협력적 필터링 추천 시스템의 정확도 향상

이 석 환* · 박 승 헌*
*인하대학교 산업공학과

Accuracy improvement of a collaborative filtering recommender system

Seog-Hwan Lee* · Seung-Hun Park*
*Inha University Industrial Engineering

Abstract

In this paper, the author proposed following two methods to improve the accuracy of the recommender system. First, in order to classify the users more accurately, the author used a EMC(Expanded Moving Center) heuristic algorithm which improved clustering accuracy. Second, the author proposed the Neighborhood-oriented preference prediction method that improved the conventional preference prediction methods, so the accuracy of the recommender system is improved. The test result of the recommender system which adapted the above two methods suggested in this paper was improved the accuracy than the conventional recommendation methods.
Keywords : Collaborative Filtering, Recommender System, EMC Heuristic

1. 서 론

추천 시스템은 상품을 추천받을 고객인 목표고객의 원하는 상품을 정확하게 예측할 수 있어야 한다. 추천 시스템이 목표고객의 원하는 상품을 정확하게 예측하지 못하면 온라인 매장에 대한 목표고객의 만족도는 낮아지게 된다. 따라서 온라인 매장은 목표고객의 만족도를 높이기 위해 추천의 정확도가 높은 추천 시스템을 개발하여 사용해야만 한다.

본 연구의 목적은 협력적 필터링 추천 시스템의 추천 정확도를 향상시키기는 것이다. 그 추천의 정확도를 향상시키기 위해 다음과 같이 두 가지 사항을 개선하고자 한다.

첫째, 협력적 필터링은 목표고객에게 상품을 추천하기 위해 이웃사용자의 선호도 정보를 이용한다. 따라서 추천의 정확도를 향상시키기 위해서는 목표고객에 대한 이웃사용자를 정확하게 선정해야한다. 목표고객에 대한 이웃사용자 선정은 기 사용자들을 유사도에 따라서 미리 몇 개의 군집으로 분류하여 목표고객과 유사도가 가장 높은 군집을 찾음으로써 선정할 수 있다. 이 방법은 군집의 정확도만을 높이면 이웃사용자 선정의 정확도를 높일 수 있기 때문에 결과적으로 추천의 정확도가 향상된다. 따라서 본 연구는 추천의 정확도를 향상시키기 위해 기존 알고리즘보다 군집의 정확도가 높은 새로운 알고리즘으로 기 사용자를 군집한다.

둘째, 추천의 정확도를 향상시키기 위해서는 상품에 대한 목표고객의 선호도를 정확하게 예측해야한다. 선호도 예측은 목표고객의 평균 선호도와 이웃사용자가 평가한 상품의 선호도를 이용한다. 기존 방법에서 평균 선호도는 목표고객의 선호도를 이용하여 계산하기 때문에 목표고객이 입력한 선호도 정보가 적을 경우 평균 선호도의 정확도가 낮아질 수밖에 없다. 따라서 본 연구는 평균 선호도 계산방법을 개선하여 목표고객의 선호도 정보가 적을 경우에도 선호도 예측의 정확도를 높일 수 있는 방법을 제시한다.

본 연구에서는 앞에서 설명한 두 가지 사항을 개선하여 추천의 정확도를 향상시킬 수 있는 방법을 다음과 같이 제안한다. 첫째, 목표고객에 대한 이웃사용자를 정확하게 선정하기 위해 군집의 정확도가 높은 알고리즘을

† 이 논문은 인하대학교 교내 연구비 지원에 의해 연구되었음.
† 교신저자: 이석환, 인하대학교 산업공학과 생산관리 연구실
Tel: 02-2610-0419, E-mail: seoghwan@inha.ac.kr
2010년 1월 18일 접수; 2010년 3월 5일 수정본 접수; 2010년 3월 12일 게재확정

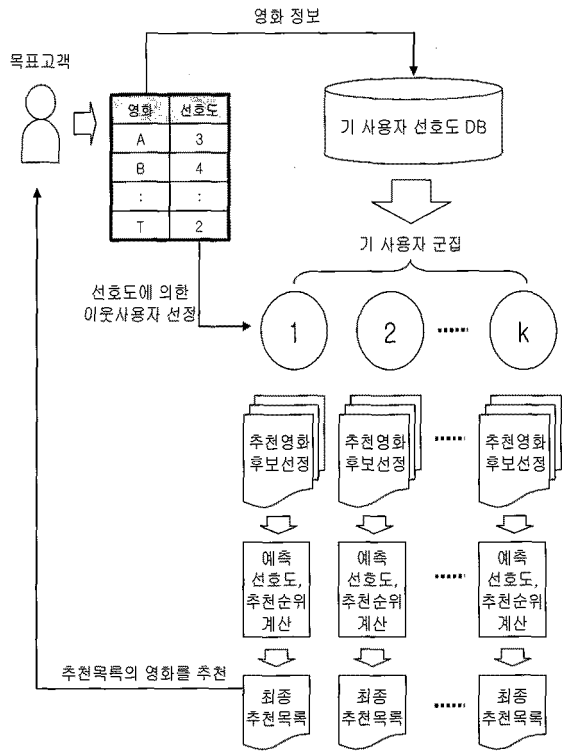
사용하여 기 사용자들을 군집한다. 본 연구에서 기 사용자 군집에 사용한 군집 알고리즘은 EMC 휴리스틱 군집 알고리즘이다[1]. EMC 휴리스틱 군집 알고리즘은 기존의 군집 알고리즘에 비해서 군집의 정확도가 높기 때문에 목표고객에 대한 이웃사용자를 보다 정확하게 선정할 수 있다. 둘째, 목표고객에 비해서 선호도 정보량이 항상 많은 이웃사용자의 선호도를 이용하여 평균 선호도를 계산함으로써 선호도 예측의 정확도를 높인다.

다음으로 본 연구에서 제안한 두 가지 방법을 실제로 고객의 선호도를 조사한 MovieLens 데이터[9]에 적용하고 그 결과를 기존의 기 사용자 군집방법 및 선호도 예측방법의 결과와 비교한다.

2. 추천 시스템

협력적 필터링 추천 시스템의 성능 향상을 위해 본 연구에서 제시한 추천 시스템은 다음과 같은 순서로 설명한다. 2.1에서는 추천 시스템의 추천과정을 설명하고 2.2에서는 추천 시스템의 정확도를 향상시키기 위해 본 연구에서 제시한 이웃사용자 선정방법과 2.4에서는 이웃사용자 중심 선호도 예측방법에 대해 설명한다.

