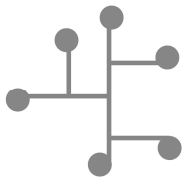


평점 정규화를 이용하여 사용자 평가 성향을 반영한 영화 추천 방법

Movie Recommendation Method Using Score Normalization Based on User Rating Tendency

김현경(Hyunkyung Kim)[1] 김현진(Hyunjin Kim)[2] 박상현(Sanghyun Park)[3]

요 약

기존의 사용자 기반 추천 방법을 이용한 영화 추천 시스템에서는 대개 다른 사용자들의 평점을 기반으로 목표 사용자의 평점을 예측하는 데에 이용하였지만 사용자 개개인의 평가 성향은 반영하지 않아 평점 데이터의 객관성을 확보하기에는 어려운 점이 있었다. 본 논문에서는 기존의 사용자 기반 추천 방법과 항목 기반 추천 방법을 바탕으로 한 항목 간 선호도 차이를 이용한 추천 방법을 토대로 사용자 개인의 평가 성향을 반영한 새로운 영화 추천 시스템을 제안한다. 많은 사용자들의 평점 데이터가 사용자의 성향에 따라 치우쳐 있어 다른 사용자의 평점 예측에 이용되기에는 다소 어려운 점이 있었다. 따라서 사용자들의 평가 성향을 바탕으로 데이터를 정규화하였고 항목 간 선호도 차이를 이용하여 평점을 예측하는 시스템을 구현하였다. 실험 결과 제안한 시스템은 기존의 시스템에 비해 추천의 정확도가 향상되었다. 따라서 본 연구의 제안 방법은 사용자의 평점 결정 성향을 반영함으로써 다양한 콘텐츠에 대한 사용자의 평가를 보다 정확하게 예측하여 사용자 개인에 맞는 영화 추천을 가능하게 할 것으로 기대된다.

주제어: 추천 시스템, 데이터 마이닝, 협업 필터링, 사용자 성향, 평점 정규화

1 연세대학교 컴퓨터과학과, 학부생.
2 연세대학교 컴퓨터과학과, 박사과정.
3 연세대학교 컴퓨터과학과, 교수, 교신저자. (sanghyun@yonsei.ac.kr)
+ 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (NRF-2015R1A2A1A05001845)
+ 논문접수: 2016년 3월 27일, 심사완료: 2016년 5월 17일, 게재승인: 2016년 7월 1일.

Abstract

Existing movie recommender systems generally use rating data of other users to predict the rating of target user. However, it is hardly possible to guarantee the objectivity of rating data since the rating tendency of individual user is not considered. In this paper we propose a new recommendation method which took into account rating tendency of each individual user using score normalization based on rating difference between items. We have found many users have biased rating tendency and their rating data was affected. So we have normalized those rating data to get better prediction results. The results of experiments indicate that the proposed system has relatively improved performance compared to the previous recommender system in terms of prediction accuracy. Consequently, the proposed system is expected to enable improved movie recommendation for each individual by weighing rating tendency using score normalization.

Keywords: Recommender System, Data Mining, Collaborative Filtering, Rating Tendency, Score Normalization

1. 서 론

최근 최근 스마트기기가 대중화로 시간과 공간의 제약 없이 멀티미디어 콘텐츠를 재생할 수 있게 되면서 영상 콘텐츠에 대한 수요 역시 늘고 있다. 이러한 수요에 부응하여 실시간으로 막대한 양의 영상 콘텐츠가 공급되고 있지만 사용자들은 수많은 콘텐츠들 사이에서 자신에게 맞는 콘텐츠를 찾는 데 어려움을 겪고 있다 [1].

많은 사용자들은 자신에게 적합한 콘텐츠를 찾기 위해 영화 콘텐츠에 대한 정보를 제공해주는 TV 프로그램이나 온라인 커뮤니티를 이용하고 있다. 하지만 이러한 방식은 사용자의 개인적인 취향을 반영하기 어려우며, 추천의 범위가 제한적이어서 추천의 질이 떨어진다는 단점이 있기 때문에 사용자 개인에 알맞은 영화 추천을 하기에는 어려움이 있다 [2].

이와 같은 문제를 해결하기 위해 Netflix [3] 와 IMDb[4] 등에서는 사용자들에게 자동화된 추천시스템을 제공하였고, 우리나라에서도 Watcha [5] 등의 사이트가 등장하면서 영화 추천 시스템에 대한 연구가 활발히 진행되고 있다.

영화 추천 시스템에서는 크게 협업 필터링과 항목 기반 추천의 두 가지 방식을 사용하고 있다. 협업 필터링의 경우 다른 사용자들의 선호도 정보에 기반하여 항목을 추천하는 방식으로 사용자의 취향을 반영할 수 있다 [6]. 이와는 달리 항목 기반 추천은 항목 간의 유사성을 기반으로 사용자가 좋아할 만한 항목을 추천한다. 추천의 정확도가 높은 협업 필터링 방식은 항목 기반 추천 방식에 비해 널리 사용되고 있다 [7]. 그러나 이 방식은 데이터가 충분하지 않은 경우 사용자 간의 유사도 계산의 신뢰도를 확보하기 어렵다. 이런 데이터 희박성 문제를 보완하기 위해 나이, 성별 등의 인구통계학적 정보를 사용하기도 한다[8-9].

