

협력적 여과 시스템에서 사용자 변동 계수를 이용한 기본 평가값 예측

(Default Voting using User Coefficient of Variance in Collaborative Filtering System)

고 수 정 [†]

(Su Jeong Ko)

요 약 협력적 여과 시스템에서 대부분의 사용자들은 모든 아이템에 대하여 선호도를 평가하지 않으므로 인하여 사용자-아이템 행렬은 희박성을 나타내며, 또한 사용자가 평가하지 않은 아이템으로부터 결측치가 발생한다. 일반적인 결측치 예측 방법은 특정 대상의 사용자가 평가하지 않은 결측치를 이 사용자와 비슷한 흥미를 갖는 사용자들의 평가값을 기반으로 예측하나, 기본 평가값 예측 방법은 사용자-아이템 행렬의 결측치를 특정 사용자가 아닌 전체 사용자에 대하여 예측한다. 기본 평가값 예측 방법 중 가장 많이 사용되는 방법은 아이템 평균이나 사용자 평균을 이용한 방법이다. 그러나 이 방법은 아이템이나 사용자의 특성, 또한 데이터 집합의 분포 특성을 전혀 고려하지 않는다는 문제점을 갖는다. 본 논문에서는 이러한 문제점을 해결하기 위하여 데이터 집합에 나타난 사용자의 변동 계수를 이용하는 기본 평가값 예측 방법을 제안한다. 제안한 방법에서는 수식을 이용하여 자동적으로 사용자 변동 계수의 임계값을 선택하고, 그 임계값에 따라 사용자 평균에서 아이템 평균으로 전환하여 사용자들의 결측치에 대한 기본 평가값을 결정한다. 그러나 사용자 변동 계수들의 분포 정보로 인하여 사용자 변동 계수와 임계값이 항상 일정한 관계를 유지하는 것이 아니므로, 제안된 방법에서는 임계값을 선택하기 위하여 사용자 변동 계수의 평균과 변동 계수의 분포 정보를 병합한다. 제안된 방법은 사용자가 영화에 대하여 평가한 MovieLens 데이터 집합을 대상으로 평가되었으며, 기존의 기본 평가값 예측 방법보다 그 성능이 우수함을 보인다.

키워드 : 기본 평가값 예측, 사용자 변동 계수, 협력적 여과 시스템

Abstract In collaborative filtering systems most users do not rate preferences; so User-Item matrix shows great sparsity because it has missing values for items not rated by users. Generally, the systems predict the preferences of an active user based on the preferences of a group of users. However, default voting methods predict all missing values for all users in User-Item matrix. One of the most common methods predicting default voting values tried two different approaches using the average rating for a user or using the average rating for an item. However, there is a problem that they did not consider the characteristics of items, users, and the distribution of data set. We replace the missing values in the User-Item matrix by the default voting method using user coefficient of variance. We select the threshold of user coefficient of variance by using equations automatically and determine when to shift between the user averages and item averages according to the threshold. However, there are not always regular relations between the averages and the thresholds of user coefficient of variances in datasets. It is caused that the distribution information of user coefficient of variances in datasets affects the threshold of user coefficient of variance as well as their average. We decide the threshold of user coefficient of variance by combining them. We evaluate our method on MovieLens dataset of user ratings for movies and show that it outperforms previously default voting methods.

Key words : default voting, user coefficient of variance, collaborative filtering system

· 이 연구는 인덕대학 학술연구비 일부 지원에 의하여 수행되었음

† 정 회 원 : 인덕대학 컴퓨터소프트웨어과 교수

sjko@induk.ac.kr

논문접수 : 2005년 5월 26일

심사완료 : 2005년 9월 13일

1. 서 론

협력적 여과 기술은 사용자-아이템 행렬의 고차원성과 데이터 집합에 나타난 결측치로 인한 희박성으로 인

하여 그 한계를 나타낸다[1,2]. 결측치가 없는 사용자-아이템 행렬을 구성하기 위해서는 사용자들이 모든 아이টে에 대해 선호도를 평가해야 할지라도 대부분의 사용자가 그들이 흥미를 느끼는 아이টে에 한하여 평가를 하기 때문에 전혀 결측치가 없는 행렬을 구성하는 것은 불가능하다. 결측치는 사용자-아이템 행렬에 나타난 여러 평가값 중 '0'의 값과는 다르다. 일반적인 결측치 예측 방법은 새로운 사용자와 비슷한 흥미를 나타내는 사용자들을 사용자-아이템 행렬로부터 찾아서 그 사용자들의 평가값을 기반으로 새로운 사용자의 결측치를 예측하는 방법을 사용한다. 반면, 기본 평가값 예측(default voting) 방법은 모델 기반의 협력적 여과 방법들[3]을 사용하고자 할 때 사용되는 방법으로, 결측치의 일부가 아닌 모든 값을 예측할 때 사용되는 방법이다. 모델 기반의 협력적 여과 방법은 사용자 군집이나 차원 감소 등을 이용하여 사용자-아이템 행렬의 희박성과 고차원성을 해결하는 방법이다. 기본 평가값 예측 방법은 이를 위한 전처리 과정으로서 사용된다.

일반적인 결측치 예측 방법은 광범위하게 연구되어 왔으나[2] 결측치 예측의 전처리로서 사용되는 기본 평가값 예측에 관한 연구는 미흡한 편이다[1,4-8]. [4]는 모든 사용자가 평가하지 않은 몇몇 수에 대한 아이টে의 기본 평가값으로 동일한 값을 배정하였다. 이 방법은 단순한 방법이나 사용자에 대한 평가 특성을 고려하지 않았다는 단점을 갖는다. [5]는 50개의 아이টে보다 적은 수의 아이টে를 평가한 사용자들에 대하여 이들을 공통 집합으로 구분한 뒤, 그들의 가중치를 감소시키고 그 가중치의 의해 기본 평가값을 계산하였다. [6]은 사용자들이 마우스로 클릭한 아이টে 평가값의 평균을 기반으로 평가 히스토리를 확장하여 이를 기본 평가값을 결정하였다. [5,6]의 방법은 단지 적은 수의 사용자들을 대상으로 하여 사용자-아이টে 행렬의 희박성을 제거하는 경우에만 적합하다는 단점을 갖는다. 또 다른 방법으로, [1]은 아이টে에 대한 평균과 사용자에 대한 평균을 사용하는 접근 방식을 사용하였다. 이 방법은 사용자-아이টে 행렬에 사용자 평균이나 아이টে 평균 중 하나의 값을 사용하여 행렬의 희박성을 해결하는 가장 대중화된 방법 중 하나이며, 또한 아이টে 평균을 사용하는 것이 보다 좋은 결과를 보였음을 기술하였다. 그러나 기본 평가값으로 확실적인 아이টে 평균값을 사용하는 방법은 아이টে이나 사용자의 특성, 또한 데이터 집합의 분포 특성을 전혀 고려하지 않은 방법이므로, 이에 대한 보완이 필요하다.