2.1 추천과정



<그림 2-1> 추천 시스템 흐름

협력적 필터링 기법을 적용한 추천 시스템의 성능을 향상시키기 위해서는 서론에서 언급한 것과 같이 목표고객에 대한 이웃사용자를 정확하게 선정하며 상품에 대한 목표고객의 선호도를 정확하게 예측해야한다. 본 연구에서는 이와 같은 두 가지 사항을 개선함으로써 기존의 협력적 필터링 기법에 비해서 추천의 정확도를 향상시킬 수 있는 방법을 제시하여 시스템으로 구축하였다. <그림 2-1>은 구축한 추천 시스템의 흐름을 그림으로 나타낸 것이고 <그림 2-2>는 그 추천과정을 단계별로 나타낸 것이다. 추천과정은 크게 사용자 군집, 추천영화 후보선정, 선호도 예측, 추천목록 생성으로 나누어진다. 이후 설명은 <그림 2-2>의 순서로 설명한다.

2.2 사용자 군집방법과 이웃사용자 선정방법

본 연구에서는 추천의 초기에 EMC 휴리스틱 군집 알고리즘으로 기 사용자를 군집하였다. EMC 휴리스틱 군집 알고리즘은 학술지에 게재되어 타 군집 알고리즘에 비해서 군집의 정확도가 높다는 것을 확인하였다[1].

기존 연구에서 이웃사용자 선정방법 중 하나는 목표고객과 모든 기 사용자간의 유사도를 계산하고 그 중에서 유사도가 높은 일부 사용자를 이웃사용자로 선정한다. 그러나 이 방법은 이웃사용자로 선정할 사람의 수가 적은 경우, 선호도 예측의 정확도가 낮아지는 문제점이 있다. 기 사용자가 군집되면 목표고객과 각 군집의 유사도를 비교하여 유사도가 가장 높은 군집을 목표고객의 이웃사용자로 선정한다.

한편으로 기 사용자의 수가 크게 증가할 경우 이웃

Step 1. 목표고객과 기 사용자에 대해서 x개 영화에 대한 선호도를 조사한다.
Step 2. Step 1에서 조사한 선호도를 기준으로 기 사용자를 군집한다. 군집에 사용하는 알고리즘은 EMC 휴리스틱 군집 알고리즘이다.
Step 3. x개 영화에 대한 목표고객의 선호도와 군집의 중심과의 거리를 계산하여 거리가 가장 가까운 군집을 목표고객의 이웃사용자로 선정한다.
Step 4. 군집별 추천영화 후보를 선정한다.
Step 5. Step 4에서 선정된 영화에 대해서 목표고객의 선호도를 예측한다.
Step 6. 추천순위를 계산하고 추천순위가 높은 순서로 N개의 추천영화 목록을 생성한다.

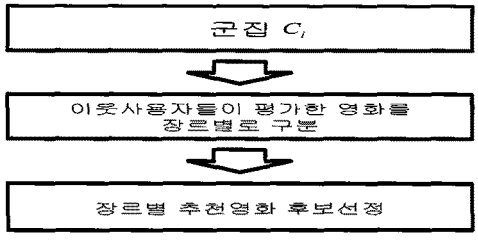
<그림 2-2> 추천과정

사용자 선정에 군집을 사용하는 방법은 목표고객과 모든 기 사용자에 대해 유사도를 계산하는 방법에 비해서 이웃사용자 선정에 소요되는 시간을 크게 단축시킬 수 있다[2]. 기 사용자 군집 및 이웃사용자 선정은 <그림 2-2>에서 Step 1~3에 해당하는 것으로 Step 1과 Step 2는 기 사용자를 군집하는 단계이고 Step 3은 목표고객의 이웃사용자를 선정하는 단계이다.

2.3 추천영화 후보선정

추천영화 후보선정은 목표고객에게 영화를 추천하기 전에 목표고객이 선호할 가능성이 높은 영화를 선정하는 단계이다. 추천영화의 후보 수는 추천할 영화 편수의 몇 배의 영화를 선정한다. 일반적으로 영화의 추천은 목표고객이 선호하는 장르의 영화를 추천하기 때문에 추천영화의 후보도 목표고객이 선호하는 장르에서 선정한다. 그러나 선호 장르를 이용한 추천은 추천영화의 장르가 줄어든다는 문제점이 있다. 즉, 목표고객이 특정 장르의 영화에 관심을 보일 경우 그 목표고객은 다른 장르의 영화를 추천받을 수 없다. 본 연구에서는 이 문제를 해결하기 위해 목표고객에게 모든 장르의 영화를 다양하게 추천하는 방법을 사용하였다. 본 연구에서는 추천과정 초기에 기 사용자를 몇 개의 군집으로 분류하고 그 군집 중에서 목표고객과 유사도가 가장 높은 군집을 목표고객의 이웃사용자로 선정하였다. 따라서 추천영화의 후보선정은 목표고객의 이웃사용자가 평가한 영화를 대상으로 이루어진다. 먼저 이웃사용자가 평가한 영화에 대해서 장르를 구분한 후 각 장르별로 몇 편의 영화를 추천영화의 후보로 선정한다.

이때 추천영화 후보는 이웃사용자들의 선호도 합이 높은 영화를 우선적으로 선정한다. 그 이유는 일반적인 추천 시스템에서 상품의 추천은 상품을 구매한 기 사용자의 선호도를 조사하여 그 중에서 선호도가 높은 상품을 우선적으로 추천하기 때문이다. 따라서 추천영화 후보는 이웃사용자들의 선호도 합이 높은 영화를 우선적으로 선정하였다. <그림 2-3>은 특정 군집에 대해서 추천영화 후보를 선정하는 과정으로 <그림 2-2>의 Step 4에 해당한다.



<그림 2-3> 추천영화 후보 선정과정

2.4 이웃사용자 중심 선호도 예측방법

추천의 정확도는 추천할 영화에 대한 목표고객의 선호도를 정확하게 예측할수록 높아진다. 선호도 예측은 <그림 2-2>에서 Step 5로 Step 4에서 선정한 추천영화 후보에 대해서 목표고객에 대한 선호도를 예측하는 단계이다. 일반적으로 영화 추천 시스템은 목표고객이 특정 영화에 대해서 갖는 선호도를 예측하고 그 중에서 예측 선호도가 가장 높은 영화를 추천한다. 이때 선호도 예측은 식 (1)을 사용한다.

$$R_{A,m} = \overline{R_A} + \frac{\sum_{j=1}^{c} w(A,j)(R_{j,m} - \overline{R_j})}{\sum_{j=1}^{c} |w(A,j)|} \qquad (1)$$

A: 목표고객
j: 이웃사용자
$\overline{R_A}$: 목표고객 A의 평균선호도
$\overline{R_j}$: 이웃사용자 j의 평균선호도
c: 이웃사용자의 수
m: 영화
$w(A,j)$: 목표고객 A와 이웃사용자 j의 유사도 가중치
$R_{j,m}$: m영화에 대한 이웃사용자 j의 선호도

