0.9336			12	0.9311		
	CMA	비적합	931	0.7697	-2.099	0.036*	930	0.7607	-1.798	0.072
		적합	12	0.9210			12	0.8918		
이탈 제4형	NBCFA	비적합	700	0.7694	-3.889	0.000**	699	0.7646	-2.421	0.016*
		적합	243	0.8411			243	0.8126		
	CMA	비적합	700	0.7544	-3.485	0.001**	699	0.7483	-2.702	0.007**
		적합	243	0.8214			243	0.8030		
이탈 제5형	NBCFA	비적합	896	0.7670	-8.675	0.000**	895	0.7599	-7.300	0.000**
		적합	47	1.1865			47	1.1023		
	CMA	비적합	896	0.7512	-8.682	0.000**	895	0.7474	-5.649	0.000**
		적합	47	1.1610			47	1.0479		
이탈 제6형	NBCFA	비적합	910	0.7738	-6.416	0.000**	909	0.7641	-8.655	0.000**
		적합	33	1.1764			33	1.1329		
	CMA	비적합	910	0.7577	-6.246	0.000**	909	0.7511	-6.047	0.000**
		적합	33	1.1553			33	1.0730		

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

표본 t검정을 실시하였다. 다음 <표 3>은 dataset 1과 dataset 2에서 χ^2분포적합도 검정에 따라 분류된 집단 간의 독립 2표본 t검정 결과이다.

<표 3>에서 증가형태의 분포와 감소형태의 분포를 가정한 이탈 제 3형과 제 4형의 분석결과는 타 유형의 결과보다 상대적으로 차이가 작게 나타났다. 훈련집합의 선호도 평가치 분포형태와 대칭형의 분포 유형을 가정한 이탈 제 1형도 제 3형과 제 4형의 분석결과보다는 평균의 차가 크게 나타났지만 균등분포의 유형으로 가정한 이탈 제 2형과 "V"자 형태의 분포유형으로 가정한 이탈 제 5형, "W"자 형태의 분포유형으로 가정한 이탈 제 6형의 분석결과보다는 상대적으로 평균의 차가 작게 나타났다. 분석결과 훈련집합의 분포형태에서 이탈할 것으로 가정한 6개의 분포형태에 따라 분류된 고객 집단 간에는 대부분 통계적으로 유의한 차이가 있음을 알 수 있으며 이탈 제 2형과 제 5형, 제 6형에서 분류 집단 간에 MAE의 평균의 차가

크게 나타남을 알 수 있다. 실험결과를 바탕으로 가설 2를 채택할 수 있다.

4.3 가설검정 결과를 이용한 분석

가설 1과 가설 2에서 얻어진 검정결과를 바탕으로 각 가설에서 선별된 고객들 중 공통적으로 선별된 고객들을 분류하였다. 표준편차를 이용한 선별 고객의 수가 가장 많은 235명이므로 이를 기준으로 가설 2에서 선별된 고객의 포함 여부를 확인하였다. 다음 <표 4>는 가설 2의 이탈 제 2형, 제 5형, 제 6형에서 선별된 고객이 가설 1에서 선별된 고객에 포함된 비율이다.

<표 4> 가설 2에서 선별된 고객이 가설 1에서 선별된 고객에 포함된 비율

가설 2	dataset1	dataset2
제2형	68.79%	73.76%
제5형	95.74%	95.74%
제6형	93.94%	100.00%

<표 4>에서 균등분포의 분포유형으로 가정한 이탈 제2형은 표준편차에 의해 선별된 고객에 포함된 비율이 약 70% 전후로 나타났으며 이탈 제 5형과 제 6형의 경우 90% 이상의 비율로 나타나 표준편차에 의한 선별과 매우 밀접한 관계가 있음을 알 수 있다. 가설 1과 가설 2에서 공통적으로 선별된 고객을 선별한 결과 dataset1과 dataset2, 모두에서 동일한 고객 20명이 선정되었다. 다음 <표 5>는 가설 1과 가설 2에서 공통으로 선별된 집단과 그렇지 않은 집단 간 독립2표본 t검정 결과이다.