위의 항목 기반 추천 방식과 협업 필터링 방식을 혼합한 방식에 대한 연구도 많이 이루어지고 있다. 논문 Oh et al. [10] 에서는 사용자들의 항목 선호도 정보를 이용하여 항목 간 평점 차의 평균을 구하고, 이를 바탕으로 새로운 항목에 대한 목표 사용자의 평점을 예측하는 추천 방법을 제안하였다. 그런데 사용자들은 평가 성향이 서로 다르기 때문에 평점을 높게 주는 경향이 있는 사용자와 평점을 낮게 주는 경향이 있는 사용자의 점수가 같을 때에도 이 점수가 같은 의미를 갖는다고 보기 어렵다. 이는 사용자 개개인의 서로 다른 평가 성향을 반영하지 않은 데이터를 이용해서 다른 목표 사용자의 평점을 예측하기에는 어려움이 따른다는 것을 의미한다.

따라서 본 논문에서는 항목 간 선호도 차이를 이용한 영화 추천 방법 [7] 을 기반으로 사용자의 평가 성향을 반영한 평점 정규화를 통해 다른 사용자들의 평점 데이터의 객관성을 확보하여 평점 예측의 정확도를 높인 추천 방법을 제안하고자 한다.

2. 관련 연구

2.1 추천 시스템

추천 시스템은 고객이 관심을 가지고 있는 상품에 대한 정보나 고객 개인의 정보를 분석하여 고객의 요구에 맞는 항목을 추천해주는 시스템이다 [11]. 전자상거래나 다양하고 방대한 양의 콘텐츠가 있는 웹 사이트에서 콘텐츠 추천은 사용자가 자신이 원하는 상품을 찾고 구매하도록 돕는 데에 중요한 역할을 한다 [12].

2.2 개인화된 추천 기법

사용자들 개개인에 적합한 추천을 돕기 위해 개인

성향과 협업 필터링에 관련된 많은 추천 기법들이 연구되었으며 [13] 개인화된 추천 기법에는 다음과 같은 방법들이 있다.

1) 항목 기반 추천 기법(Contents based recommendation)

항목 기반 추천 기법은 사용자의 평점이 높은 항목과 유사한 새로운 항목을 추천한다 [14]. 이 추천 방식은 데이터가 작은 경우에도 새로운 항목을 추천하는 데에 어려움이 없다는 장점이 있다. 이 방법은 항목의 특성이 유사하다면 그에 대한 사용자의 선호도 역시 유사할 것이라는 가정을 전제로 항목을 추천한다. 하지만 이러한 추천 방식은 실제 사용자의 선호도를 전혀 반영하지 못한다는 단점이 있다.

2) 협업 필터링(Collaborative Filtering)

협업 필터링은 사용자 사이의 유사도를 계산하여 목표 사용자와 유사도가 높은 사용자들이 높은 점수를 준 항목들을 추천하는 방식으로 [15], 추천의 정확도가 비교적 높기 때문에 항목 기반 추천 방식 보다 널리 사용되고 있다. 그러나 이 방식은 영화에 대한 사용자들의 평점 정보를 이용하여 유사도를 계산하고 그에 따라 추천이 진행되기 때문에 사용자의 평점 데이터가 부족한 경우에는 사용자간 유사도 계산의 신뢰도가 떨어진다는 한계점이 존재한다 [16].

3) 인구 통계학적 추천(Demographic Recommendation)

나이, 성별 등의 사용자 정보를 바탕으로 특정 나이 대 혹은 성별 집단의 성향 분석을 통해 추천을 수행하는 시스템이다 [17]. Recio-Garcia et al. [18]에서는 협업 필터링을 기반으로 집단 단위의 성격을 고려하여 추천하는 시스템을 제안한 바 있다. 인구통계학적 속성은 다양한 항목 중에서 특정 집단에 속하는 사용자들의 성향을 분석하기 쉽다는 장점이 있다.

4) 오피니언 마이닝을 이용한 추천 방법(Opinion Mining)

Tripadvisor[19], Expedia[20], Amazon[21]과 같은 전자상거래 사이트들에서 주로 이용하는 방법으로 사용자의 리뷰나 코멘트를 수집, 분석하여 사용자들에게 개인화된 맞춤형 서비스를 제공하고자 하는 추천 방법이다. 자연어 처리를 통해 사용자의 리뷰나 코멘트로부터 사용자의 선호도를 분석하기 쉽다는 장점이 있다 [22-24].

5) 혼합형 추천 방법(Hybrid Recommendation)

항목 기반 추천 기법과 협업 필터링 기법을 혼합한 방식에 대한 연구로 두 가지 방식의 장점을 결합하여 기존의 두 가지 방법보다 추천의 정확도가 향상되었다 [25].

6) 항목 간 선호도 차이를 이용한 영화 추천 방법

항목 간 평점 차이의 평균을 계산하고, 이를 바탕으로 목표 사용자의 평점을 예측한다 [10]. 항목 간의 평점 차를 이용한다는 점에서 항목 기반 추천 기법과 유사하며 항목 간의 평점 차이는 같은 항목에 평점을 준 사용자들 사이의 평점 유사도에 기반한다는 점에서 협업 필터링과도 유사한 점을 갖는다. 이 방법 또한 혼합형 추천 방법처럼 두 방식을 혼합하여 향상된 성능을 보인다.

3. 제안하는 방법

본 논문에서는 논문 Oh et al. [10]에서 제안한 방법을 기반으로 사용자의 평점 결정 성향을 반영하여 새로운 평점 예측 방법을 제안하고자 한다.