본 논문에서 제안한 방법은 사용자 변동 계수(coefficient of variance)와 변동 계수의 분포 정보를 기반으로 사용자 평균이나 아이টে 평균을 사용하는 기본 평가

값 예측 방법이다. 이 방법은 모든 결측치를 채우기 위해 사용자 평균이나 아이টে 평균을 확실히 사용하여 사용하는 기존 방법의 문제점을 해결한다. 사용자의 변동 계수가 낮을 경우 이것은 사용자가 아이টে들에 대하여 독특한 성향으로 평가할 확률이 적다는 것을 의미한다. 따라서 이와 경우에 있어서 결측치로서 아이টে 평균보다는 사용자 평균을 사용하는 것이 보다 효율적이다. 반면, 사용자의 변동 계수가 높을 경우 그 사용자는 다양한 값으로 아이টে에 대해 평가할 확률이 높으므로, 기본 평가값으로 아이টে 평균을 사용하는 것이 바람직하다. 제안된 방법은 기본 평가값 예측을 위해 변동 계수의 임계값을 기준으로 사용자 평균에서 아이টে 평균으로 전환한다. 그러나 데이터 집합에 있는 사용자 변동 계수의 평균과 그 임계값이 항상 일정한 관계를 갖는 것은 아니다. 왜냐하면 사용자 변동 계수는 그 평균 뿐만 아니라 분포 정보가 임계값에 영향을 주기 때문이다. 따라서 임계값을 결정하기 위해서는 변동 계수의 평균과 변동 계수의 분포 정보를 병합하는 과정이 필요하다. 본 논문에서는 사용자 평균에서 아이টে 평균으로 전환하는 임계값을 수식을 이용하여 자동으로 결정하는 알고리즘을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 사용자 변동 계수를 이용한 기본 평가값 예측 방법의 구조도를 보인다. 3장에서는 사용자 변동 계수와 그 임계값에 대해 기술하며, 4장에서는 기본 평가값 예측 방법을 기술한다. 5장에서는 제안된 방법의 성능을 분석하고, 마지막으로 6장에서는 결론과 향후 연구 과제를 제시한다.

2. 사용자 변동 계수를 이용한 기본 평가값 예측 방법의 구조도

본 논문에서는 사용자-아이টে 행렬을 $R = \{r_{ij}\} (i=1, 2, \dots, n \text{ and } j=1, 2, \dots, m)$ 로 표현한다. 행렬 R 은 m 개의 아이টে의 집합과 n 명의 사용자 집합으로 구성되며 r_{ij} 는 아이টে d_j 에 대한 사용자 u_i 의 선호도 값을 나타낸다. 특히 아이টে의 집합은 $I = \{d_j\} (j=1, 2, \dots, m)$ 로서 표현하며 사용자 집합은 $U = \{u_i\} (i=1, 2, \dots, n)$ 로서 표현한다. 아이টে의 선호도는 0~1까지 0.2간격으로 6개의 단계로 구성된다. 0.5보다 큰 값을 갖는 아이টে는 사용자가 흥미를 보이는 아이টে로서 분류된다. 그림 1은 행렬 R 에 대한 모든 결측치를 본 논문에서 제안한 기본 평가값 예측 방법에 의해 채우는 구조도를 나타낸다. 그림 1의 구조도는 사용자 변동 계수를 계산하는 부분, 사용자 변동 계수의 임계값을 결정하는 부분, 그리고 기본 평가값을 예측하는 부분으로 총 3단계로 구성된다.

그림 1의 사용자 변동 계수를 구하는 단계에서는 사용자-아이টে 행렬에서 각 사용자를 중심으로 아이টে에

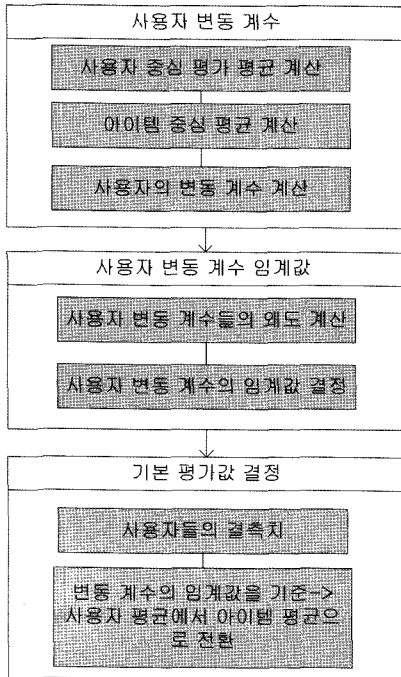


그림 1 기본 평가값 예측을 위한 구조도

대하여 평가한 값의 평균을 계산하고, 이를 이용하여 각 사용자의 변동 계수를 계산한다. 또한 각 아이템을 중심으로 그 아이템 평가에 대한 평균을 계산한다. 사용자 평균이나 아이템 평균 중 어느 값을 기본 평가값으로 사용할 것인가를 결정하기 위해 MovieLens 데이터 집합[9]에서 100명의 사용자씩을 임의로 선택하여 5개의 모집단을 만들었다. 이들 데이터 집합의 실제 데이터를 결측치로 가정한 후에, 이에 대한 기본 평가값으로 사용자 평균과 아이템 평균을 각각 사용하였다. 그 결과, 2개의 집단에서는 사용자 평균을 사용하였을 경우 결측치 오차가 낮아지는 결과를 보였으며, 나머지 집단에서는 아이템 평균을 사용하였을 경우 결측치 오차가 낮아지는 결과를 보였다. 이와 같은 결과는 사용자들이 아이템에 대하여 평가한 경향에 따라 두 가지 값 중 하나의 값을 기본 평가값으로 결정해야 한다는 결론을 내릴 수 있다. 본 논문에서 제안한 방법에서는 사용자들이 아이템에 대하여 평가한 경향을 판단하는 기준으로 데이터 점들의 변동의 정도를 나타내는 변동 계수를 사용한다.