<표 5>에서 가설 1과 가설 2에서 공통으로 선별된 집단과 그렇지 않은 집단 간의 MAE의 평균은 통계적으로 유의한 차이가 있음을 알 수 있다. 또한 두 집단의 평균 MAE는 가설 1과 가설 2에서 선별된 집단의 평균 MAE 보다 크다는 것을 알 수 있다. 그러나 t값은 가설 2의 결과보다 약간 커짐을 알 수 있다. 이는 공통적으로 선별된 고객들 대부분의 개인 MAE가 크지만 일부 MAE가 상대적으로 작은 사용자도 선별될 가능성이 있음을 보여준다. 그러나 공통적으로 선별된 고객들의 MAE의 평균을 보면 전체에서 매우 큰 MAE를 갖는 고객들이 선정되었음을 알 수 있다.

4.4 선별 고객의 예측 성능 향상을 위한 예측 방법의 제안

가설검정 결과를 통하여 선별된 고객들의 특징은 선호도 평가치의 표준편차가 크면서 훈련집합의 선호도 평가치 분포 유형에서 벗어난 평가치 분포를 갖는 고객들이다. 선호도 평가치를 기

<표 5> 공통 선별집단과 비 선별집단 간 독립 2표본 t검정 결과

dataset	알고리즘	집단구분	빈도	MAE 평균	t값	유의확률
dataset1	NBCFA	비 선별	923	0.7781	-5.271	0.000**
		공통 선별	20	1.2408		
	CMA	비 선별	923	0.7621	-5.024	0.000**
		공통 선별	20	1.2130		
dataset2	NBCFA	비 선별	922	0.7682	-7.545	0.000**
		공통 선별	20	1.1819		
	CMA	비 선별	922	0.7550	-6.207	0.000**
		공통 선별	20	1.1006		

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

반으로 선호도를 예측하는 추천시스템의 경우 고객들에게 제시되는 선호도 평가치의 범위는 절대적 기준으로 제시되지만 고객에 따라서는 이 기준이 다르게 판단될 수 있다.즉 고객에 따라 제시되는 선호도 평가치 범위에 부여되는 선호 정도가 다르기 때문에 이를 고객별로 표준화시켜 표준화된 선호도 평가치를 이용한 분석의 필요성을 제기할 수 있다. 고객별로 선호도 평가치를 표준화시킬 경우 긍정적 선호도 혹은 부정적 선호도로 만으로 평가한 경우 이들의 편차를 늘려주는 효과가 있으며 반대로 편차가 크게 평가한 고객의 경우 편차를 줄여줄 것으로 예상할 수 있다. 연구의 결과를 통하여 선별된 고객들의 선호도 예측 성능 향상을 위하여 선별고객의 선호도 평가치를 표준화시켜 평가치의 편차를 줄이는 방법을 제안한다. 제안 방법의 적용을 위해 훈련집합에서 선별고객이 평가한 선호도 평가치를 고개별로 표준화시켜 표준화된 평가치를 이용하여 실험집합의 영화에 대한 선호도를 예측하고 예측 결과를 다시 표준화 이전의 형태로 환원하여 예측 결과의 오차를 분석하였다. 다음 <표 6>은 가설 1에 의해 선별된 고객들의 표준화 이전의 고객별 MAE와 표준화 이후의 고객별 MAE의 대응평균 검정 결과이다.

<표 6>에서 실험 dataset1, 2 모두에서 4집단으로 선별된 고객들의 MAE가 선호도 평가치의 표준화 이후 선호도를 예측한 결과가 표준화 이전의 예측 결과보다 개선되었음을 알 수 있으며 CMA에 비하여 NBCFA의 결과가 상대적으로 개선 정도가 큼을 알 수 있다. 실험 dataset에 따라서는 dataset1의 결과가 표준화에 따른 개선 정도가 크게 나타났으며 dataset2에서의 개선 정도가 상대적으로 낮음을 알 수 있다. 결과에서 실험 dataset에 따라 표준화에 따른 영향의 차이가 발생할 수 있지만 개선 경향에 있어서는 유사한 결과를 보이고 있다. <표 7>은 가설 2에 의해 선별된 실험 dataset1의 고객들의 표준화 이전의 고객별 MAE와 표준화 이후의 고객별 MAE의 대응평균 검정 결과이다.