3.1 항목 간 선호도 차이를 이용한 영화 추천 방법

Oh *et al.* [10]에서는 사용자들의 평점 정보를 바탕으로 항목 간 평균 평점 차이를 구하고, 이를 이용

해서 새로운 항목에 대한 목표 사용자의 평점을 예측한다.

1) 항목 간 선호도의 차이 계산

이 방식에서는 먼저 사용자들의 항목 평가 점수를 바탕으로 항목 간 평점 차이의 평균을 구한다. 두 항목 i와 j에 대한 사용자들의 선호도 차이의 평균 $d_{i,j}$는 다음 식을 이용하여 구할 수 있다.

$$d_{i,j} = \frac{\sum_{a \in U_i \cap U_j}(r_{a,i} - r_{a,j})}{|U_i \cap U_j|} \quad (1)$$

위 식은 U_i, U_j로 표현된 주어진 두 항목을 모두 평가한 사용자들의 두 항목 평점이 얼만큼 차이가 나는지를 $r_{a,i} - r_{a,j}$로 나타낸다.

2) 선호도 예측

식 (1)에서 계산한 평점의 평균 차이를 이용하여 새로운 항목 i에 대한 사용자 u의 평점 $\hat{r}_{u,i}$를 구하는데, 이는 다음 식을 이용하여 계산한다.

$$\hat{r}_{u,i} = \frac{\sum_{j \in I_u}(r_{u,j} + d_{i,j})}{|I_u|} \quad (2)$$

위의 식에서 항목 j를 이용해 새로운 항목 i에 대한 사용자 u의 평점 예측값을 구할 수 있는데, 이는 사용자 u가 평점을 매긴 항목 j에 대한 선호도 $r_{u,j}$에 항목 i와 항목 j의 평균 선호도 차이 $d_{i,j}$를 더해서 구한다. 이렇게 구해진 각 항목 j를 이용한 예측값들을 단순 평균하여 $\hat{r}_{u,i}$를 계산하는데, 이는 각각의 예측값들에 대한 중요도를 모두 상수 1로 간주한 경우이다.

3) 가중치로 항목의 중요도를 표현

위의 경우처럼 각각의 예측값들의 중요도를 1로 적용한 것에 비해서 $d_{i,j}$를 구하는 데 사용한 데이터의 개수, 즉 i 와 j 항목 모두를 평가한 사용자의 수 $|U_i \cap U_j|$를 항목 j의 가중치로 적용하여 위 식 (2)에 곱해준 경우 각 항목 j의 상대적인 중요도를 반영할 수 있다. 가중 평균을 적용한 식은 다음과 같다.

$$\hat{r}_{u,i} = \frac{\sum_{j \in I_u}\{(r_{u,j} + d_{i,j}) \cdot |U_i \cap U_j|\}}{\sum_{j \in I_u}|U_i \cap U_j|} \quad (3)$$

3.2 사용자의 평점 결정 성향을 이용한 추천 방법

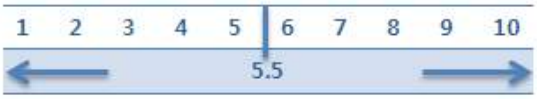
위의 항목 간 선호도 차이를 이용한 추천 방법에 사용자의 평점 결정 성향을 반영하여 평점 예측의 정확도를 높인다.

1) 사용자의 평가 성향을 반영한 평점 정규화

사용자가 콘텐츠의 평점을 결정하는 방식은 사용자 개인의 성향에 따라 조금씩 차이가 있다. 가령 영화를 평가할 때 주제, 극중 인물, 내용 전개 방식, 재미, 영상미 5가지를 기준으로 보는 사용자 u1과 u2가 있을 때, u1라는 사용자는 영화 평가 기준 중 재미만 충족되면 10점 만점 중 무조건 10점을 주는 반면에, u2라는 사용자는 5가지 기준 중 한 가지만 미달되더라도 6점 미만의 점수를 준다고 가정한다. 이처럼 평가 성향이 서로 다른 경우 이들의 평점 데이터를 다른 사용자의 평점을 예측하는 데 사용하려면 보다 객관성이 있는 데이터로 만들어주는 과정이 필요하다. 이러한 평점 데이터를 사용자의 평가 성향에 비추어 정규화 해준다면 또 다른 사용자 u3에게 보다 정확한 추천을 해줄 수 있게 된다. 평점 정규화는 사용자들의 평점 데이터가 중앙값으로부터 양쪽으로 분포되어 있는 정도를 조정해주는 과정으로, 정규화 과정은 다음과 같이 진행된다.

A. 중앙값 5.5 기준 정규화

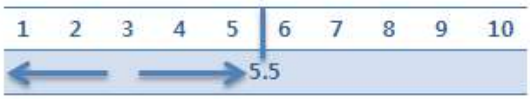
사용자가 각 항목에 대해 1부터 10까지의 점수를



[그림 1] 중앙값을 기준으로 한 평점 정규화

매길 수 있을 때, 중앙값인 5.5점을 중심으로 사용자의 점수 최대값과 최소값이 각각 10점, 1점이 되도록 분포를 조정한다. 예컨대 사용자 u1의 평점 최대값이 8.0일 때 u1의 점수 7.0을 정규화한다면 중앙값 5.5를 중심으로 오른쪽에 위치한 점수 7을 $5.5+\left\{(7-5.5)\times\frac{10-5.5}{8-5.5}\right\}$를 계산하여 만점이 10점일 때의 점수로 바꾸어주는 것이다. 최소값의 경우도 같은 방식으로 적용된다. 평점의 최소값이 2.0이라면 중앙값 5.5를 넘지 않는 점수 3.0을 $5.5-\left\{(5.5-3)\times\frac{5.5-1}{5.5-2}\right\}$으로 조정해주어 사용자의 평점을 정규화한다.