다음 단계에서는 각 사용자들에 대한 기본 평가값을 결정하기 위하여 사용자 변동 계수의 임계값을 계산한다. 사용자-아이템 행렬에 속한 사용자들의 변동 계수 평균에 따라 사용자 평균에서 아이템 평균으로 전환되는 임계값이 달라지나 사용자 변동 계수의 분포 또한 그 임계값에 영향을 미친다. 따라서 임계값을 결정하기

위하여 사용자 변동 계수의 분포 정도를 나타내는 왜도 (skewness of distribution)를 사용자 변동 계수 평균의 보정값으로서 사용한다. 이러한 임계값은 사용자 변동 계수의 평균과 그 분포의 규칙성을 이용하여 자동으로 결정한다.

마지막 단계에서는 변동 계수의 임계값을 기준으로 사용자가 평가하지 않은 결측치를 기본 평가값으로 채운다. 이를 위하여 각 사용자의 변동 계수가 임계값보다 작다면 사용자 평균을, 아닐 경우 아이템 평균을 사용한다.

3. 사용자 변동 계수와 변동 계수의 임계값

본 장에서는 사용자가 아이템에 대하여 평가한 값을 기반으로 사용자의 변동 계수를 계산하고, 변동 계수의 평균과 변동 계수들의 분포를 나타내는 왜도를 병합하여 변동 계수의 임계값을 계산하는 과정을 기술한다.

3.1 사용자 변동 계수

변동 계수는 집합에 분포한 데이터의 분포가 얼마나 다양한 가에 대한 정도를 나타낸다[10]. 이 변동 계수는 데이터 집합의 평균을 고려하므로 다른 용어로 상대 표준 편차라고 불리기도 한다. 식 (1)은 변동 계수를 계산하기 위한 식이다.

$$CV = (S / X) \times 100 \quad (1)$$

식 (1)에서 CV는 변동 계수를 의미하고, S는 표준 편차, 그리고 X는 사용자가 아이템에 대하여 평가한 값의 평균을 나타낸다.

변동 계수의 값이 커질수록 데이터 집합의 변동성은 커진다. 즉, 사용자가 아이템에 대하여 평가한 값의 변동 계수가 작다는 것은 평가한 값의 변동성이 낮다는 의미이며, 이는 또한 사용자가 평가하지 않은 아이템에 대한 평가값 역시 평균과 가까운 범위 내에서 평가할 가능성이 높음을 의미한다. 따라서 이러한 경우에는 결측치를 채우기 위한 기본 평가값으로서 사용자 평균을 사용하는 것이 유리하다. 반면, 사용자가 아이템들에 대하여 평가한 값들의 변동 계수가 클 경우에는 결측치에 대한 기본 평가값으로서 아이템 평균을 사용하는 것이 보다 합리적이다. 왜냐하면, 평가하지 않은 아이템에 대해 사용자가 어느 정도의 값으로 평가할 것인가에 대하여 예측하기 어렵기 때문이다. 이와 같은 이론을 적용하기 위하여 변동 계수의 임계값을 사용한다. 즉, 사용자 변동 계수가 임계값 이상이면 아이템 평균을 사용하고, 임계값 이하이면 사용자 평균을 사용한다.

3.2 변동 계수의 임계값

결측치를 예측해야 하는 데이터 집합들은 모두 다른 평가 정보를 갖고 있다. 따라서 집합들에 대한 변동 계수의 임계값은 데이터 집합의 평가 정보를 반영해야 한

다. 이론상으로는 각 집합의 사용자들의 평균 변동 계수의 평균이 낮으면 사용자들이 아이টে들에 대해 평가한 값의 변동이 작으므로 변동 계수 임계값은 상대적으로 높아질 것이다. 사용자 변동 계수가 임계값 이상인 경우에만 아이টে 평균을 사용하기 때문이다. 그러나 각 집합에 대한 사용자들의 평균 변동 계수의 평균과 변동 계수의 임계값은 반드시 역비례하지 않는다. 이와 같은 결과는 사용자 변동 계수의 분포가 데이터 집합마다 다르므로 변동 계수의 전체 평균만을 이용하여 임계값을 결정하기 어렵기 때문이다. 따라서 사용자들의 평균 변동 계수에 그 변동 계수의 분포 정보를 추가하여 변동 계수의 임계값을 계산한다.

본 논문에서는 사용자 변동 계수의 분포 정도를 구하기 위하여 식 (2)의 왜도를 사용하였다. 식 (2)의 왜도는 [11]에서 사용되었으며, 통계학에서 통상적으로 사용되고 있는 수식이다.

$$skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{cv_i - CV}{s} \right)^3 \quad (2)$$

식 (2)에서 CV는 사용자 변동 계수의 평균을 나타내고, S는 변동 계수의 표준 편차, n은 사용자-아이টে에 있는 전체 사용자의 수이다.