<표 7>에서 가설 2의 분포 유형에 따른 분류 기준에서 선별의 효과가 상대적으로 떨어졌던 유형 1, 3, 4를 제외한 유형 2, 5, 6의 경우에서 표준화를 통한 개선효과가 나타남을 알 수 있다. 결과에서 선호도 평가치의 표준화를 통하여 통계적으로 유의한 개선효과를 얻은 분포의 유형은 각각 균등분포의 유형, "V"형태의 분포 유형, "W" 형태의 유형임을 알 수 있다. 실험 dataset2에서도 유사한 결과를 얻을 수 있으나 실험 dataset1의 결과에비하여 큰 효과를 얻지는 못하였다. 이는 실험 dataset의 구성에 따라 표준화의 효과가 증감할 수 있지만 그 경향에 있어서 유사한 결과를 보이고 있음을 알 수 있다.

<표 6> 가설 1에 의해 선별된 고객의 표준화 예측 결과의 대응평균 검정 결과

dataset	알고리즘	구분	평균	빈도	평균차	t값	유의확률
dataset1	NBCFA	비 표준화	0.9787	235	0.0148	5.7836	0.0000**
		표준화	0.964				
	CMA	비 표준화	0.9528		0.0117	4.4192	0.0000**
		표준화	0.9412				
dataset2	NBCFA	비 표준화	0.9653	235	0.0102	4.9709	0.0000**
		표준화	0.9551				
	CMA	비 표준화	0.9311		0.0072	3.6155	0.0004**
		표준화	0.9239				

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

<표 7> 가설 1에 의해 선별된 실험 dataset1의 표준화 예측 결과의 대응평균 검정 결과

분포유형	알고리즘	구분	MAE 평균	빈도	평균차	t값	유의확률
1형	NBCFA	비 표준화	0.9738	35	-0.0024	-0.5678	0.5739
		표준화	0.9762				
	CMA	비 표준화	0.9472		-0.0006	-0.1463	0.8845
		표준화	0.9479				
2형	NBCFA	비 표준화	1.0461	141	0.0128	3.6399	0.0004**
		표준화	1.0333				
	CMA	비 표준화	1.0203		0.0122	3.1821	0.0018**
		표준화	1.0081				
3형	NBCFA	비 표준화	0.9336	12	0.0134	1.5840	0.1415
		표준화	0.9202				
	CMA	비 표준화	0.921		0.0130	1.2972	0.2211
		표준화	0.9079				
4형	NBCFA	비 표준화	0.8411	243	0.0046	2.6709	0.0081**
		표준화	0.8365				
	CMA	비 표준화	0.8214		0.0016	0.8718	0.3842
		표준화	0.8197				
5형	NBCFA	비 표준화	1.1865	47	0.0381	3.9916	0.0002**
		표준화	1.1484				
	CMA	비 표준화	1.161		0.0294	2.8431	0.0066**
		표준화	1.1316				
6형	NBCFA	비 표준화	1.1764	33	0.0391	3.1670	0.0034**
		표준화	1.1373				
	CMA	비 표준화	1.1553		0.0327	2.4578	0.0196*
		표준화	1.1226				

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

다음 <표 8>은 <표 5>에서 공통으로 선별된 고객들의 표준화에 의한 예측 성능향상의 결과에 대한 대응평균 검정 결과이다.

<표 8>에서 선호도 평가치의 표준화를 통한 선호도 예측 결과가 표준화 이전의 선호도 예측 결과에 비해 향상됨을 알 수 있지만 실험 dataset의 구성에 따라 그 정도가 달라짐을 알 수 있다. 공통으로 선별된 고객들의 표준화에 따른 영향은 실험 dataset1에서는 그 개선 정도가 통계적으로 유의한 개선이 있는 것으로 분석되었고 dataset2에서는 개선의 정도가 통계적으로는 유의하지 않지만 향상되어 있음을 결과에서 알 수 있다.