식 (1)에서 평균선호도 $\overline{R_A}$는 군집의 기준으로 선정한 영화에 대해서 목표고객 A가 입력한 선호도를 평균한 것이다. 선호도의 예측은 목표고객의 평균 선호도와 이웃사용자의 선호도 가중치를 합하여 계산한다. 평균 선호도는 일부 영화에 대해서 목표고객이 입력한 선호도를 평균한 것이기 때문에 목표고객의 실제 평균 선호도와는 차이가 있을 수 있다. 평균 선호도가 목표고객의 실제 평균 선호도와 차이가 있을 경우 선호도 예측의 정확도는 낮아질 수 있고 동시에 추천의 정확도도 낮아질 수 있다. 본 연구에서는 평균 선호도가 목표고객의 실제 평균 선호도에 근접할 수 있도록 목표고객의 평균 선호도 대신에 이웃사용자의 평균 선호도를 사용하였다. 그 방법은 다음과 같다. 이웃사용자의 선호도는 목표고객과 유사할 뿐 아니라 목표고객이 평가하지 않은 영화의 선호도 정보도 가지고 있다. 따라서 선호도 예측 식에서 목표고객의 평균 선호도 대신에 이웃사용자들의 평균 선호도를 사용하면 더 많은 선호도를 선호도 예측에 반영할 수 있다. 한편 목표고객은 영화의 속성에 따라서 서로 다른 평균 선호도를 가질 수 있다. 예를 들면, 어떤 목표고객은 특정 장르의 영화 또는 특정 감독 및 주연배우 등에 대해서 높거나 낮은 평균 선호도를 가질 수 있다. 이 경우 목표

고객의 평균 선호도는 영화마다 차이가 있게 된다. 따라서 본 연구에서는 목표고객에 대해서 동일한 평균 선호도를 적용하지 않고 선호도를 예측하려는 영화별로 서로 다른 평균 선호도가 적용되도록 하였다. 이 방법은 선호도를 예측하려는 영화에 대해서 이웃사용자의 평균 선호도를 계산하고 이것을 해당 영화의 평균 선호도로 정하는 방법으로 이웃사용자 중심 선호도 예측방법이라고 부른다. 기존의 선호도 예측방법은 한 명의 목표고객에 대해서 동일한 평균 선호도가 적용되었지만 본 연구의 이웃사용자 중심 선호도 예측방법은 동일한 목표고객이라도 영화에 따라서 서로 다른 평균 선호도가 적용된다. 식 (2)는 이웃사용자 중심 선호도 예측방법으로 선호도를 예측하는 식이다. 식 (2)에서 평균 선호도는 앞에서 설명한 것과 같이 이웃사용자가 m영화에 대해서 평가한 선호도의 평균이다.

$$R_{A,m} = \frac{\sum_{j=1}^{c} R_{j,m}}{c} + \frac{\sum_{j=1}^{c} w(A,j)(R_{j,m} - \overline{R_j})}{\sum_{j=1}^{c} |w(A,j)|} \quad (2)$$

A : 목표고객
j : 이웃사용자
$\overline{R_j}$: 이웃사용자 j의 평균 선호도
c : 이웃사용자의 수
m : 영화
$w(A,j)$: 목표고객 A와 이웃사용자 j의 유사도 가중치
$R_{j,m}$: m영화에 대한 이웃사용자 j의 선호도

평균 선호도를 계산할 때 목표고객 A의 모든 이웃사용자를 사용할 수도 있으나 상관관계가 너무 낮은 이웃사용자의 선호도를 평균 선호도에 반영할 경우 평균 선호도의 정확도가 낮아질 수 있다. 본 연구에서는 평균 선호도의 정확도를 높이기 위해 피어슨 상관계수가 0.5 이상인 이웃사용자의 선호도만을 사용하여 평균 선호도를 계산하였다.

식 (2)에서 $w(A,j)$는 피어슨 상관계수이다. 피어슨 상관계수는 상관계수가 높은 이웃사용자가 평가한 영화에는 높은 선호도 가중치를 부여하고 상관계수가 낮은 이웃사용자가 평가한 영화에는 낮은 선호도 가중치를 부여하기 위해 사용한다. 이때 모든 이웃사용자의 선호도 가중치를 적용할 수도 있지만 상관관계가 너무 낮은 이웃사용자들의 선호도 정보를 반영할 경우 선호도 예측의 정확도가 낮아질 수 있다. 본 연구에서는 선호도 예측의 정확도를 높이기 위해 0.5 이상의 상관관계를 갖는 이웃사용자에 대해서만 선호도 가중치를 적용하였다.

한편 군집의 정확도가 높더라도 이상치가 존재하기 때문에 부의 상관을 갖는 이웃사용자가 존재할 수 있다. 목표고객과 부의 상관을 갖는 이웃사용자는 목표고객의 이웃사용자라고 할 수 없기 때문에 부의 상관을 갖는 이웃사용자는 선호도 예측에서 제외하였다. 종합하면 이웃사용자 중심 선호도 예측방법에서 평균 선호도 계산과 선호도 가중치는 정의 상관을 가지고 상관계수가 0.5이상인 이웃사용자의 선호도만을 사용한다.

3. 추천 시스템 성능평가

본 연구에서 제시한 두 가지 추천방법(2.2, 2.4 참조)을 추천 시스템에 적용함으로써 기존의 추천 시스템보다 추천의 정확도가 향상된다는 것을 실험을 통하여 확인한다. 먼저 3.1에서는 추천실험의 과정을 설명한다.

3.2.1에서는 기 사용자 군집에 k-means, 타부 탐색 알고리즘과 EMC 휴리스틱 군집 알고리즘을 사용하는 경우 추천의 정확도를 비교한다. 3.2.2에서는 선호도 예측에서 기존의 선호도 예측방법과 본 연구에서 제안한 이웃사용자 중심 선호도 예측방법을 사용했을 경우 선호도 예측의 정확도를 비교한다.

3.1 추천실험 과정

본 연구에서 제안한 추천 시스템의 성능평가를 위해 <그림 3-1>와 같은 절차로 추천실험을 진행하였다. 추천실험은 실험의 정확도를 높이기 위해 3회 반복하여 진행하였다.

그 반복 과정은 <그림 3-1>의 Step 2부터 Step 7이다. 즉, 3회의 각 실험마다 영화 10편을 랜덤으로 선정하고 선정한 영화를 군집의 기준으로 사용하여 추천을 수행하였다.

Step 1은 전체 사용자를 train 집단과 test 집단으로 분리하는 과정이다. test 집단은 목표고객이고 train 집단은 기 사용자이다. train 집단은 군집을 통해서 목표고객의 이웃사용자가 된다. 두 집단의 비율은 80:20으로 train 집단이 745명이고 test 집단이 189명이다. 즉, 목표고객인 test 사용자 189명에 대해서 영화를 추천하고 그 결과를 분석한다.