B. 평점 범위 최소값과 중앙값 사이의 데이터 정규화

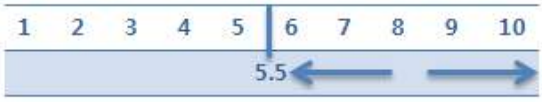


[그림 2] 사용자 평점 범위가 최소값–중앙값 사이인 경우

사용자의 평점 범위가 최소값과 중앙값 사이에만 위치하는 경우 사용자 평점의 최대값은 전체 평점 범위의 중앙값, 사용자 평점의 최소값은 전체 평점 범위의 최소값이 되도록 사용자 평점 범위의 중앙값을 전체 평점 범위에서 최소값부터 중앙값에 해당하는 범위의 중앙값, 즉 최소값 1점부터 중앙값 5.5점 사이의 중앙값인 3.25점에 맞추어서 정규화한다.

C. 평점 범위 중앙값과 최대값 사이의 데이터 정규화

위의 B 경우와 마찬가지로 사용자의 평점 범위가 국소적인 경우로, 중앙값과 최대값 사이의 평점 데이터만 존재하는 사용자의 경우 사용자 평점의 최대값은 전체 평점 범위의 최대값, 사용자 평점의 최소값은 전체 평점 범위의 중앙값이 되도록 사용자 평점 범위의 중앙값을 기준으로 정규화한다.



[그림 3] 사용자 평점 범위가 중앙값–최대값 사이인 경우

위의 평점 정규화를 표로 나타내면 다음과 같다.

[표 1] 평점 분포에 따른 평점 정규화 방법

유저의 평점 분포	정규화 방법
평점 최고점>5.5점 평점 최하점<5.5점	중앙값 5.5점 기준 정규화
평점 최고점<5.5점	평점 범위 최소값과 중앙값 사이의 데이터 정규화
평점 최하점>5.5점	평점 범위 중앙값과 최대값 사이의 데이터 정규화

2) 항목 간 선호도 차이를 이용한 방법에의 적용

논문 Oh et al. [10]에서 제안한 항목 간 선호도 차이를 이용한 추천 방법을 기반으로 사용자 평점 결정 성향에 따른 평점 정규화를 적용한다. 정규화 과정을 거친 평점 데이터는 Oh et al. [10]의 항목 간 선호도 차이 식 (1)에 적용한다.

$$d_{i,j}=\frac{\sum_{a\in U_i\cap U_j}(r_{a,i}-r_{a,j})}{|U_i\cap U_j|} \quad (1)$$

위 식에서 항목 i와 j에 대한 평점 데이터가 있는 사용자들 각각의 항목 i와 j의 평점 차를 합산하는데 이 때 와 의 값에 사용자 개인의 평점 결정 성향에 비추어 평점 정규화를 적용시킨 데이터는 목표 사용자의 평점 예측함에 있어 보다 객관적인 지표로써 사용할 수 있다.

4. 실험 결과

본 논문에서 제안한 방법이 실제로 사용자 평점 예측의 정확도를 향상시키는지 알아보기 위하여 실험을 진행하였다.

4.1 실험 환경

본 연구에서 사용한 실험 데이터는 MovieLens 100K dataset [26] 으로 사용자의 수가 706명, 항목 수 즉 영화 수가 8,570편, 선호도 평가 데이터의 수가 100,023개이다. 선호도 평가 데이터는 튜플(사용자 ID, 항목 ID, 선호도)로 구성되어 있으며, 선호도 평가는 0.5점부터 5점까지 0.5점 단위로 이루어져 있다. MovieLens에서 제공하는 데이터셋의 일부는 다음과 같다.

userId	movieId	rating
1	6	2
1	22	3
1	32	2
1	50	5
1	110	4
1	164	3
1	198	3
1	260	5
1	296	4
1	303	3

[그림 4] MovieLens dataset의 일부분

실험에서는 기존 연구 [10]에서 사용한 항목간 선호도 차이를 이용한 방법만을 적용시켰을 때에 비해 사용자의 평점 결정 성향을 반영하여 평점을 정규화한 경우 정확도가 얼마나 향상되는지 알아보았다.

실험은 MovieLens dataset에 존재하는 선호도 평가 데이터 100,023개에 대해 제안 방법을 적용하여 선호도를 예측하고 원래 데이터와 비교하는 방식으로 진행하였다.

4.2 선호도 예측 성능의 평가

본 연구에서 사용한 성능 측정 지표는 MAE(Mean Absolute Error)로 다음 식과 같이 정의된다.

$$d_{i,j} = \frac{\sum_{i=1}^{N} |r_{u,i} - \hat{r}_{u,i}|}{N} \quad (4)$$

즉, 주어진 선호도 평가 데이터에 들어있는 선호도 $r_{u,i}$와 식 (2)에 의해 구해진 선호도 $\hat{r}_{u,i}$의 차이를 누적하고 데이터의 수로 나누어 평균을 구한다. MAE(Mean Absolute Error)는 오차의 평균을 구하는 방식으로 오차를 구할 때 일반적으로 많이 사용되는 통계 지표이다.