가설 1과 가설 2에 의해 선별된 고객은 공통적으로 선호도 평가치의 표준편차가 큰 특징을 가지고 있으며 표준편차의 영향을 줄여 선호도 예측 결과의 향상을 위해 선별 고객의 선호도 평가치를 표준화시켜 선호도를 예측 할 경우 향상된 결과를 얻을 수 있음을 통계적 분석을 통하여 알 수 있다. 본 연구에서는 선호도 예측의 정확도 향상을 위하여 선호도 평가치의 표준화에 의한 예측 방법을 제안하였지만 최종적으로 공통 선별된 20명의 고객에게는 표준화의 효과가 크지 않

<표 8>가설 1과 가설 2에서 공통으로 선별된 고객들의 표준화 영향에 대한 대응평균 검정 결과

dataset	알고리즘	구분	MAE 평균	빈도	평균차	t값	유의확률
dataset1	NBCFA	비 표준화	1.2408	20	0.0481	2.8146	0.0111*
		표준화	1.1927				
	CMA	비 표준화	1.213		0.0456	2.2238	0.0385*
		표준화	1.1674				
dataset2	NBCFA	비 표준화	1.1819	20	0.0184	2.0365	0.0559
		표준화	1.1635				
	CMA	비 표준화	1.1006		0.0072	0.5851	0.5654
		표준화	1.0934				

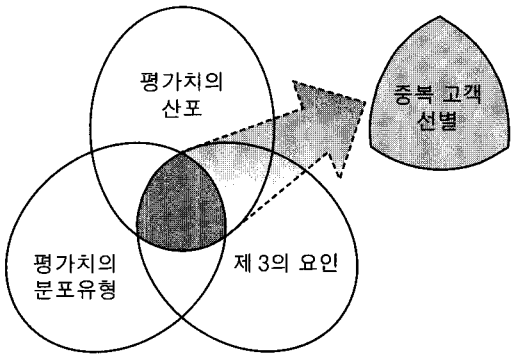
* : $p < 0.05$, ** : $p < 0.01$.

음을 통하여 선호도 예측 정확도 향상을 위하여 다른 접근법의 시도가 필요함을 알 수 있다.

4.5 실험 결과의 요약

가설검정을 통한 실험결과 협력적 필터링 기법에서 선호도 예측 이전에 주어진 사전정보를 이용하여 고객의 선호도 예측 결과를 사전에 평가할 수 있는 가능성이 있음을 알 수 있다. 가설 1의 검정 결과 개인의 선호도 평가 패턴에서 선호도 평가치의 극값이 크다는 것은 이웃 고객의 선호도 평가치를 바탕으로 선호도 예측이 이루어지는 협력적 필터링 기법에서 그 오차가 커질 수 있으며 이러한 유형의 고객들은 시스템에서 정확한 선호도 예측 결과를 얻기 어려울 수 있다는 점을 시사한다. 가설 2의 검정에서 시스템 전체의 선호도 평가치의 분포와 다른 형태의 선호도 분포를 보이는 고객들은 시스템 내의 대다수 고객들의 성향과 차이가 있어 정확한 예측이 어렵다는 점을 시사하고 있다. 또한 가설 2의 이탈 제 2형, 제 5형, 제 6형의 분포는 일반적으로 분포의 표준편차가 커질 수 있는 분포 유형으로 가설 1에 의해 선별된 고객들과 중복될 가능성이 크다는 것을 알 수 있었다. <표 4>의 결과에서 알 수 있듯이 가설 2에서 MAE가 큰 분류집단의 고객들은 가설 1에서 분류된 고객들과 많은 부분 중복되어 있음을 알 수 있다. 가설 2의 이탈 제 2형을 제외한 이탈 제 5형과 이탈 제 6형에 의해 분류된 고객들은 대부분이 가설 1에서 분류된 고객들과 중복됨을 알 수 있다. 또한 가설 1과 가설 2의 검정결과를 확장하여 각 가설검정에서 분류된 고객들 중 중복된 고객들만을 선별한 집단의 MAE가 그렇지 않은 집단의 MAE보다 크다는 것을 알 수 있으며 선호도 평가치의 표준편차의 영향을 줄일 수 있으면 개선된 선호도 예측 결과를 얻을 수 있음을 제안 방법의 선호도 예측 결과를 통하여 알 수 있었다. 실험결과를 토대로 선호도 예측 성능이 낮을 것으로 예상되는 고객의 사전 평가에 대한 방법으로 다음 <그림 5>와 같은 개략적인 선별기준을 제시할 수 있다.