Step 2는 train 집단을 군집하기 위해 군집의 기준으로 사용할 영화를 선정하는 단계이다. 군집은 선호도를 기준으로 수행하기 때문에 선호도가 적게 입력된 영화로 군집을 수행하게 되면 군집의 정확도가 낮아진다.

따라서 군집의 기준으로 사용할 영화의 선정은 하나의 영화에 대해서 150명 이상의 고객이 선호도를 입력한 영화 중에서 랜덤으로 10개를 선정하였다. 이 실험은 이미 선호도가 알려진 사용자들을 대상으로 하는 실험이다. 따라서 목표고객인 test 사용자들은 랜덤으로 선정한 10개 영화의 선호도 이외는 입력된 선호도가 없다고 가정한다.

Step 1. 전체 사용자를 train 사용자와 test 사용자로 분리
Step 2. 군집의 기준으로 사용할 영화 10개를 랜덤으로 선정
Step 3. Step 2에서 선정한 10개 영화의 선호도를 기준으로 train 사용자를 군집
Step 4. test 사용자의 선호도와 Step 3에서 생성한 군집의 중심간 거리를 계산하여 가장 가까운 거리의 군집을 test 사용자의 이웃사용자로 결정
Step 5. 군집별 추천영화의 후보를 선정
Step 6. Step 5에서 선정한 영화에 대해서 test 사용자의 예측 선호도를 계산
Step 7. test 사용자별 추천영화 후보에 대해서 추천순위를 계산하고 추천순위가 높은 순서로 10개의 영화를 추천

<그림 3-1> 추천 실험 절차

Step 3은 train 사용자를 군집하는 단계이다. 군집 알고리즘은 EMC 휴리스틱 군집 알고리즘을 사용한다.

군집의 기준은 Step 2에서 선정한 10개의 영화에 대해서 train 사용자가 평가한 선호도이다. 군집의 수는 5개부터 15개까지 증가시키면서 수행한다.

Step 4는 목표고객인 test 사용자의 이웃사용자를 선정하는 단계이다. test 사용자의 이웃사용자 선정은 Step 2에서 선정한 10개 영화에 대한 test 사용자의 선호도와 Step 3에서 생성한 군집의 중심간 거리를 계산하여 짧은 거리를 갖는 군집을 목표고객의 이웃사용자로 선정한다.

Step 5는 목표고객인 test 사용자에게 추천할 영화의 후보들을 선정하는 단계이다. 추천영화 후보는 test 사용자의 이웃사용자가 평가한 영화를 대상으로 한다. 추천영화 후보의 수가 증가하면 그만큼 선호도 예측에 필요한 계산량이 증가한다. 따라서 추천 시스템에 부하를 주지 않는 추천영화 후보의 수가 필요하다. 본 연구에서는 추천영화 후보의 수를 결정하는 방법으로 많은 이웃사용자가 평가한 영화 중에서 선호도 합이 큰 영화를 장르별로 3편 선정한다. 여기서 MovieLens 데이터는 중복 장르가 존재하므로 중복 장르를 재분류하기 위해 clementine 8.1의 TwoStep군집[8]을 사용하여 19개의 장르로 재분류 한다. 그리고 재분류한 19개의 장르별로 3개의 영화를 추천영화 후보로 선정한다. 이 같은 방법을 적용하면 19개 장르별로 3개의 영화를 추천하게 되므로 목표고객별 총 57개의 추천영화 후보가 선정된다.

Step 6은 Step 5에서 선정한 추천영화 후보에 대해서 test 사용자들의 선호도를 예측하는 단계이다. 선호도의 예측은 본 연구에서 제시한 이웃사용자 중심 선호도 예측방법을 사용한다.

Step 7은 Step 6에서 계산된 예측 선호도로 추천순위를 정하는 단계이다. 추천순위는 예측 선호도 값이 큰 순서로 10개를 선정하여 최종 추천목록을 생성한다.

추천목록이 생성되면 추천 시스템의 성능을 평가하기 위해 추천목록의 영화에 대한 예측 선호도와 test 사용자가 입력한 실제 선호도를 비교하여 추천의 정확도를 평가한다.

본 연구에서 제시한 추천 시스템의 성능은 MAE(Mean Absolute Error)를 사용하여 평가한다. MAE는 예측된 평가치들이 목표고객의 실제 평가치들과 평균적으로 어느 정도 유사한지를 나타내는 지표이다. 추천 시스템의 성능은 MAE 값이 작을수록 우수하다고 평가한다[5]. 따라서 MAE 값을 계산하게 되면 추천 시스템이 얼마나 정확하게 목표고객의 선호도를 예측했는지 알 수 있다. MAE는 식 (3)과 같다.

$$MAE = \frac{\sum_{m=1}^{n} |P_{A_m} - E_{A_m}|}{n} \qquad (3)$$

P_{A_m} : 목표고객 A의 실제 선호도
E_{A_m} : 목표고객 A의 예측 선호도
n : 추천영화 중에서 목표고객 A가 실제로 관람한 영화 편수

P_{A_m}는 목표고객 A가 m영화에 대해 평가한 실제 선호도이고 E_{A_m}는 추천 시스템에 의해서 예측된 선호도이다. n은 목표고객인 test 사용자에게 추천한 영화 중에서 test 사용자가 실제로 평가한 영화의 개수이다.

3.2 실험결과 및 분석

본 연구에서는 추천의 정확도를 향상시키기 위해 기 사용자들을 보다 정확하게 군집할 수 있는 EMC 휴리스틱 군집 알고리즘을 사용하였고 이웃사용자 중심 선호도 예측방법을 제시하였다. 여기서는 추천 시스템에 EMC 휴리스틱 군집 알고리즘과 이웃사용자 중심 선호도 예측방법을 적용할 경우 추천의 정확도가 향상된다는 것을 실험을 통하여 확인한다.

3.2.1 군집 알고리즘에 따른 추천 정확도

첫 번째 실험은 기 사용자 군집에 k-means, 타부 탐색, EMC 휴리스틱 군집 알고리즘을 사용하는 경우 추천의 정확도를 비교하는 것이다. 본 연구의 추천 시스템은 이웃사용자를 선정하기 위해 기 사용자를 군집한다. 따라서 기 사용자를 군집하는 알고리즘의 변경이 추천 정확도에 어느 정도 영향을 주는지 실험하였다.

이 실험은 <그림 3-1>의 추천실험 절차에서 Step 3에 해당한다. 실험의 목적은 군집 알고리즘의 차이가 추천의 정확도에 어느 정도 영향을 주는지 평가하는 것이므로 Step 6의 선호도 예측과 Step 7의 추천순위 결정은 기존 방법을 사용하여 진행하였다.