사용자 변동 계수들이 정상분포 상태에 있을 경우 왜도는 0이며, 대칭적인 데이터는 0에 가까운 왜도를 갖는다. 500명의 사용자를 대상으로 실험한 결과, 사용자의 변동 계수 분포에 대한 왜도는 0보다 크고 3보다 작은 값을 나타내었다. 양의 값을 갖는 왜도는 그래프의 모양이 오른쪽으로 기울어졌다는 것을 나타내며, 오른쪽으로 기울어졌다는 것은 오른쪽 꼬리부분이 왼쪽 꼬리 부분보다 더 무겁다는 것을 의미한다. 사용자 변동 계수의 평균은 1보다 작은 범위의 값을 갖는다. 따라서 왜도를 사용자들의 평균 변동 계수를 보정하는 값으로 변환하기 위해서는 그 보정값이 1보다 작은 값을 가져야 한다. 이와 같이 임계값이 사용자 변동 계수에 역비례한다는 이론과 사용자 변동 계수가 1보다 같은 보정값을 가져야 한다는 이론을 기반으로 본 논문에서는 임계값을 식 (3)과 같이 정의한다. 식 (3)에서 사용자 변동 계수에 대한 보정값을 1보다 작게 만들기 위하여 변동 계수의 분포 정보를 나타내는 왜도의 값 중에서 1보다 큰 값을 갖는 값에 한하여 그 값에 로그를 취한다. 식 (3)은 임계값에 대한 사용자 변동 계수와 왜도와의 관계를 새롭게 정의한 식이다.

$$Threshold_{cv} = 1 - (CV + C_difference)$$

$$C_difference = \log skewness, (skewness > 0) \quad (3)$$

식 (3)에서 $Threshold_{cv}$ 는 사용자 변동 계수의 임계값을 나타내며, $C_difference$ 는 사용자 변동 계수의 보정

값을 나타낸다.

반면, 왜도가 0보다 크고, 1과 같거나 작은 범위에 속할 경우, 이에 대한 로그의 값은 음의 값이 나오거나 계산할 수 없는 경우가 발생한다. 이와 같은 음의 값은 평균 사용자 변동 계수에 대한 보정값으로 사용할 수 없다. 따라서 이와 같은 범위의 왜도에 대해서는 새로운 함수가 정의되어야 한다. 따라서 본 논문에서는 왜도가 0보다 크고 1과 같거나 작은 경우 식 (4)를 적용하여 평균 사용자 변동 계수에 대한 보정값을 정의한다.

$$C_differece = 10^{skewness - e^{0.88}} \quad (0 < skewness \leq 1) \quad (4)$$

그림 2는 식 (4)를 유도하기 위한 그림이다. 그림 2는 왜도가 0보다 크고 1과 같거나 작은 경우의 보정값을 구하는 새로운 함수를 정의하기 위하여 함수 $y = \log(skewness)$ 와 이의 역함수인 $y = 10^{skewness}$ 를 그린 그래프이고, 각각 'Differential correction1'과 'Differential correction1'로 표기한다.

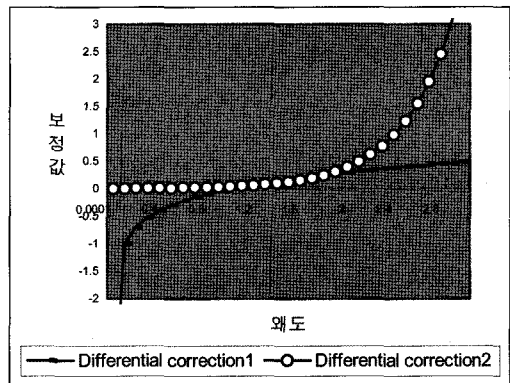


그림 2 왜도가 0보다 크고 1보다 같거나 작은 경우의 보정값을 구하기 위한 그래프

그림 2에서 $y = \log(skewness)$ 의 함수가 그 역함수 $y = 10^{skewness}$ 와 만날 수 있도록 $y = 10^{skewness}$ 를 수평 이동한다. 그 결과, $y = 10^{skewness - e}$ 의 함수를 재정의할 수 있다. 왜도의 값이 1보다 큰 최소의 값을 갖는 k를 계산할 경우, $k = 0.88$ 의 값에서 두 함수는 1보다 큰 최소의 값으로 접점을 갖는다. 따라서 식 (4)와 같은 식으로 유도된다.

4. 기본 평가값 예측

3장에서는 사용자들의 변동 계수와 그 변동 계수들의 왜도를 계산하고, 이를 기반으로 변동 계수의 임계값을 결정하였다. 본 장에서는 그 임계값에 따라 기본 평가값을 예측하는 방법의 예를 기술한다. 그림 3의 알고리즘은 변동 계수와 왜도를 계산하여 임계값을 결정 한 후 기본 평가값을 예측하는 알고리즘이다.

```

Algorithm Default_Voting(R)
R={rij} // 사용자-아이템 행렬
Default_Value={dvi} //사용자 ui :기본 평가값
Thresholdcv //사용자 변동 계수의 임계값
cvi // 사용자 ui의 변동 계수
Skewness //사용자 변동 계수에 대한 왜도

/* 사용자 변동 계수 평균 계산 */
     $CV \leftarrow (\sum_i cv_i) / n$ 
/* 사용자 변동 계수에 대한 왜도 계산 */
     $skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \frac{(cv_i - CV)^3}{s}$ 
/* 임계값 정의 */
    Thresholdcv=1-
        (CV+ Correction_difference(Skewness))

/* 각 사용자에 대한 기본 평가값 결정 */
For each user ui
    If cvi > Thresholdcv
        then dvi <- the item averages //아이템평균
        else dvi <- the user averages //사용자평균
Endfor
End Default_Voting

Function Correction_difference(Skewness)
If Skewness ≤ 1 then
     $C\_difference \leftarrow 10^{Skewness - e^{0.88}}$ 
Else
     $C\_difference \leftarrow \log Skewness$ 
Return C_difference
End Correction_difference
    
```

그림 3 기본 평가값 예측을 위한 알고리즘

그림 3에서는 사용자-아이템 행렬을 정의하고, 사용자에 대한 기본 평가값들과 사용자 변동 계수의 임계값을 위한 변수를 정의한다. 다음으로, 정의된 사용자-아이템 행렬에 속한 사용자들 각각의 변동 계수와 전체 사용자들에 대한 변동 계수의 왜도를 정의한다. 그 다음 단계에서는 계산한 왜도를 기반으로 사용자-아이템 행렬의 사용자 변동 계수 평균의 보정값을 계산한다. 계산후,

사용자 변동 계수 평균과 보정값을 기반으로 사용자 변동 계수의 임계값을 결정한다. 마지막으로, 사용자 각각에 대하여 그 사용자의 변동 계수가 임계값보다 큰 경우는 아이템 평균을, 아닐 경우는 사용자 평균을 기본 평가값으로 결정한다.