<그림 5>에서 제시된 평가치의 산포와 평가치의 분포유형을 통한 선별기준은 본 연구에서 제시



<그림 5> 사전정보에 의한 예측 성과가 낮은 고객의 선별기준

되고 있으며 기타 제3의 요인으로 기존 연구에서 제시되었던 고객별 선호도 평가 패턴에서 시간에 따라 발생하는 Run의 길이가 선호도 예측 정확도에 미치는 영향에 관한 연구와 선호도 평가치의 특정 선호도 평가치의 발생 빈도에 따라 선호도 예측 성과가 낮은 고객들을 선정할 수 있는 기준 등을 제시할 수 있다[Lee *et al.*, 2007a; Lee *et al.*, 2007b].

V. 결론 및 시사점

본 연구는 협업필터링에서 선호도 예측 이전에 고객이 평가한 선호도 평가치의 특성을 이용하여 선호도 예측 오차가 클 것으로 예상되는 고객을 선별할 수 있는 방법을 가설검정을 통하여 제시하였다. 선호도 예측 이전의 사전정보인 고객의 선호도 평가치를 이용한 선호도 예측 오차의 사전평가 방법은 선호도 예측 이전에 예측 오차가 큰 고객들을 분류할 수 있기 때문에 이들의 특성을 파악하는데 중요한 자료를 제공할 수 있다. 가설검정의 결과에 따르면 표준편차를 이용한 방법은 집단 구분의 수에 따라 943명의 고객들 중 예측 오차가 클 것으로 예상되는 고객을 달리 정할 수 있으며 선호도 평가치의 분포를 이용한 방법은 MovieLens 100K dataset을 이용한 본 논문에서 이탈 제 2형은 141명, 제 5형은 47명, 제 6형은 33명을 선별 할 수 있었다. 또한 선별방법에서 공통으로 선별된 고객들은 dataset1과 dataset2에서 모두 20명의 선별되었으며 이들의 선호도 예측 오차는 선별되지 않은 고객들과 비교하여 오차가 큼을 통계적으로 확인하였다. 본 연구를 통하여 다음과 같은 차기 연구의 주제를 제시할 수 있다.

첫째, 선별된 선호도 예측 정확도가 낮은 고객들이 선호도 예측 시스템의 예측 성능 향상에 장애가 되는가? 본 연구의 결과를 통하여 선별된 고객의 예측 성과가 낮은 것을 알 수 있었으나 이들이 선호도 예측 시스템의 성능에 악영향을 주는지 혹은 시스템의 다른 고객들의 선호도에 영향을 받아 성과가 나쁘게 나왔는지에 대하여는 분석이 이루어지지 않았다. 차기 연구에서는 이들이 예측 시스템에 영향을 주는지 혹은 영향을 받는지에 대한 연구가 필요하다.

둘째, 분류된 고객들의 선호도 평가 특성을 파악할 수 있는가? 선호도 예측 시스템의 성능을 저하시키는 시스템 노이즈는 크게 자연적 노이즈와 인위적 노이즈로 구분하여 정의하는데 선별된 고객들을 이 두 가지 유형의 노이즈로 구분할 수 있는지에 대한 연구가 필요하다. 이는 인위적 노이즈로 분류될 경우 악의적 의도를 지닌 고객들인 추천시스템에 대한 공격자로 분류할 수 있기 때문에 추천시스템에서 선호도 예측 과정에서 이들의 평가치에 대한 적절한 조치를 사전에 취할 수 있으며 자연적 노이즈일 경우 선호도 평가치 수집과정이나 고객의 시스템에 대한 잘못된 지식 혹은 습관 등에 대하여 인지시킬 수 있다.