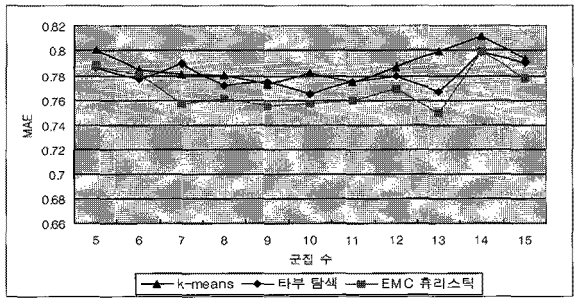
<표 3-1> k-means, 타부 탐색, EMC 알고리즘에 의한 MAE (1회)

군집	k-means		타부 탐색		EMC 알고리즘	
	제곱오차	MAE	제곱오차	MAE	제곱오차	MAE
5	60662.40	0.800	60251.81	0.787	59119.84	0.788
6	59371.09	0.785	57438.22	0.777	56578.62	0.782
7	57347.81	0.782	56575.43	0.790	55298.86	0.757
8	56814.05	0.781	55143.93	0.772	54146.97	0.762
9	55800.46	0.773	54116.13	0.776	52173.15	0.756
10	54917.95	0.783	53762.57	0.765	51147.05	0.757
11	54085.46	0.775	52848.93	0.776	50021.24	0.760
12	53271.32	0.787	51505.08	0.780	49284.51	0.770
13	53219.90	0.799	50693.43	0.767	47680.88	0.750
14	52243.98	0.811	49323.83	0.798	47795.54	0.799
15	52571.80	0.793	50010.97	0.790	47521.90	0.778
평균	55482.38	0.788	53788.21	0.780	51888.05	0.769

<표 3-1>은 1회 차 실험으로 k-means, 타부 탐색, EMC 휴리스틱 군집 알고리즘을 사용하여 기 사용자를 군집하고 그 결과로 목표고객의 이웃사용자를 선정하여 추천한 결과를 비교한 표이다. 음영부분은 두 가지 방법 중에서 가장 좋은 제곱오차와 MAE를 나타낸 것이다. EMC 휴리스틱 군집 알고리즘의 MAE가 타 군집 알고리즘에 비해서 향상된 군집 수는 7로 최대 3.20%가 향상되었다. 반면 군집 수 5, 6, 14에서는 타 알고리즘을 사용한 경우의 MAE가 EMC 휴리스틱 군집 알고리즘에 비해서 향상되었다. 모든 군집 수에 대한 평균 MAE는 k-means 알고리즘을 사용한 경우 0.788, 타부 탐색 알고리즘을 사용한 경우 0.780, EMC 휴리스틱 군집 알고리즘을 사용한 경우 0.769로 EMC 휴리스틱 군집 알고리즘을 사용한 경우가 k-means 알고리즘에 비해서 2.41%, 타부 탐색 알고리즘에 비해서 1.41% 향상되었다.

MAE 변화의 경향을 보기 위해 <표 3-1>의 결과를 <그림 3-2>와 같이 그래프로 나타내었다. MAE는 세 가지 알고리즘 모두 군집의 수가 증가할수록 향상되는 경향을 보이다가 군집 수 14에서 증가한다.

<표 3-2>는 2회 차 실험으로 1회 차 실험과는 다른 영화로 기 사용자를 군집하고 추천한 결과이다. 음영부분은 두 가지 방법 중에서 가장 좋은 제곱오차와 MAE를 나타낸 것이다. EMC 휴리스틱 군집 알고리즘의 MAE가 타 군집 알고리즘에 비해서 향상된 군집 수는 6, 10, 11, 13으로 최대 3.13% 향상되었다.



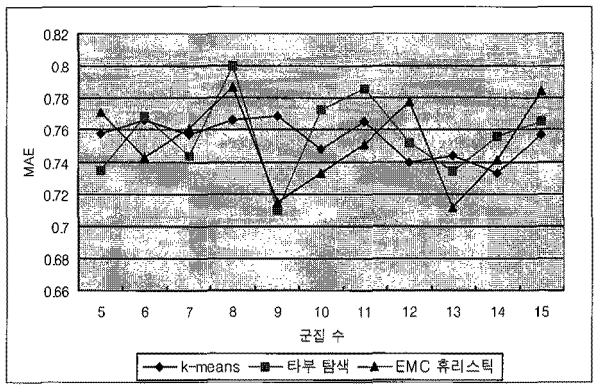
<그림 3-2> k-means, 타부 탐색, EMC 알고리즘에 의한 MAE (1회)

반면 군집 수 5, 7, 8, 9, 12, 14, 15에서는 타 군집 알고리즘을 사용한 경우의 MAE가 EMC 휴리스틱 군집 알고리즘에 비해서 향상되었다. 모든 군집 수에 대한 평균 MAE는 k-means 알고리즘을 사용한 경우 0.755, 타부 탐색 알고리즘을 사용한 경우 0.757, EMC 휴리스틱 군집 알고리즘을 사용한 경우 0.752로 EMC 휴리스틱 군집 알고리즘을 사용한 경우가 k-means 알고리즘에 비해서 0.40%, 타부 탐색 알고리즘에 비해서 0.66% 향상되었다. 1회 차 실험보다 MAE가 향상되지 않은 이유는 알고리즘 간의 제곱오차 차이가 크지 않기 때문이다.

MAE 변화의 경향을 보기 위해 <표 3-2>의 결과를 <그림 3-3>과 같이 그래프로 나타내었다. MAE는 세 가지 군집 알고리즘 모두 군집 수가 증가함에 따라서 향상되는 경향을 나타낸다. 군집 수에 따른 MAE 변화 폭은 타부 탐색 알고리즘과 EMC 휴리스틱 군집 알고리즘의 경우 크게 나타났지만 k-means 알고리즘은 크지 않았다. 이와 같은 결과는 휴리스틱 군집 알고리즘(타부 탐색, EMC)으로 기 사용자를 군집할 경우 군집의 기준으로 사용하는 영화에 따라서 MAE의 변화 폭이 커질 수 있다는 것을 나타낸다.

