협업필터링 기반 추천시스템에서 유사도 계산의 신뢰성 향상 방안

이건우*, 전동엽*, 하지운**, 김형욱**, 김상욱**§
*한양대학교 컴퓨터전공
**한양대학교 컴퓨터·소프트웨어학과
e-mail:{lgw0915, dongyeoup, jiwoonha, ook0810, wook}@hanyang.ac.kr

An Approach to Improve the Credibility of Similarity Calculation in CF-based Recommender Systems

Gun Woo Lee*, Dong Yeoup Jeon*,
Jiwoon Ha**, Hyung-ook Kim**, Sang-Wook Kim**§
*Division of Computer Sciences and Engineering, Hanyang University
**Department of Computer and Software, Hanyang University

요 약

협업 필터링 기반 추천 시스템에서는 이웃 사용자를 정확하게 찾는 것이 추천 정확도에 핵심적인 영향을 미친다. 그러나 기존의 유사도 척도는 사용자가 공통으로 평가한 아이템만을 고려하여 유사도를 계산하기 때문에 이러한 아이템이 적은 사용자 간의 유사도가 부정확하게 계산되는 문제가 있다. 본 논문에서는 이러한 문제를 극복하기 위해 공통으로 평가하지 않은 아이템을 함께 고려하여 유사도를 계산하는 방안을 제안한다. 또한, 실험을 통해 제안하는 방안이 협업 필터링 기반 추천 시스템의 정확도 향상에 기여함을 보인다.

1. 서론

협업 필터링(collaborative filtering) 기법은 추천 시스템 분야에서 널리 연구 및 사용되고 있는 방법으로, 내용 정보 및 적용 분야에 관계없이 합리적인 정확도를 얻을 수 있는 장점이 있다. 협업 필터링 기반 추천 시스템은 서로 유사한 사용자들은 특정 아이템에 대해 유사한 선호도를 가질 것이라는 가정에 기반하고 있다[1]. 협업 필터링 기반 추천 시스템은 크게 사용자 기반 추천 시스템과 아이템 기반 추천 시스템이 존재한다[1,2]. 본 논문에서는 편의상 사용자 기반 추천 시스템에 대해서만 설명한다[1].

추천 시스템에서 아이템을 추천 받는 사용자를 타겟 사용자, 타겟 사용자의 선호도를 예측하고자 하는 아이템을 타겟 아이템이라 칭한다. 협업 필터링 기반 추천 시스템에서는 타겟 사용자의 타겟 아이템에 대한 선호도를 예측하기 위해 타겟 사용자와 유사한 이웃 사용자들을 찾고, 이웃 사용자들의 타겟 아이템에 대한 선호도를 종합하여 타겟 사용자의 타겟 아이템에 대한 선호도를 예측한다[1,2].

협업 필터링 기반 추천 시스템에서 사용자 간 유사도를 계산할 때 가장 널리 사용되는 척도로는 피어슨 상관계수(Pearson correlation coefficient)와 코사인 유사도(cosine similarity)가 있다. 위의 유사도 척도들은 유사도 계산 시, 두 사용자가 공통으로 평가한 아이템들만을 고려한다[1,2]. 이로 인해, 공통으로 평가한 아이템들이 많은 두 명의 사용자보다, 극히 적은 수의 아이템만을 공통으로 평가한 두 명의 사용자가 더 유사한 것으로 계산되는 경우가 발생한다.

본 논문에서는 이러한 기존의 유사도 계산 방안의 한계를 극복하기 위해, 두 사용자가 공통으로 평가하지 않은 아이템들을 함께 고려하여 유사도를 계산하는 방안을 통해 협업 필터링 기반 추천 시스템의 추천 정확도를 향상시키는 방안을 제안한다.

2. 기존 환경에서의 유사도 측정 결과 분석

위에서 언급한 바와 같이, 피어슨 상관계수와 코사인 유사도는 두 사용자가 공통으로 평가한 아이템만을 고려하여 두 사용자 간의 유사도를 계산한다. 이로 인해, 직관적으로 타당하지 않은 계산 결과가 도출되는 경우가 발생한다.