셋째, 분류된 고객들의 선호도가 시스템내의 다른 고객들과 전혀 다른 성향을 가지고 있을 경우 이들의 예측 오차를 줄일 수 있는가? 분류된 고객들이 시스템의 공통적 성향에 많이 벗어나 선호도 예측의 성과가 낮더라도 이들을 위한 선호도 예측 개선을 위하여 새로운 알고리즘을 개발하여 특이 성향의 고객들을 위한 선호도 예측 방법에 대한 연구가 필요하다.

마지막으로, 메모리 기반 협업필터링 알고리즘을 이용한 추천시스템은 추천대상 상품에 대한 목표고객의 선호도를 예측하기 위해 선호도 예측에 필요한 이웃 고객들을 선정한다. 이웃 선정과정을 통하여 선정된 고객들의 선호정보를 이용하기 때문에 웹상에서의 실시간 추천에서 이웃 선정과정과 고객 간 유사도 가중치 계산에 소요되는 시간이 매우 커질 가능성이 크다. 그렇기 때문에 사전에 이웃 고객의 선정과 유사도 가중치에 계산에 대한 과정이 이루어져야 실시간 추천이 가능하다. 차기 연구로 실시간 추천을 위한 일괄처리 방법을 통하여 예측 시간의 단축에 관한 연구가 필요하다.

〈참 고 문 헌〉

[1] 김경재, 김병국, "데이터 마이닝을 이용한 인터넷 쇼핑몰 상품추천시스템," 한국지능정보시스템학회논문지, 제11권, 제1호, 2005, pp. 191-205.

[2] 김용수, "비정형화된 속성의 학습을 통한 자동화된 내용 기반 필터링 기법의 개발," *Journal of the Korean Data Analysis Society*, Vol. 8, No. 4, 2006, pp. 1615-1624.

[3] 김종우, 배세진, 이홍주, "협업 필터링 기반 개인화 추천에서의 평가자료의 희소 정도의 영향," *경영정보학연구*, Vol. 14, No. 2, 2004, pp. 131-149.

[4] 김재경, 안도현, 조윤호, "Development of a Personalized Recommendation Procedure Based on Data Mining Techniques for Internet Shopping Malls," *한국지능정보시스템학회논문지*, 제9권, 제3호, 2003, pp. 177-191.

[5] 손재봉, 서용무, "협업 필터링 시스템에서 Degree of Match를 이용한 성능향상," *Information Systems Review*, 제8권, 제3호, 2006, pp. 139-154.

[6] 심장섭, "K-means 군집화와 순차 패턴 기법을 사용하는 VLDB 기반의 추천 시스템 설계," 충북대학교, 박사학위논문, 2005.

[7] 이석준, 이희춘, "협업 필터링 추천에서 대응평균 알고리즘의 예측 성능에 관한 연구," *Information Systems Review*, 제9권, 제1호, 2007, pp. 85-103.

[8] 이석준, 김선옥, 이희춘, "추천시스템에서 Run 특이자가 예측 정확도에 미치는 영향에 관한 연구," *한국인터넷정보학회 추계학술발표대회*, 2007a, pp. 299-302.

[9] 이석준, 김선옥, 이희춘, "협력적 필터링에서 평가치의 Run 특이자와 예측 정확도의 관계에 관한 연구," *Journal of the Korean Data Society*, Vol. 9, No. 4, 2007b, pp. 2043-2054.

[10] 이희춘, 이석준, "대응평균 알고리즘을 이용한 협력적 필터링 추천시스템의 성능향상," *한국경영정보학회 2006 추계컨퍼런스*, 2006, pp. 208-214.

[11] 한국인터넷진흥원, "한국인터넷백서 2007," 한국인터넷진흥원, 2007.

[12] Adomavicius, G., A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Transactions on Knowledge and DATA Engineering*, Vol. 17, No. 6, 2005, pp. 734-749.