<표 3-2> k-means, 타부 탐색, EMC 알고리즘에 의한 MAE (2회)

군집	k-means		타부 탐색		EMC 알고리즘	
	제곱오차	MAE	제곱오차	MAE	제곱오차	MAE
5	57071.31	0.758	57193.26	0.735	57103.99	0.771
6	55299.45	0.766	55410.83	0.769	54708.39	0.743
7	53367.13	0.757	53469.14	0.744	52906.96	0.761
8	52503.94	0.766	51997.79	0.800	51711.39	0.787
9	50616.33	0.769	50919.83	0.710	49532.17	0.715
10	49689.79	0.748	49457.42	0.773	49054.57	0.733
11	49513.12	0.765	48784.44	0.786	48271.39	0.751
12	48093.04	0.740	46813.09	0.752	46779.38	0.778
13	47202.77	0.744	46656.38	0.734	45819.01	0.711
14	46580.31	0.733	46044.66	0.756	45554.20	0.741
15	44862.40	0.757	45803.99	0.765	44954.54	0.784
평균	50436.33	0.755	50231.89	0.757	49672.36	0.752



<그림 3-3> k-means, 타부 탐색, EMC 알고리즘에 의한 MAE (2회)

<표 3-3> k-means, 타부 탐색, EMC 알고리즘에 의한 MAE (3회)

군집	k-means		타부 탐색		EMC 알고리즘	
	제곱오차	MAE	제곱오차	MAE	제곱오차	MAE
5	58805.04	0.773	55053.53	0.766	56067.94	0.782
6	57258.57	0.752	53666.17	0.760	52606.48	0.751
7	55331.88	0.765	49797.55	0.752	51263.06	0.763
8	55010.40	0.742	48452.66	0.768	48886.76	0.796
9	49486.02	0.738	48248.16	0.737	47224.06	0.740
10	49181.89	0.735	46699.15	0.736	45958.86	0.735
11	47864.61	0.740	46183.31	0.730	45233.04	0.747
12	46732.92	0.763	45949.88	0.742	44589.90	0.743
13	46008.95	0.735	44693.92	0.723	43825.76	0.728
14	45698.39	0.750	43364.12	0.750	43304.10	0.732
15	45260.71	0.740	42278.49	0.748	41834.38	0.722
평균	50603.68	0.749	47671.54	0.747	47344.94	0.749

<표 3-3>은 3회 차 실험으로 1, 2회 차 실험과는 다른 영화로 기 사용자를 군집하고 추천한 결과이다. 음영부분은 두 가지 방법 중에서 가장 좋은 제곱오차와 MAE를 나타낸 것이다. EMC 휴리스틱 군집 알고리즘의 MAE가 타 군집 알고리즘에 비해서 향상된 군집 수는 6, 10, 14, 15로 최대 2.43% 향상되었다. 반면 군집 수 5, 7, 8, 9, 11, 12, 13에서는 타 알고리즘을 사용한 경우의 MAE가 EMC 휴리스틱 군집 알고리즘에 비해서 향상되었다. 모든 군집 수에 대한 평균 MAE는 k-means 알고리즘을 사용한 경우 0.749, 타부 탐색 알고리즘을 사용한 경우 0.747, EMC 휴리스틱 군집 알고리즘을 사용한 경우 0.749로 세 가지 알고리즘의 MAE는 큰 차이가 없었다. 그러나 1, 2회 차 실험과는 다르게 타부 탐색 알고리즘을 사용한 경우가 EMC 휴리스틱 군집 알고리즘을 사용한 경우에 비해서 0.27% 향상되었다. 그 이유는 타부 탐색 알고리즘의 제곱오차가 EMC 휴리스틱 군집 알고리즘에 비해서 우수한 경우가 있었기 때문이다.

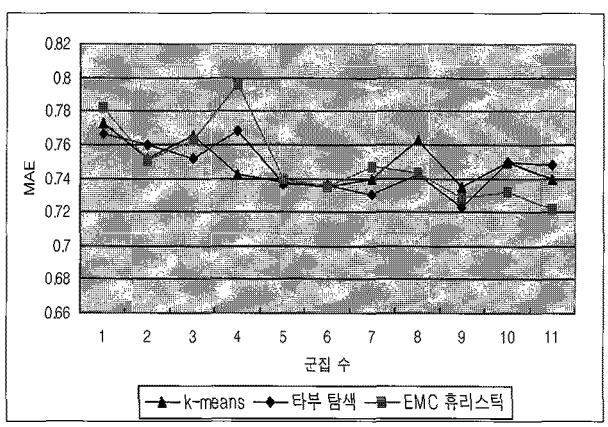
MAE 변화의 경향을 보기 위해 <표 3-3>의 결과를 <그림 3-4>와 같이 그래프로 나타내었다. MAE는 세 가지 알고리즘 모두 군집 수가 증가하면서 점차로 향상되는 경향을 나타내고 있다. 군집 수에 따른 MAE 변화 폭은 세 가지 알고리즘 모두 그 폭이 크지 않았다.

이 실험의 결과에서 3조 실험에 대한 평균 MAE는 k-means 알고리즘을 사용한 경우 0.764, 타부 탐색 알고리즘을 사용한 경우 0.761, EMC 휴리스틱 군집 알고리즘을 사용한 경우 0.757로 EMC 휴리스틱 군집 알고리즘을 사용한 경우가 k-means 알고리즘에 비해서 0.92%, 타부 탐색 알고리즘에 비해서 0.53% 향상되었다.

3.2.2 선호도 예측방법에 따른 추천 정확도

두 번째 실험은 선호도 예측에서 기존의 선호도 예측방법을 사용했을 경우와 본 연구에서 제시한 이웃사용자 중심 선호도 예측방법을 사용했을 경우 선호도 예측의 정확도를 비교하는 것이다. 첫 번째 실험에서 EMC 휴리스틱 군집 알고리즘의 성능을 확인하였으므로 기 사용자 군집은 EMC 휴리스틱 군집 알고리즘을 사용하여 수행한다. 이 실험은 <그림 3-1>의 추천실험 절차에서 Step 6에 해당한다. Step 6의 선호도 예측에서는 군집의 수를 5부터 15까지 증가시키면서 기존 선호도 예측방법과 이웃사용자 중심 선호도 예측방법을 사용하여 추천을 한다.

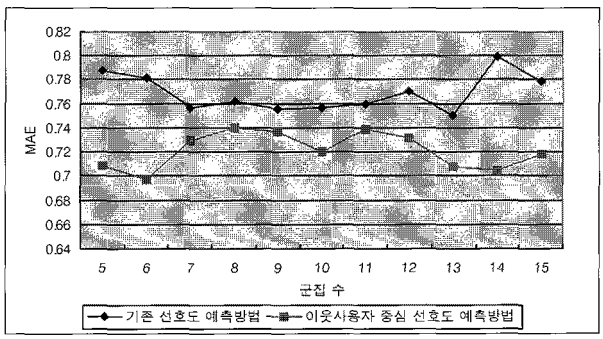
<표 3-4>는 본 연구에서 제안한 이웃사용자 중심 선호도 예측방법과 기존 선호도 예측방법을 사용하여 추천한 결과를 비교한 표이다. 음영부분은 두 가지 방법 중에서 가장 좋은 MAE를 나타낸 것이다. 1회 차 실험의 MAE는 모든 군집 수에서 이웃사용자 중심 선호도 예측방법을 사용한 경우가 기존 선호도 예측방법을 사용한 경우에 비해서 향상되었다. MAE가 크게 향상된 군집 수는 5, 6, 14로 각각 11.76%, 10.87%, 10.03% 향상되었다. 그 외의 군집 수에서는 최소 2.51%에서 최대 7.71%까지 향상되었다.