본 논문에서 제안하는 방법과 Random approach 및 Oh *et al.* [10]에서 사용한 방법을 비교한 결과는 다음 표와 같다.

[표 2] 다른 추천 시스템과의 성능 비교

	Cumulative Error	Mean Absolute Error
Random approach	157938.50000	1.579038
Oh *et al.*[10]	50487.19640	0.504755
Ours	50117.05365	0.501055

MAE(Mean Absolute Error)의 경우 목표 사용자의 평점을 랜덤 함수를 이용하여 예측하는 Random approach에 비해 오차가 1.078점 줄어들었고 이것은 Random approach 오차의 68%가 감소한 값이다. Oh *et al.* [10]에서 제안한 방법을 이용하였을 때보다는 오차가 약 0.0037점 줄어들었다. 누적 오차는 제안한 방법을 적용시킨 경우 Oh *et al.* [10]의 방법과 비교했을 때 총 100,023개 데이터에 대해 오차가 약 370점 낮게 나타났다.

이는 Oh *et al.* [10]에서 제안한 항목간 선호도 차이만을 이용하여 평점 예측을 한 경우에 비해서 본 논문에서 제안한 방법과 같이 사용자의 평점 결정 성향에 따라 데이터를 분류하여 각각의 사용자의 성향에 알맞게 평점 정규화를 적용시킨 후 평점 예측을 했을 때 그 평점 예측의 정확도가 향상되었음을 나타낸다. 즉 사용자의 평점 결정 성향에 따른 평점 정규화를 적용시키면 목표 사용자의 평점을 보다 정확하게 예측할 수 있게 된다.

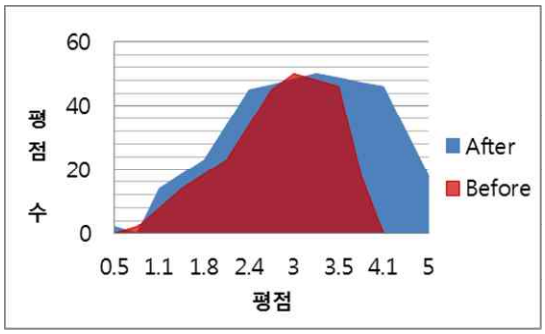
4.3 Case study

A. 평점 정규화의 영향을 크게 받은 유저 데이터

먼저 본 논문에서 제안한 사용자 평점 결정 성향에 따른 평점 정규화를 적용하였을 때 영향을 많이 받을 사용자들을 알아보기 위해 MovieLens dataset의 706명의 사용자 중 먼저 평점의 최소값과 최대값의 범위가 가장 좁은 상위 10명의 사용자들을 추출하였다. 그리고 그 10명의 사용자 중에서도 본 논문에서 제안한 방법을 통해 정규화 과정을 거쳤을 때 기존의 평점에 비해 점수에 가장 많은 변화가 있었던 사용자 3명에 대해서 Case Study를 진행하였다.

1) 598번 사용자

598번 사용자는 제안한 방법에 따라 정규화 과정을 거쳤을 때 평점 분포의 변화가 다음과 같이 변하였다.



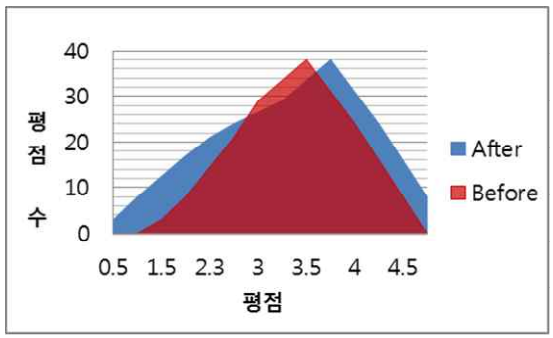
[그림 5] 598번 유저의 평점 정규화

598번 사용자는 198개의 평점 데이터를 보유한 사용자로 평점의 최고점은 4점, 최하점은 1점이었다. 이 사용자는 198편의 영화에 대해 평점을 부여했음에도 평점의 최고점은 5점이 아니라 4점이었다. 이런 경우 이 사용자의 평점 최고점인 4점을 만점으로 보고 제안 방법에 따라 정규화 해주었을 때 기존의 평점 데이터(붉은색 그래프)보다 완만해진 데이터(파란색 그래프)를 볼 수 있다.

598번 사용자의 평점 데이터를 본 연구의 제안 방법에 따라 정규화하였을 때 198개의 평점 데이터 중 84.8점이 변하였다. 정규화 과정을 통해 객관성을 확보한 평점 데이터가 목표 사용자의 평점 예측에 사용되어 성능을 향상시킬 수 있었음을 확인할 수 있다.

2) 304번 사용자

304번 유저의 평점 분포는 제안한 방법에서의 정규화 여부에 따라 다음과 같이 변하였다.