[13] Balabanovic, M., Y. Shoham, "Fab: contentbased, collaborative recommendation," *Communications of the ACM*, Vol. 40, Issue 3, 1997, pp. 66-72.

[14] Breese, J.S., D. Heckerman, C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, 1998, pp. 43-52, Madison, Wisconsin.

[15] Burke, R., B. Mobasher, R. Bhaumik, C. Williams, "Segment-Based Injection Attacks against Collaborative Filtering Recommender Systems," *In Proceedings of the Fifth IEEE International Conference on Data Mining*, 2005, pp. 577-580.

[16] Claypool, M., A. Gokhale, T. Miranda, P. Murnikov, D. Netes and M. Sartin, "Combining content-based and collaborative filters in an online newspaper," *In Proceedings of ACM SIGIR Workshop on Recommender Systems: Algori-*

thms and Evaluation, University of California, Berkeley, Aug. 1999.

[17] Deshpande, M., G. Karypis, "Item-based top-N recommendation algorithms," *ACM Transactions on Information Systems*, Vol. 22, No. 1, 2004, pp. 143-177.

[18] Herlocker, J., J. Konstan, J. Riedl, "An Empirical Analysis of Design Choices in Neighborhood Based Collaborative Filtering Algorithms," *Information Retrieval*, Vol. 5, No. 4, 2002, pp. 287-310.

[19] Hill, W.L., S.M. Rosenstein, G. Furnas, "Recommending and Evaluating Choices in A Virtual Community of use," *In Proceedings of the SIGCHI conference on Human factors in computing systems*, 1995, pp. 194-201.

[20] Huseyin P. and D. Wenliang, "SVD-based Collaborative Filtering with Privacy," *In Proceedings of the 2005 ACM symposium on Applied computing*, 2005, pp. 791-795.

[21] Kim, T.H. and S. B. Yang, "An Improved Neighbor Selection Algorithm in Collaborative Filtering," *IEICE TRANS. INF. & SYST.*, Vol. E88-D, No. 5, 2005, pp. 1072-1076.

[22] Lee, S.J., S.O. Kim, H.C. Lee, "Pre-Evalua tion for Detecting Abnormal Users in Recommender System," *Journal of the Korean Data & Information Science Society*, Vol. 18, No. 3, 2007a, pp. 619-628.

[23] Lee, S.J., S.O. Kim, H.C. Lee, " 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, 2007b, pp. 659-668.

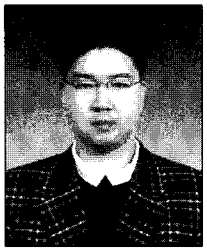
[24] O'Mahony, M.P., N.J. Hurley, G.C. M. Silvestre, "Detecting noise in recommender system databases," *In Proceedings of the 11th international conference on Intelligent user interfaces*, 2006, pp. 109-115.

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

[26] Sarwar, B.M., J. Konstan, A. Borchers, J. Herlocker, B. Miller, J. Riedl, "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System," *In Proceedings of the 1998 Conference on Computer Supported Cooperative Work*, Nov. 1998.

[27] Shardanand, U. and P. Maes, "Social Information Filtering: Algorithms for Automating 'Word of Mouth'," *In Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, 1995, pp. 210-217.

◆ 저자소개 ◆



이석준 (Lee, Seok Jun)
상지대학교 산업공학과 및 산업환경대학원에서 석사를 마치고 일반대학원 경영학 박사학위를 취득하였으며 상지대학교 생산기술연구소 연구원을 역임하고 현재 상지대학교 경영학과 겸임교수로 재직 중이다. 관심분야는 전자상거래, 추천시스템, 데이터 마이닝 등이다.

김선옥 (Kim, Sun Ok)
서강대학교 대학원 수학과에서 박사학위를 취득하였으며 현재 한라대학교 정보통신공학부 교수로 재직 중이다. 관심분야는 멀티미디어 시스템, 정보보안, 개인화 서비스 등이다.

◈ 이 논문은 2007년 08월 13일 접수하여 1차 수정을 거쳐 2007년 11월 14일 게재확정되었습니다.