표 1은 결측치가 있는 사용자-아이템 행렬을 나타낸다.

표 1에서 d1,d2,...,d14는 사용자-아이템 행렬에 나타난 결측치를 나타낸다. 이들 d1,d2,...,d14는 사용자 변동 계수를 이용한 기본 평가값 예측을 위한 알고리즘에 따라 그 결측치들을 예측한다. 이와 같이 희박성을 제거한 사용자-아이템 행렬은 새로운 사용자가 발생하였을 경우 그 사용자에게 아이템을 추천하고자 하는 목적으로 사용하는 모델 기반의 예측 알고리즘의 정확도를 높일 수 있다.

표 1의 사용자-아이템 행렬의 결측치들을 제거하기 위하여 식 (3)과 식 (4)에 의해 사용자 변동 계수의 임계값을 구한다. 이 임계값을 기준으로 사용자 평균에서 아이템 평균으로 전환하여 표 1에 나타난 사용자-아이템 행렬의 결측치를 기본 평가값으로 채운다. 그림 4은 기본 평가값을 구하는 알고리즘을 이용하여 계산된 기본 평가값을 나타낸다.

그림 4에서와 같이 표 1에 있는 d1,d2,...,d14의 결측치는 각각 0.733,0.733,...,0.4로 채워진다.

5. 성능 평가

본 논문에서 제안한 방법의 성능을 평가하기 위하여 기존에 제안되었던 기본 평가값 예측 방법들과 비교하였다. 또한, 제안한 방법이 결측치와 희박성의 문제점을 효과적으로 보완할 수 있다고 가정하고, 그들의 매개변수에 대한 사용의 민감성을 실험하였다. 사용자 변동 계수를 사용한 기본 평가값 예측 방법의 성능을 평가하기 위해 평균 절대 오차(MAE)[12,13]를 사용하였다. 성능 평가 결과의 분석은 대응일치 t-검증(paired t-test) [3,14]과 95%의 신뢰도 수준에서 본페로니 절차(Bonferroni procedure for multiple comparison statistics)의 에이노바(ANOVA) 분석[15,16]을 사용한다. MAE는 사용자-아이템 행렬에서 실제 평가값과 예측값 사이의 차이를 기반으로 그 정확도를 측정한다. MAE는 추천의 정확도를 측정하기 위해 가장 통상적으로 사용하는 방법이며, 가장 쉬운 방법이기도 하다. n개의 결측치를 갖는 데이터 집합에서 실제 평가 값이 ai이고, 예측된 값이 pi 인 경우, MAE는 식 (5)에 의해 계산된다.

$$MAE = \frac{\sum_{i=1}^n a_i - p_i}{n} \tag{5}$$

표 1 결측치가 있는 사용자-아이템 행렬의 예

	1	2	3	4	5	6	7	8	9	10	11	12	13	평균	사용자 변동 계수
1	d1	0.6	0.4	0.8	0.2	0.2	0.4	0.8	0.2	0.6	0.8	0.4	0.8	0.517	0.480
2	0.8	0.4	0.4	0.8	0.8	0	1	0.6	0.6	1	0	d2	0.6	0.583	0.575
3	1	0.8	0.8	0.8	0.8	d3	d4	0.8	0.6	0.8	0	1	1	0.764	0.367
4	0.2	0	0.4	0.4	d5	0	0.6	0	0.4	0	0.4	0.4	d6	0.255	0.867
5	0.6	d7	0.4	0.8	0.4	0	0.4	1	0	0.8	0.8	1	0.4	0.550	0.623
6	0.8	0	0.2	1	0	0	0.6	0.4	d8	1	1	0.4	1	0.533	0.789
7	0.6	0.6	0.6	0.8	0	0.2	0.6	0.8	0.6	0.8	d9	0.8	d10	0.582	0.447
8	1	0.8	0.8	0.8	0.2	0.4	0.8	0.8	0	0.8	1	1	d11	0.700	0.464
9	1	0.8	0.8	0	d12	0.2	0.6	0	0.8	0.8	0.8	1	0.4	0.600	0.603
10	0.6	0.6	1	0.8	d13	0	0.8	0.2	d14	0.8	0.8	0.6	0.8	0.636	0.462
평균	0.73	0.51	0.58	0.7	0.34	0.11	0.64	0.54	0.4	0.74	0.62	0.73	0.71	0.572	0.568

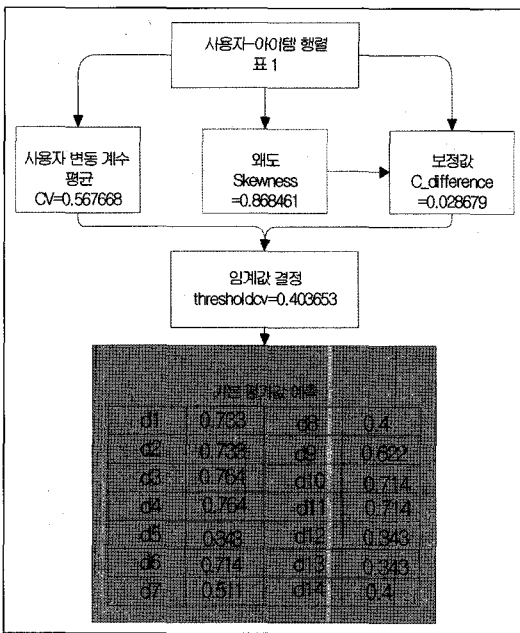


그림 4 표 1의 사용자-아이템 행렬의 기본 평가값 예측

사용자 변동 계수를 이용한 기본 평가값 예측 방법의 성능을 평가하기 위해 GroupLens Research Center의 MovieLens 평가 데이터를 사용하였다. DEC Systems Research Center는 협력적 여과 알고리즘의 실험을 위해 영화에 대해 18개월동안 추천 서비스를 실시하였다. MovieLens 평가 데이터는 이러한 데이터를 기반으로 하고 있다. MovieLens 데이터 집합은 6040의 사용자들이 3960의 영화에 대해, 총 1000000의 평가를 하였다. 기존의 협력적 여과 연구는 다양한 사용자 집합을 대상으로 실험을 하였다. 예를 들어 [17]의 연구에서는 943명을 대상으로, [6]는 1400명을 대상으로, [4]에서는 5000명을 대상으로 실험을 하였다. 본 논문에서는 1000명의 사용자를 데이터 집합으로부터 무작위로 선택하였

으며, 그 사용자들은 0에서 1까지 0.2의 간격으로 30개의 영화보다 더 많은 영화에 대하여 평가하였다.