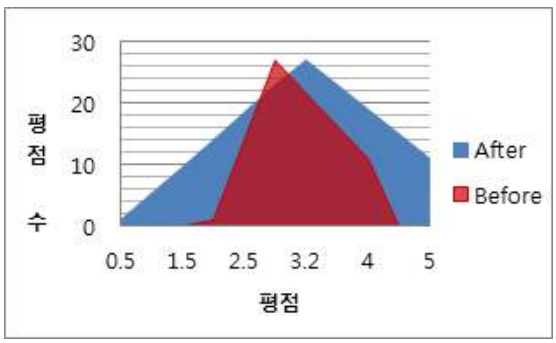


[그림 6] 304번 유저의 평점 정규화

304번 유저는 131개의 평점 데이터를 가진 사용자로 평점의 최고점은 4.5점, 최하점은 1.5점으로 다른 유저들과 비교했을 때 최하점이 높은 편인 사용자이다. 이 사용자의 경우도 마찬가지로 평점 정규화 과정을 거친 후의 데이터가 조금 더 완만해졌음을 알 수 있다. 이렇게 정규화하였을 때 304번 유저의 131개 데이터 중 47점이 변하였다. 정규화 과정에 따라 평점 데이터가 영향을 많이 받았음을 알 수 있다.

3) 397번 사용자

397번 사용자의 평점 분포 변화는 다음과 같다.



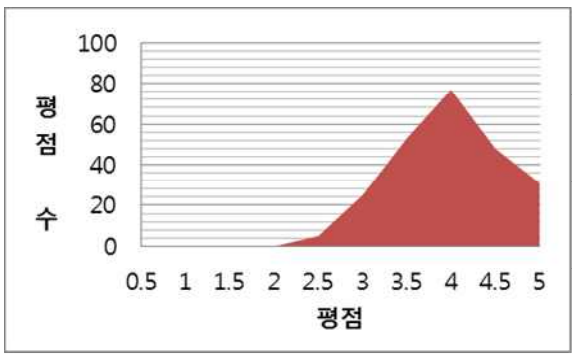
[그림 7] 397번 유저의 평점 정규화

397번 사용자의 경우 39개의 평점 데이터 중 최고점은 4점 최하점은 2점으로 데이터의 중앙값인 3점이 최빈값이었다. 이 사용자는 데이터가 중앙에 몰려있는 경향을 보였는데 평점 정보를 정규화하였을 때 39개의 평점 데이터 중에서 17.9점이 변하였다. 위의 그래프에서 치우쳐져 있던 평점 분포가 정규화 이후 매우 완만해졌음을 알 수 있다.

세 번째 사용자의 경우 앞선 두 명의 사용자에 비해 그래프의 너비가 다소 작은 것을 볼 수 있는데 이는 사용자가 평가한 평점 데이터의 개수가 1번과 2번에 비해 적기 때문이다. 사용자의 평점 데이터의 양이 많을수록 정규화 과정 이후 전체 평점 변화에 미치는 영향이 큰 것은 당연한 사실이다. 이런 관점에서 볼 때 세 명의 사용자 중 비교적 평점 데이터 수가 많은 598번과 304번 사용자가 데이터 수가 적은 편인 397번 사용자보다 정규화 과정에 따른 전체 평점 변화에 더 많은 기여를 한다는 사실을 알 수 있다.

B. 평점 데이터의 편향성이 큰 유저들의 데이터 비교

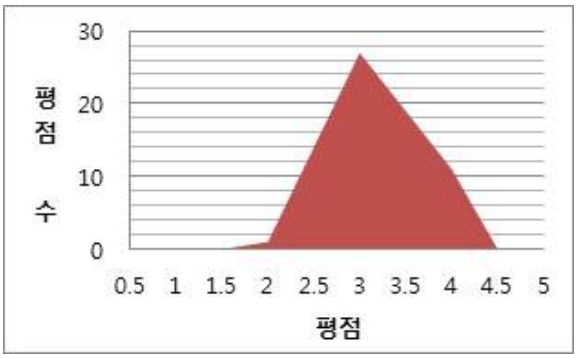
정규화 과정이 어떻게 사용자들의 평점 데이터에 객관성을 부여할 수 있는지 잘 보여주는 두 명의 유저 평점 데이터가 있다.



[그림 8] 44번 유저의 평점 분포

먼저 44번 유저의 경우에는 평점을 대체로 후하게 주어서 평점 데이터가 높은 점수에 치우쳐 있다. 즉 높은 점수를 주는 경향이 강한 사용자이다.

반면에 397번 유저는 0.5점부터 5점까지의 평점 범위 중 최고점은 4점, 최하점은 2점으로 데이터가 중앙값에 몰려서 분포하는 경우이다.



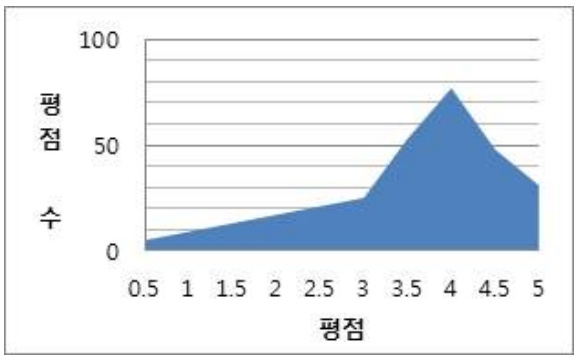
[그림 9] 397번 유저의 평점 분포

이런 경우 44번과 397번 두 유저의 3점이 과연 같은 의미의 3점이라고 볼 수 있을까? 그렇게 보기 어려울 것이다. 44번 유저는 239개의 평점 데이터를 보유한 사용자이다. 데이터 개수가 충분히 많음에도 불구하고 평점 데이터가 치우친 정도가 강한 경향을 보이는 것으로 볼 때 44번 사용자에게 있어 3점은 상당히 낮은 점수임이 자명하다. 반면에 397번 유저의 경우에는 3점의 의미가 중간의 느낌이 강하다. 39개의 평점 데이터 중 27개의 데이터가 3점에 분포하고 있음을 그래프를 통해 알 수 있다.