<그림 3-4> k-means, 타부 탐색, EMC 알고리즘에 의한 MAE (3회)

<표 3-4> 선호도 예측방법에 따른 추천결과 비교 (1회)

군집 수	기존 선호도 예측방법	이웃사용자 중심 선호도 예측방법
5	0.788	0.709
6	0.782	0.697
7	0.757	0.730
8	0.762	0.740
9	0.756	0.737
10	0.757	0.720
11	0.760	0.739
12	0.770	0.732
13	0.750	0.708
14	0.799	0.705
15	0.778	0.718
평균	0.769	0.721



<그림 3-5> 선호도 예측방법에 따른 MAE 변화 (1회)

모든 군집 수에 대한 평균 MAE는 이웃사용자 중심 선호도 예측방법을 사용한 경우 0.721이고 기존 선호도 예측방법을 사용한 경우 0.769로 이웃사용자 중심 선호도 예측방법이 기존 선호도 예측방법보다 6.24% 향상되었다.

MAE 변화의 경향을 보기 위해 <표 3-4>의 결과를 <그림 3-5>와 같이 그래프로 나타내었다. 기존 선호도 예측방법을 사용한 경우의 MAE는 군집 수가 증가하면서 작아지다가 군집 수 14부터 다시 커지고 있다. 반면 이웃사용자 중심 선호도 예측방법을 사용한 경우의 MAE는 군집 수가 증가하면서 커지다가 군집 수 13부터 다시 작아지고 있다. 이 같은 MAE의 변화로 인해서 군집 수가 5, 6, 14, 16에서는 이웃사용자 중심 선호도 예측방법이 기존 선호도 예측방법에 비해서 큰 차이로 향상되었다.

<표 3-5>는 2회 차 실험으로 음영부분은 두 가지 방법 중에서 가장 좋은 MAE를 나타낸 것이다. MAE는 1회 차 실험과 동일하게 모든 군집 수에서 이웃사용자 중심 선호도 예측방법을 사용한 경우가 기존 선호도 예측방법을 사용한 경우에 비해서 향상되었다.

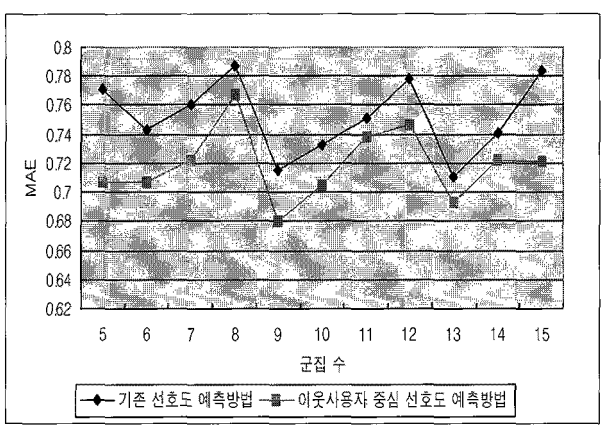
MAE가 크게 향상된 군집 수는 5, 15로 각각 8.30%, 8.04% 향상되었다. 그 외의 군집 수에서는 최소 1.73%에서 최대 5.12%까지 향상되었다. 모든 군집 수에 대한 평균 MAE는 기존 선호도 예측방법을 사용한 경우 0.752이고 이웃사용자 중심 선호도 예측방법을 사용한 경우 0.719로 이웃사용자 중심 선호도 예측방법이 기존 선호도 예측방법보다 4.39% 향상되었다.

MAE 변화의 경향을 보기 위해 <표 3-5>의 결과를 <그림 3-6>과 같이 그래프로 나타내었다. MAE 변화는 1회 차 실험과는 다르게 두 방법 모두 군집 수 8, 12에서 커지고 9, 13에서 갑자기 작아지는 경향을 보이고 있다.

<표 3-6>은 3회 차 실험으로 음영부분은 두 가지 방법 중에서 가장 좋은 MAE를 나타낸 것이다. MAE는 군집 수 5, 13, 14를 제외한 군집 수에서 이웃사용자 중심 선호도 예측방법을 사용한 경우가 기존 선호도 예측방법을 사용한 경우에 비해서 향상되었다. MAE가 크게 향상된 군집 수는 8로 9.30% 향상되었다. 그 외의 군집 수에서는 최소 0.40%에서 최대 8.92%까지 향상되었다. 모든 군집 수에 대한 평균 MAE는 기존 선호도 예측방법을 사용한 경우 0.749이고 이웃사용자 중심 선호도 예측방법을 사용한 경우 0.730으로 이웃사용

<표 3-5> 선호도 예측방법에 따른 추천결과 비교 (2회)

군집 수	기존 선호도 예측방법	이웃사용자 중심 선호도 예측방법
5	0.771	0.707
6	0.743	0.707
7	0.761	0.722
8	0.787	0.768
9	0.715	0.680
10	0.733	0.705
11	0.751	0.738
12	0.778	0.747
13	0.711	0.693
14	0.741	0.722
15	0.784	0.721
평균	0.752	0.719



<그림 3-6> 선호도 예측방법의 종류에 따른 MAE 변화 (2회)

자 중심 선호도 예측방법이 기존 선호도 예측방법보다 2.54% 향상되었다.

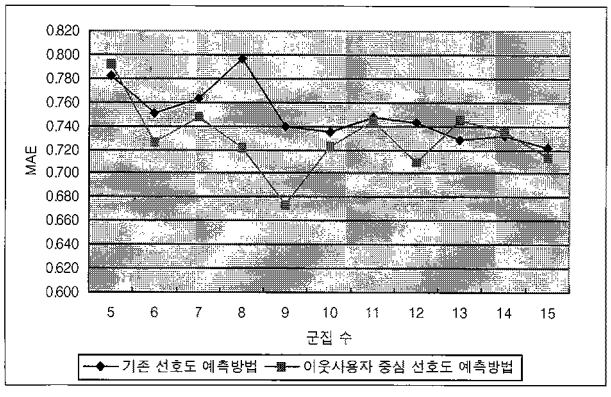
MAE 변화의 경향을 보기 위해 <표 3-6>의 결과를 <그림 3-7>과 같이 그래프로 나타내었다. MAE는 기존 선호도 예측방법을 사용했을 경우 군집 수가 증가하면서 점차 작아지고 있다. 또한 이웃사용자 중심 선호도 예측방법을 사용했을 경우도 군집 수가 증가하면서 다소 작아지는 경향을 나타내고 있다. 군집 수 5, 13, 14에서는 기존 선호도 예측방법을 사용한 경우가 이웃사용자 중심 선호도 예측방법을 사용한 경우에 비해서 향상되었다.