1000명의 사용자들은 기본 평가값 예측 방법의 성능을 평가하기 위해 10개의 데이터 집합으로 나눈다. 그 10개의 데이터 집합은 dataset1, dataset2,...,dataset10으로 표기하였다. 사용자의 평가 정보는 각 데이터 집합마다 다르며, 그 분포는 기본 평가값 예측 방법의 정확도에 영향을 미친다. 따라서, 그 데이터 집합을 나눔에 의해 데이터 집합 각각의 분포 정보를 고려하였다. 각 데이터 집합은 훈련 데이터 집합과 테스트 데이터 집합으로 구분하였다. 훈련 데이터 집합은 사용자-아이템 행렬로부터의 실제 데이터 집합이고, 테스트 데이터 집합은 기본 평가값 예측 방법의 정확도를 평가하기 위하여 사용되었다. 사용자 변동 계수를 이용한 기본 평가값 예측 방법(Using_CV)은 사용자 평균을 사용하는 기본 평가값 예측 방법(USER_A), 아이템 평균을 사용하는 기본 평가값 예측 방법(ITEM_A), 선호도 평가값의 중립값을 이용하는 방법(Neutral)과 비교하였다. 그 중립값은 사용자 아이템 행렬에 속한 모든 선호도 평가값들의 평균이다. 부가적으로 본 논문에서는 사용자 변동 계수의 임계값을 0.1로부터 0.9까지 0.1의 간격으로 변동시킴에 의해 기본 평가값 예측 방법의 성능에 대한 민감성을 평가하였다. 표 2는 10개의 테스트 데이터 집합에서 사용자 변동 계수의 임계값을 0.1로부터 0.9로 변화시켰을 때의 기본 평가값 예측 방법의 MAE를 보인다.

표 2에서 이탤릭 문자는 각 데이터 집합에서 가장 낮은 MAE값을 나타낸다. 예를 들어, 첫 번째 데이터 집합에서는 임계값이 0.2일 경우 가장 높은 정확도를 보였다.

표 3은 표 2의 방법에서와 같이 본 논문에서 정의한 수식을 사용하지 않고 실제로 임계값을 변화시켜가면서 가장 높은 정확도를 보인 임계값을 사용하였을 경우와 본 논문에서 정의한 식 (3)과 식 (4)를 사용하여 계산된 임계값을 사용하였을 경우의 MAE를 나타낸다.

표 2 임계값을 변화시킴에 따른 MAE

임계값	1	2	3	4	5	6	7	8	9	10	Mean
0.1	0.2383	0.212	0.1912	0.1827	0.1956	0.1363	0.2264	0.1355	0.193	0.1599	0.1870
0.2	0.2369	0.1586	0.2166	0.1600	0.1956	0.1428	0.1983	0.1496	0.1731	0.1713	0.1802
0.3	0.2369	0.1586	0.2159	0.1549	0.1896	0.1398	0.1879	0.1198	0.1747	0.1708	0.1748
0.4	0.2753	0.1649	0.218	0.1505	0.2146	0.1376	0.1814	0.1334	0.1692	0.1592	0.1804
0.5	0.2826	0.1649	0.2195	0.1499	0.1981	0.1406	0.1762	0.1363	0.1687	0.1579	0.1794
0.6	0.267	0.1671	0.2195	0.1481	0.1981	0.1366	0.1762	0.138	0.1726	0.1410	0.1764
0.7	0.2478	0.1847	0.2195	0.154	0.1954	0.1339	0.1821	0.138	0.1693	0.1448	0.1769
0.8	0.2478	0.2215	0.2195	0.1646	0.1955	0.1338	0.1967	0.138	0.1693	0.1692	0.1855
0.9	0.2478	0.2215	0.2195	0.1646	0.1999	0.1338	0.1967	0.138	0.1693	0.1692	0.1860

표 3 임계값 변화에 의한 방법과 제안된 식에 의한 방법의 성능 비교

	임계값 변화에 의해 결정된 최적의 임계값	R_V(threshold)의 MAE	제안된 방법에 의해 결정된 임계값	Using_CV의 MAE
dataset1	0.2	0.23692381	0.18323	0.236923810
dataset2	0.5	0.164883898	0.52638	0.164789841
dataset3	0.1	0.191155538	0.31465	0.193556000
dataset4	0.6	0.148107331	0.590707	0.148230000
dataset5	0.7	0.19540210	0.52254	0.198135000
dataset6	0.8	0.133830520	0.58926	0.136700000
dataset7	0.6	0.176172957	0.60656	0.176300000
dataset8	0.3	0.119824435	0.38777	0.128848004
dataset9	0.5	0.168709841	0.64610	0.170410000
dataset10	0.6	0.141000690	0.65495	0.143810214
Mean		0.1676		0.1698

그림 5는 표 3을 기반으로 정의된 수식을 이용한 임계값을 사용하였을 경우와 실제로 가장 높은 정확도를 나타낸 임계값을 사용하였을 경우의 MAE를 나타내며, 이 결과를 통하여 정의된 수식의 정확도를 판단할 수 있다.