이렇게 두 사용자의 평점 분포가 다를 때 두 사용자의 서로 다른 평가 성향을 고려하지 않고서는 두 데이터를 비교하기가 어렵기 마련이다. 따라서 정규화 과정을 통해 사용자들의 평가 성향을 고려하여 사용자의 데이터에 객관성을 부여할 필요가 있다.

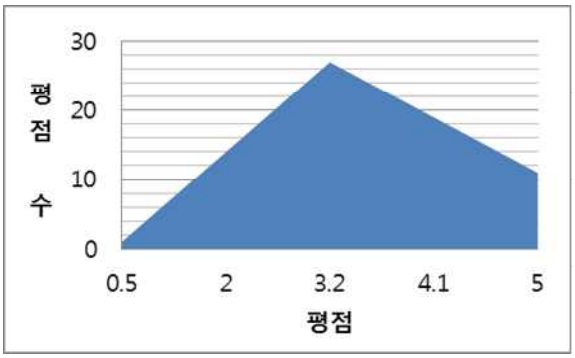
이들 사용자들의 평점 데이터를 정규화한 경우 평점 분포는 아래와 같다.

44번 유저의 경우 최하점인 2.5점을 평점 범위의 최하점인 0.5점에 맞추어 평점 범위의 중앙값 2.75



[그림 10] 44번 유저의 정규화된 평점 분포

점 이하의 점수를 정규화하였다. 44번 사용자에게 있어 낮은 점수에 해당하는 점수들을 사용자의 평가 성향을 반영하여 변환한 것이다.



[그림 11] 397번 유저의 정규화된 평점 분포

397번 유저의 경우에는 0.5-5점의 평점 범위 중 유저의 평점이 2-4점에 몰려 있어 치우친 정도를 보정하였다. 중앙값을 기준으로 최하점과 최고점을 각각 0.5점과 5점으로 조정하여 변환한 결과이다.

개인의 성향에 따른 평점 결정 특성을 고려하지 않을 경우 평점 분포가 그림 8과 9처럼 다소 치우친 경향을 보인다. 어떤 집단은 대부분의 영화에 대해 높은 평점을 주는 반면에, 또 다른 집단은 주로 낮은 점수로 영화를 평가하는 경향을 가질 수 있기 때문이다. 이런 경우 본 논문에서 제안하는 정규화 과정을 거칠 경우 평점의 편향성 문제가 해소될 수 있으며, 그림 10, 11과 같이 정규화 전에 비해 완만한 평점 그래프를 보이게 된다.

위의 사용자들과 같이 평점 범위가 고르게 분포되어 있지 않고 사용자의 평점 결정 성향에 따라 한쪽으로 치우친 경향을 보이는 경우에는 본 연구의 제안 방법에 따라 정규화 과정을 적용시키면 해당 사용자들의 데이터의 객관성을 확보하여 목표 사용자의 평점을 예측할 때 정확도가 높은 결과를 얻을 수 있다.

5. 결 론

본 논문에서는 항목 간 선호도 차이 정보를 바탕으로 사용자의 평점 결정 성향을 반영하여 사용자의 선호도를 예측하는 영화 추천 방법을 제안하였다. 사용자들이 영화에 대해 평가 점수를 결정할 때 그 척도가 각 개인마다 다양하므로 평가 점수의 분산 정도가 차이가 나게 된다. 이에 따라 사용자들이 정한 평점을 정규화하여 보다 객관적인 지표로 사용할 필요가 있다.

제안한 방법에서는 사용자들의 평점 결정 성향에 따라 평점을 정규화하였고, 항목 간 선호도 차이 정보에 정규화된 평점을 적용하여 목표 사용자의 평점을 예측하였다. 그 결과 정규화를 적용시킨 경우 평점 예측의 오차가 감소하였다.

제안한 방법은 데이터 수가 적은 초기에는 희박성 문제에서 자유롭지 못하다는 문제점이 있다. 초기에 사용자 데이터가 충분치 않은 경우 사용자 정보를 이용한 네트워크를 생성하여 평점을 Network Propagation의 방법으로 예측하게 하는 등 향후 부족한 데이터를 채우는 방법에 대한 연구가 진행된다면 희박성 문제를 개선하여 성능을 향상시킬 수 있을 것으로 기대된다.

참고 문헌

[1] George Lekakos, Petros, “A hybrid approach for movie recommendation,” Multimedia Tools and Applications, vol. 36, Issue 1-2, pp. 55-70, 2008.

[2] Boo-Sung Kim, Heera Kim, Jaedong Lee, Jee-Hyong Lee, “Movie Recommendation System Using Collaborative Filtering Based on Demographic Information,” Proceedings of KIIS Fall Conference, vol. 23, pp. 63-64, 2013.

[3] Netflix. https://www.netfilx.com

[4] IMDb. https://www.imdb.com

[5] Watcha. https://watcha.net

[6] Hee-Choon Lee, Seok-Jun Lee, Sun-Ok Kim, "A Study on improvements of prediction accuracy using additional information in collaborative filtering," Proceeding of The KITS Conference 2009, pp. 349-352, 2009.