이 실험의 결과에서 3조 실험에 대한 평균 MAE는 기존의 선호도 예측방법을 사용한 경우 0.757이고 이웃사용자 중심 선호도 예측방법을 사용한 경우 0.723으로 이웃사용자 중심 선호도 예측방법이 기존의 선호도 예측방법에 비해서 4.41% 향상되었다.

<표 3-7>은 지금까지의 두 가지 실험결과를 요약한 표이다. MAE는 3조의 실험에 대해서 모든 군집 수에 대한 평균 MAE를 나타내었다. <표 3-7>과 같이 MAE는 본 연구에서 제안한 방법을 사용함으로써 점차 향상되는 것을 확인할 수 있다.

<표 3-6> 선호도 예측방법에 따른 추천결과 비교 (3회)

군집 수	기존 선호도 예측방법	이웃사용자 중심 선호도 예측방법
5	0.782	0.792
6	0.751	0.726
7	0.763	0.747
8	0.796	0.722
9	0.740	0.674
10	0.735	0.723
11	0.747	0.744
12	0.743	0.709
13	0.728	0.745
14	0.732	0.735
15	0.722	0.714
평균	0.749	0.730



<그림 3-7> 선호도 예측방법의 종류에 따른 MAE 변화 (3회)

<표 3-7> 제안방법별 MAE 변화

제안방법	평균 MAE	
	적용 전	적용 후
EMC 알고리즘	0.761	0.757
EMC 알고리즘과 이웃사용자 중심의 선호도 예측방법	0.757	0.723

4. 결 론

본 연구에서는 협력적 필터링 기법을 적용한 추천 시스템에서 추천의 정확도를 향상시키기 위해 다음과 같은 두 가지 방법을 제시하였다.

첫째, EMC 휴리스틱 군집 알고리즘으로 기 사용자를 군집하였다. EMC 휴리스틱 군집 알고리즘은 실제로 고객의 선호도를 조사한 MovieLens 데이터를 사용한 추천실험에서 추천 시스템의 성능평가 척도인 MAE를 향상시켰다. 실제로 기 사용자 군집의 실험에서 EMC 휴리스틱 군집 알고리즘의 군집 성능을 평가하는 척도인 제곱오차는 k-means 알고리즘에 비해서 5.86% 향상되었고 타부 탐색 알고리즘에 비해서 3.02% 향상되었다. 이렇게 하여 향상된 군집의 효율은 결과적으로 추천 시스템의 성능을 향상시킬 수 있었다.

둘째, 기존의 선호도 예측방법을 개선한 이웃사용자 중심 선호도 예측방법을 제시하였다. 추천 시스템에서 이웃사용자 중심 선호도 예측방법을 사용하여 선호도를 예측한 결과 추천 시스템의 성능평가 척도인 MAE는 기존의 선호도 예측방법을 사용한 경우에 비해서 4.41%가 향상되었다. 이와 같이 추천 시스템의 성능이 향상된 이유는 본 연구가 새롭게 제안한 이웃사용자 중심의 선호도 예측방법을 사용하여 평균 선호도 계산의 정확도를 높였기 때문이다. 기존의 선호도 예측방법은 목표고객이 입력한 선호도만을 사용하여 평균 선호도를 계산하였다. 이 방법은 목표고객이 입력한 선호도의 양이 적을 경우 선호도 예측의 정확도를 저하시킬 수 있다. 본 연구에서는 이 문제를 해결하기 위해 평균 선호도 계산에서 목표고객에 비해 항상 많은 선호도 정보량을 갖는 이웃사용자의 선호도를 사용하였다. 그 결과 평균 선호도의 정확도가 향상되어 결과적으로 추천 시스템의 성능을 향상시킬 수 있었다.

본 연구에서 제시한 두 가지 방법 중에서 이웃사용자 중심 선호도 예측방법은 추천 시스템의 성능을 가장 크게 향상시켰다. 그 이유는 기존의 선호도 예측 방법에 비해서 선호도 예측에 사용하는 정보량을 크게 늘렸기 때문이다. EMC 휴리스틱 군집 알고리즘도 추천 시스템의 성능을 향상시켰으나 그 향상 폭은 이웃사용자 중심 선호도 예측방법에 비해서 크지 않았다.

5. 참 고 문 헌

[1] 이석환, 박승헌, 군집의 효율향상을 위한 휴리스틱알고리즘, 대한안전경영과학회, 제11권 제3호, 2009, 157-166
[2] 이재식, 박석두, 장르별 협업필터링을 이용한 영화 추천 시스템의 성능 향상, 한국지능정보시스템학회, 제13권 제4호, 2007, 65-78
[3] 정경용, 이정현, 개인화 추천 시스템의 예측 정확도 향상을 위한 사용자 유사도 가중치에 대한 비교 평가, 전자공학회, 제42권 제6호, 2005, 63-73
[4] Ahn, H. J. A new similarity measure for collabo rative filtering to alleviate the new user cold-starting problem. Information Sciences, 178, 2008, 37-51
[5] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, Item-Based Collaborative Filtering Recommendation Algorithms, Association for Co mputing Machinery, 2001, 285-295
[6] Badrul Sarwar, Sparsity, scalability, and distribution in recommender systems. PhD thesis, University of Minnesota, 2001
[7] Herlocker, J., Konstan, J. A., & Riedl, J. E xplaining collaborative filtering recommendation s. In ACM 2000 conference on computer-supported collaborative work, 2000, 241-250
[8] SPSS, Clementine 8.0 user's guide, SPSS, 2003
[9] MovieLens dataset, URL : http://movielens.umn.edu

저 자 소 개

이 석 환

인하대학교 산업공학과에서 공학사, 공학석사 학위를 취득하였다. 주요 관심 분야는 데이터 마이닝이다.

주소: 인천광역시 남구 용현동 253, 인하대학교 산업공학과

박 승 헌

인하대학교 금속공학과에서 공학사, 일본 Keio대학 관리공학과에서 공학석사 및 공학박사 학위를 취득 하였다. 현재 인하대학교 산업공학과 교수로 재직 중이다. 주요 관심 분야는 FMS와 각종 생산시스템의 설계 및 운영, 인터넷 마케팅과 데이터 마이닝 등이다.

주소: 인천광역시 남구 용현동 253, 인하대학교 산업공학과