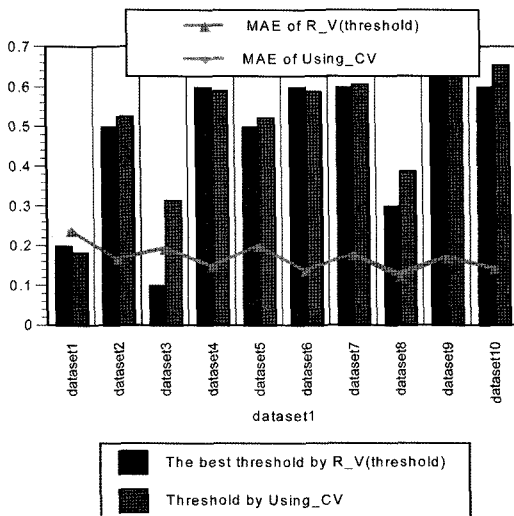


그림 5 Using_CV와 R_V(threshold)와의 MAE 비교

표 3과 그림 5에서 제안된 방법인 Using_CV를 R_V(threshold)의 방법과 비교하였을 경우, 그 평균 오차(ϵ)가 0.00217이다. 만약 평균 오류가 0인 경우, 식 (3)과 식 (4)를 적용한 결과는 실제 실험 결과와 차이가 전혀 없으므로 완벽한 식이라고 할 수 있다. 그러나 실제 데이터는 항상 유동적이기 때문에 오차가 발생한다. 따라서 10개의 데이터 집합에서 실제 실험 결과와 논문에서 정의한 수식을 적용한 결과는 다소의 차이를 보인다. 특히 표 3에서 10개의 집합을 대상으로 하였을 때 그 차의 평균이 0.00217으로 매우 작은 오차를 보였다. 이를 10개의 데이터 집합에 대한 평균 0.1699에 대한 비율로 계산하면, 2.17%의 오차이다. 이러한 결과는 사용자 변동 계수의 임계값이 식 (3)와 식 (4)만으로 결정이 가능함을 보인다. 이와 같이 거의 오차가 없다는 것을 보이기 위하여 표 4와 같이 대응일치 t-검증과 95%의 신뢰도 수준에서의 다중 비교의 본페로니 절차의 에이노바 분석을 사용하여 통계적인 신뢰도를 분석하였다. 표 4의 오른쪽 상단부는 대응일치 t-검증을 이용한 각각의 방법의 신뢰도를 나타내고, 왼쪽 하단부는 에이노바 분석을 통해 각 방법들이 신뢰성이 있는가의 여부를 보인다.

표 4에서 Using_CV와 R_V(threshold)는 MAE의 값

표 4 t-검증과 에이노바를 사용함에 의한 신뢰도 수준

	R_V (threshold)	Using_CV
R_V (threshold)	no	0.115325678
Using_CV	no	no

에 있어서 큰 차이를 나타내지 못하므로, 두 방법이 차이가 있는가에 대한 신뢰도 분석에 있어서 그 차이는 큰 의미가 없다는 결과를 보인다. 따라서, 본 논문에서 정의한 식을 사용한 방법과 실제로 변화시킴에 의해 결정된 가장 높은 정확도를 나타낸 임계값을 사용한 방법과의 차이는 거의 없다는 결과를 얻었으므로 정의된 수식이 오류가 없음을 보인다.

표 5는 10개의 데이터 집합에서 사용자 변동 계수를 이용한 기본 평가값 예측 방법(Using_CV)을 사용자 평균을 사용하는 기본 평가값 예측 방법(USER_A), 아이템 평균을 사용하는 기본 평가값 예측 방법(ITEM_A), 그리고 선호도 평가값의 증립값을 이용하는 방법(Neutral)과 비교하였을 경우 그 성능의 결과를 보인다.

그림 6은 Using_CV, ITEM_A, USER_A, 그리고 Neutral의 성능을 보이고 있으며, 표 5를 기반으로 하였다.

표 5와 그림 6에서 사용자 변동 계수의 임계값을 사용하였을 경우 나머지 다른 방법들보다 보다 우수함을 보인다. 사용자 변동 계수는 USER_A, Neutral, 그리고 ITEM_A의 방법보다 보다 좋은 성능을 보인다. 그리고 USER_A의 방법은 ITEM_A와 Neutral의 방법보다 다소의 차이에서 우세함을 보인다.

표 6은 대응일치 t-검증과 95%의 신뢰도 수준에서의 다중 비교의 본페로니 절차의 에이노바 분석을 사용한 결과의 통계적인 신뢰성을 나타낸다.

표 6의 오른쪽 상단부는 대응일치 t-검증을 이용한 각각의 방법들간의 신뢰도를 나타내고, 왼쪽 하단부는 에이노바 분석을 통해 각 방법들간에 신뢰성이 있는가

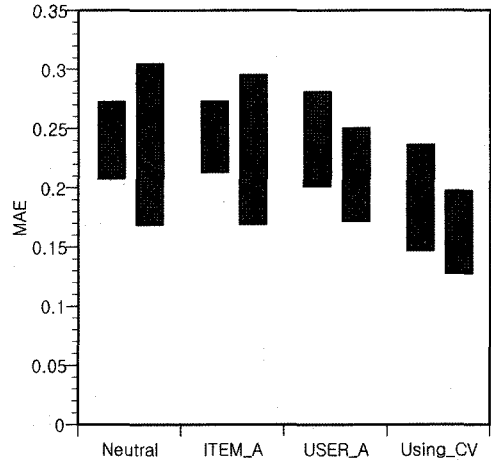


그림 6 기본 평가값 예측 방법들의 성능 비교

표 6 제안된 기본 평가값 방법의 성능 비교의 신뢰도 수준

	Using_CV	USER_A	Neutral	ITEM_A
Using_CV	no	0.000000089	0.000001250	0.000001207
USER_A	yes	no	0.477790526	0.381414232
Neutral	yes	no	no	0.842677588
ITEM_A	yes	no	no	no

의 유무를 보인다. 표 6에서 USER_A, Neutral, 그리고 ITEM_A의 방법들 간에는 성능에 있어서 큰 의미를 보이지 않으므로 그들간에는 그 결과에 있어서 신뢰성이 없다는 결과를 보이나 Using_CV와 나머지 방법들과는 그 비교 결과에 있어서 신뢰성이 있음을 보인다.