[7] P. Melville, R. J. Mooney and R. Nagarajan, “Content-Boosted Collaborative Filtering,” Proceedings of the SIGIR-2001 Workshop on Recommender Systems, 2001.

[8] G.Lekakos and G.M.Giaglis, "Improving the Prediction Accuracy of Recommendation Algorithms : Approaches Anchored on Human Factors," Interacting with Computers, vol. 18, pp. 410-431. 2006.

[9] M. J. Pazzani, “A framework for Collaborative, Content-based and Demographic Filtering,” Artificial Intelligence Review, vol. 13, pp. 393-408, 1999.

[10] Se-Chang Oh, Min Choi, “A Movie Recommendation Method Using Rating Difference Between Items,” Journal of Korea Institute of Information and Communication Engineering, vol. 17, No. 11, pp. 2602-2608, 2013.

[11] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, “Item based Collaborative Filtering Recommendarion Algorithms,” Proceedings of the 10th International World Wide Web Conference, pp. 285-295, 2001.

[12] H. Ji, J. Li, C. Ren, and M. He, "Hybrid collaborative filtering model for improved recommendation," Service Operations and Logistics, and Informatics (SOLI), 2013 IEEE International Conference, pp. 142-145, 2013.

[13] G. Guo, "Integrating Trust and similarity to Ameliorate the Data Sparsity and Cold Start for Recommender Systems," RecSys ’13 Proceedings of the 7th ACM conference on Recommender Systems, pp. 451-454, 2013.

[14] S. H. Jo, “Weight Recommendation Technique Based on Item Quality to Improve Performance of New User Recommendation on The Web,” Ph.D. dissertation, Hannam University Graduation School, 2008.

[15] S. J. Lee and T. R. Jeon, G. D, Baek, S. S. Kim, “A Movie Rating Prediction System of User Propensity Analysis based on Collaborative Filtering and Fuzzy System,” Journal of Korean institute of intelligent systems, Vol. 19, No. 2, pp. 242-247, 2009.

[16] Gediminas Adomavicius, Alexander 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.

[17] Sven Ewan Shepstone, Zheng-Hua Tan, Søren Holdt Jensen, “Demographic Recommendation by

means of Group Profile Elicitation Using Speaker Age and Gender Recognition," 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013) : Speech in Life Sciences and Human Societies, pp. 2827-2831, 2013

[18] J. A. Recio-Garcia, G. Jimenez-Diaz, A. Sanchez-Ruiz, B. Diaz-Agudo, "Personality Aware Recommendations to Groups," Proceedings of the 3rd ACM International Conference on Recommender Systems (RecSys), New York, USA, pp. 325-328, 2009.

[19] TripAdvisor. http://www.tripadvisor.co.kr

[20] Expedia. http://www.expedia.co.kr

[21] Amazon. http://www.amazon.com

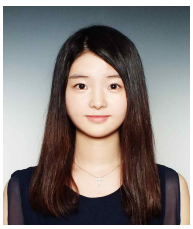
[22] C. W. Leung, S. C. Chan, F. l. Chung, "Integrating collaborative filtering and sentiment analysis: A rating inference approach," Proceedings of The ECAI 2006 Workshop on Recommender Systems, pp. 62-66, 2006.

[23] C. C. Musat, Y. Liang, B. Faltings, "Recommendation using textual opinions," Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, pp. 2684-2690, 2013.

[24] E. Cambria, B. Schuller, Y. Xia, C. Havasi, "New avenues in opinion mining and sentiment analysis," IEEE Intelligent Systems, vol. 28, Issue 02, 2013.

[25] S. Dooms, "Dynamic Generation of Personalized Hybrid Recommender Systems," Proceedings of the 7th ACM International Conference on Recommender Systems (RecSys), Hong Kong, China, pp. 443-446, 2013.

[26] GroupLens Research. MovieLens Datasets. http://grouplens.org/datasets/movielens/



김 현 경

2011년 – 현재 연세대학교 컴퓨터과학과 학부

관심분야 : 데이터 마이닝, 텍스트 마이닝, 바이오 인포매틱스

김 현 진

2010년 연세대학교 컴퓨터과학과 졸업 (학사)

2010년 – 현재 연세대학교 컴퓨터과학과 통합과정

관심분야 : 바이오 인포매틱스, 데이터 마이닝, 텍스트 마이닝, 그래프 마이닝, 데이터베이스

박 상 현

1989년 서울대학교 컴퓨터공학과 졸업 (학사)

1991년 서울대학교 대학원 컴퓨터공학과(공학석사)

2001년 UCLA 대학원 컴퓨터과학과(공학박사)

1991년 – 1996년 대우통신 연구원

2001년 – 2002년 IBM T. J. Watson Research Center Post-Doctoral Fellow

2002년 – 2003년 포항공과대학교 컴퓨터공학과 조교수

2003년 – 2006년 연세대학교 컴퓨터과학과 조교수

2006년 – 2011년 연세대학교 컴퓨터과학과 부교수

2011년 – 현재 연세대학교 컴퓨터과학과 교수

관심분야 : 데이터베이스, 데이터마이닝, 바이오 인포매틱스, 적응적 저장장치 시스템, 플래쉬메모리 인덱스, SSD