예를 들어, 아래와 같은 세 명의 사용자 A, B, 그리고 C가 있다고 가정한다. 각 열은 영화들을 나타내며, 각 행은 사용자들을 나타낸다. 칸의 숫자는 해당 사용자가 해당 영화에 부여한 평점이다.

사용자	코미디 영화									공포 영화								
A	3	2	5	3	2	1	1	5	4	3	4							
B								4	3	2	1	1	2	3	2	1	3	3
C	2	1	2	3	3	2	2	3	4	2	1							

(그림 1) 세 명의 사용자의 영화에 대한 평점.

1) 이후의 설명은 아이템 기반 추천시스템에도 동일하게 적용될 수 있다.

§ 교신저자

이 논문은 2014년도 및 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단 (No. NRF-2014R1A2A1A10054151, No. 2015R1A5A7037751), 그리고 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업 (IITP-2015-H8501-15-1013) 의 지원을 받아 수행되었음.

사용자 A와 C는 주로 코미디 영화를 보며, 사용자 B는 주로 공포 영화를 본다. 이 경우 직관적으로 생각하였을 때, 사용자 A와 B보다 A와 C가 더 서로 유사한 사용자로 볼 수 있다. 그러나 실제로 피어슨 상관계수를 통해 유사도를 계산하면, 사용자 A와 B의 유사도는 약 0.632, 사용자 A와 C의 유사도는 약 0.235로 나타난다. 마찬가지로 코사인 유사도를 통해 유사도를 계산하면, 사용자 A와 B의 유사도는 약 0.944, 사용자 A와 C의 유사도는 0.898로 나타난다. 이는 직관적으로 생각한 결과와 반대되는 결과이다.

3. 제안하는 방안을 통한 유사도 계산 결과 분석

본 논문에서는 이러한 문제를 해결하기 위해 사용자들이 공통으로 평가하지 않은 아이템도 함께 고려하여 유사도를 계산하는 방안을 제안한다.

사용자들은 아이템에 대한 자신의 관심에 따라 아이템을 선택하고, 아이템을 사용한 후에 자신의 선호도를 평점으로 나타낸다. 즉, 특정 사용자에 의해 평가되지 않은 아이템은 해당 사용자로부터 관심을 받지 못한 아이템일 가능성이 높다. 이에 착안하여, 본 논문에서는 사용자가 평가하지 않은 아이템에 대해 사용자의 관심도가 없음을 의미하는 0점을 부여하였다.

사용자	코미디 영화									공포 영화								
A	3	2	5	3	2	1	1	5	4	3	4	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	4	3	2	1	1	2	3	2	1	3	3
C	2	1	2	3	3	2	2	3	4	2	1	0	0	0	0	0	0	0

(그림 2) 0점 부여 후의 사용자들의 영화에 대한 평점.

사용자에 의해 평가되지 않은 아이템에 대해 0의 평점을 부여한 후 피어슨 상관계수를 이용하여 유사도를 계산할 경우, 사용자 A와 B의 유사도는 약 -0.088, 사용자 A와 C의 유사도는 약 0.764로 나타난다. 또한, 코사인 유사도를 통해 유사도를 계산하면, 사용자 A와 B의 유사도는 약 0.470, 사용자 A와 C의 유사도는 약 0.898로 나타난다.

이는 기존의 유사도 계산 방안을 통해 유사도를 계산했을 때에 비해 더욱 직관적으로 타당한 것이라 할 수 있다. 이로부터, 미평가 아이템을 고려하지 않고 유사도를 계산하는 것보다 미평가 아이템을 함께 고려하여 유사도를 계산하는 것이 합리적임을 알 수 있다.