반면, 본 논문에서 제안한 방법은 대부분의 협력적 여과 시스템에 사용된 방법과 같이 사용자-아이템 행렬에 나타난 희박성의 정도에 따라 성능에 있어 차이를 보인다. 특히, 희박성이 높을수록 기본 평가값을 계산하는 시간에 있어서 다른 방법들에 비하여 비경제적이다. 따라서 시간을 단축시킬 수 있는 방법에 대한 연구가 필

표 5 Using_CV, ITEM_A, USER_A, 그리고 Neutral의 성능 비교

	Neutral	ITEM_A	USER_A	Using_CV
dataset1	0.27312381	0.273333333	0.281400000	0.236923810
dataset2	0.206428571	0.246957143	0.241731086	0.164789841
dataset3	0.252963605	0.226155538	0.251714035	0.193556000
dataset4	0.208834019	0.214455102	0.202338704	0.148230000
dataset5	0.305363636	0.295912698	0.251000000	0.198135000
dataset6	0.207727273	0.178381818	0.187588550	0.136700000
dataset7	0.223888889	0.243961111	0.207689622	0.176300000
dataset8	0.169928839	0.17046102	0.173023088	0.128848004
dataset9	0.214638889	0.234168443	0.204269841	0.170410000
dataset10	0.200896739	0.193892779	0.208602484	0.143810214
평균	0.226379420	0.227767898	0.220935	0.16977028

요하다. 또한, 희박성의 정도가 매우 극심할 경우 기본 평가값의 정확도는 낮다. 따라서 기본 평가값을 적용하여 추천을 할 것인가 아닌가를 결정할 필요가 있다. 그러나 희박성 정도의 기준을 정하는 것은 용이한 일이 아니다. 따라서 이에 대한 연구 또한 필요하다.

6. 결론 및 향후 과제

협력적 여과 기술은 성공적인 추천 기술이나 사용자가 모든 아이템에 대하여 평가하지 않으므로 인하여 결측치가 발생하고, 그 결과로 희박성의 문제점을 갖는다. 기본 평가값 예측 방법은 이러한 희박성의 문제점을 해결하기 위하여 사용된 방법 중 하나이다. 기본 평가값 예측 방법에 의해 사용자-아이템 행렬의 결측치를 모두 채우고, 이를 대상으로 모델 기반의 여러 방법을 적용하였을 경우, 행렬의 고차원성 문제를 해결할 수 있다. 또한 새로운 사용자에게 보다 정확한 아이템의 추천이 가능하다는 장점을 갖는다. 본 논문에서는 기존의 기본 평가값 예측 방법 중 사용자 평균이나 아이템 평균만을 이용하는 방법의 문제점을 해결하기 위해 사용자 변동 계수를 이용한 기본 평가값 예측 방법을 제안하였다. 사용자들이 아이템에 대하여 평가한 정보를 나타내는 사용자의 변동 계수를 계산하고, 정의된 수식에 의하여 임계값을 결정한 후, 이를 기반으로 기본 평가값을 사용자 평균에서 아이템 평균으로 전환하여 사용자들의 결측치를 채운다. 제안된 방법은 MovieLens 데이터 집합을 이용하여 평가되었으며, 평가 결과 t-검정과 다변 비교 통계에 대한 본페로니 절차의 에이노바 테스트를 통하여 제안된 알고리즘이 기존의 방법보다 우세한 성능을 가짐을 보였다.

향후, 기본 평가 방법이 모델 기반의 협력적 여과 시스템에 사용되어진다면 추천의 성능이 향상될 것으로 기대한다.

참 고 문 헌

[1] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of Dimensionality Reduction in Recommender System A Case Study," Proceedings of ACM WebKDD, 2000.

[2] B. Sarwar, J. Konstan, Al Borchers, J. Herlocker, B. Miller, and J. Riedl, "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System," Proceedings of the 1998 Conference on Computer Supported Cooperative Work, 1998.

[3] David Pennock, Eric Horvitz, Steve Lawrence, and C Lee Giles, "Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-Based Approach," Proceedings of the 16th Con-

ference on Uncertainty in Artificial Intelligence, UAI 2000.

[4] John. S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proceedings of the Conference on Uncertainty in Artificial Intelligence, Madison, WI, 1998.

[5] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl, "An algorithmic framework for performing collaborative filtering," Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999.

[6] Sonny Han Seng Chee, Jiawei Han, Ke Wang, "RecTree: An Efficient Collaborative Filtering Method," Proceedings of the Third International Conference on Data Warehousing and Knowledge Discovery, September, 2001.

[7] 조선호, 김진수, 이정현, "엔트로피와 Default Voting 을 이용한 협력적 필터링에서의 사용자 유사도 측정", 한국정보과학회 추계 학술발표논문집(II), 2001.

[8] Toshihiro Kamishima, "Nantonac collaborative filtering: recommendation based on order responses," Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[9] MovieLens collaborative filtering data set, [Http://www.cs.umn.edu/Research/GroupLens/index.html](http://www.cs.umn.edu/Research/GroupLens/index.html), GROUPLENS RESEARCH PROJECT, 2000.

[10] V. Barnett and T. Lewis, Outliers in Statistical Data. John Wiley and Sons, 1994.

[11] L. Grossi, G. Gozzi, and P. Ganugi, "Distribution Analysis of Items and Ratios in Companies? Accounts using a new iterative procedure," Proceedings of compstat2002, 2002.

[12] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, "Evaluating collaborative filtering recommender systems," ACM Transactions on Information Systems (TOIS) archive, Vol. 22, No. 1, 2004.

[13] J. Delgado and N. Ishii, "Formal Models for Learning of User Preferences, a Preliminary Report," In Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-99), Stockholm, Sweden, July, 1999.

[14] Paul R. Cohen. Empirical methods for artificial intelligence. The MIT Press, 1995.

[15] Giuseppe Carenini, Rita Sharma, "Exploring More Realistic Evaluation Measures for Collaborative Filtering," AAAI 2004, 2004.

[16] Rita Sharma and David Poole, "Symmetric Collaborative Filtering Using the Noisy Sensor Model," Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, 2001.

[17] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Analysis of recommendation algorithms for

e-commerce," Proceedings of the 2nd ACM conference on Electronic commerce, 2000.

고 수 정

정보과학회논문지 : 소프트웨어 및 응용
제 32 권 제 7 호 참조