4. 제안하는 방안을 통한 추천 정확도 향상 검증

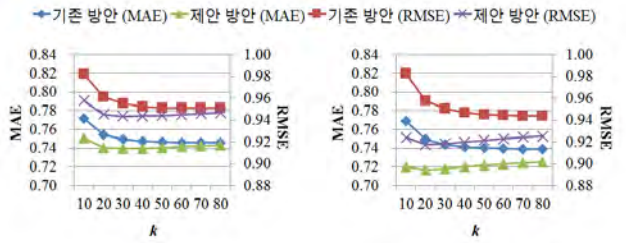
제안하는 유사도 계산 방안의 타당성을 검증하기 위해 MovieLens 데이터를 이용하여 실험을 수행하였다. MovieLens 데이터는 사용자가 943명, 영화가 1,682개이며, 사용자가 영화에 부여한 평점은 총 100,000개이다. 각 사용자는 최소 20개 이상의 영화에 1점과 5점 사이의 정수를 평점으로 부여하였다.

추천 정확도 검증을 위해 100,000개의 평점을 4:1의 비율로 트레이닝 셋과 테스트 셋으로 나누어, 사용자 기반 추천 시스템과 아이템 기반 추천 시스템 이용하여 테스트 셋에 대해 평점을 예측한다[1]. 이 때, 이웃 사용자 중 타겟 아이템에 대한 평점이 존재하지 않는 (0점으로 평점이 부여된) 사용자는 이웃 사용자에서 제외하였다.

예측한 평점과 실제 평점과의 오차 측정에는 mean absolute error(MAE)와 root-mean-square deviation(RMSE)을 이용하였으며, 5번의 교차 검증(5-cross validation)을 수행하였다.

협업 필터링 기반 추천 시스템에 제안하는 방안을 적용하였을 때, 추천 정확도가 향상되는지 검증하였다. 이를 위해, 기존 방안과 제안하는 방안을 통해 유사도를 측정하고, 각각에 대해 사용자 및 아이템 기반 추천 시스템 각각을 이용하여 평점을 예측하고, 그 정확도를 비교하였다. 이 때, 제안하는 방안이 이웃 수 k에 관계없이 항상 기존 방안으로 유사도를 계산하였을 때보다 높은 정확도를 보이는지 검증하기 위해 이웃 수 k를 10부터 80까지 변화시키며 실험을 수행하였다.

그림 3은 실험 결과를 그래프로 나타낸 것이다. 이 때, 유사도 척도로는 피어슨 상관계수를 사용하였다. 각 그래프의 x 축은 이웃 수 k를, y축은 MAE, RMSE 값을 나타낸다.



(a) 사용자 기반. (b) 아이템 기반.

(그림 3) 추천 정확도 비교.

그 결과, 모든 k 값에서 제안하는 방안을 통해 유사도를 계산하였을 때, 기존 방안에 비해 정확도가 높아지는 것을 알 수 있다. 특히, 사용자 및 아이템 기반 추천 시스템 각각에서 기존 방안의 최적의 k 값에서의 정확도와 제안하는 방안의 최적의 k 값에서의 정확도를 비교하였을 때에도 제안하는 방안을 적용하였을 때, 더 높은 정확도를 보이는 것으로 나타났다. 또한, 유사도 척도를 코사인 유사도로 사용하였을 때에도 동일한 결과가 나타났다. 이는 사용자들이 공통으로 평가하지 않은 아이템을 포함하여 유사도를 계산하는 것이 이웃의 수, 유사도 척도에 관계없이 정확도 향상에 도움이 됨을 의미한다.

5. 결론

본 논문에서는 두 사용자가 공통으로 평가하지 않은 아이템들도 고려하여 유사도를 계산하기 위해 미평가 아이템에 0점을 부여하는 방안을 제안하였다. 실험을 통해 기존 유사도 계산 방안에 비해 제안하는 유사도 계산 방법을 사용하는 것이 추천 정확도가 향상되는 것을 보였다.

참고문헌

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

[2] B. Sarwaret et al., "Item-based Collaborative Filtering Recommendation Algorithms," *WWW*, pp. 285-295, 2001.