유저의 장르 선호도를 반영한 추천

이호종, 황원석, 김상욱
한양대학교 전자컴퓨터통신공학과
e-mail: hojonglee@agape.hanyang.ac.kr

Recommendation Reflecting User Preferences on Genres

Ho Jong Lee, Won-Seok Hwang, Sang-Wook Kim
Dept. of Electronics and Computer Engineering, Hanyang University

요 약

MovieLens 를 대상으로 하는 추천 시스템에 대한 연구 중 *k*-NN 추천 방법은 정확도가 비교적 높지만 평점을 예측할 수 없는 상황이 발생할 수 있다. 본 논문에서는 기존 방법의 문제점을 해결한 장르기반 추천 방법 제안하고, 실험을 통하여 제안하는 방법이 모든 영화에 대한 평점의 예측이 가능함을 검증한다.

1. 서론

영화에 대한 정보를 제공하는 유명한 웹 사이트 중 하나로 MovieLens[1]가 있다. MovieLens 에는 영화 제목, 장르, 개봉 연도와 같은 영화의 정보와 유저가 자신이 관람한 영화에 대하여 부여한 평점 정보가 존재한다. 또한, 나이, 성별, 직업, 우편번호와 같은 유저의 간단한 신상정보가 등록되어 있다. 이러한 정보를 통해 각 유저에게 적합한 영화를 자동으로 추천해 주는 추천 시스템에 대한 다양한 연구들이 진행되어 왔다.

추천 시스템 중에서 널리 사용되는 방법은 *k*-Nearest Neighbor (*k*-NN) 추천 방법이다 [1]. 이 방법은 추천 대상인 타겟 유저가 아직 관람하지 않은 영화들에 대하여 평점을 예측한다. 예측을 위해 타겟 유저와 가장 유사한 *k* 명을 Pearson's correlation coefficient [2] 또는 cosine similarity [3]를 이용하여 찾는다. *k* 명의 유사한 유저들의 각 영화에 대한 평가와 각 유저의 유사도를 이용하여 타겟 유저의 각 영화의 평점을 계산한다.

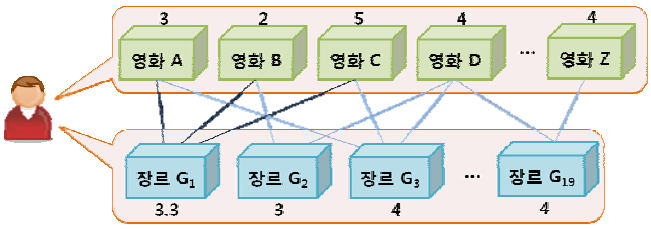
k-NN 추천 방법의 예측은 MovieLens 를 대상으로 하는 다양한 추천 시스템 중에서 정확도가 높은 것으로 알려져 있다 [4]. 그러나 기존의 *k*-NN 추천 방법은 타겟 유저와 유사한 *k* 명의 유저가 평가하지 않은 영화에 대해서는 평점을 예측하지 못하는 문제가 있다. 예를 들면, 타겟 유저의 영화 A 에 대한 평점을 예측하고 싶을 때, 타겟 유저와 유사한 유저 *k* 명이 영화 A 에 대하여 평점을 부여하지 않았다면 영화 A 의 평점을 예측하지 못한다. 특히, 새롭게 등록된 영화는 해당 영화에 대하여 평점을 부여한 유저가 거의 없기 때문에 *k*-NN 추천 방법을 통해 평점이 예측되기 어렵다.

이러한 문제를 해결하기 위하여 본 논문에서는 MovieLens 에서 제공하는 정보 중에 하나인 영화의 장르 정보를 이용하고자 한다. 즉, 유저가 각 영화에 대해 평가한 평점을 종합하여 장르에 대한 관심으로 표현하고, 이를 이용하여 기존의 *k*-NN 추천 방법에서 찾지 못하였던 유사한 유저를 추가적으로 찾고자 한다. 또한, 각 유저의 장르에 대한 관심도를 이용하여 영화를 추천함으로써 아직 평점이 부여되지 못한 영화도 평점을 예측하려 한다. 본 논문에서는 이와 같은 방식으로 영화의 평점을 예측하는 장르기반 추천 방법을 제안한다. 또한, MovieLens 데이터를 이용하여 제안하는 방법이 영화의 평점을 예측하지 못하였던 문제점에 대해 해결함을 보인다. 또한, 제안하는 방법의 정확도를 기존 방법인 *k* 최인접 이웃 추천 방법과의 비교 실험을 통해 확인한다.

2. 장르기반 추천 방법

장르기반 추천 방법에서 각 유저의 장르에 대한 선호하는 정도를 벡터로 표현하고, 이를 유저 벡터로 부른다. 한 유저에 대한 유저 벡터를 구성할 때, 하나의 장르에 대한 선호도는 그 유저가 그 장르에 속한 영화들에 부여한 평점의 평균으로 결정된다. 그리고 영화는 여러 장르에 속할 수 있기 때문에 여러 장르의 선호도에 영향을 미친다.

그림 1 은 여러 영화에 대하여 평점을 부여한 유저에 대하여 유저 벡터를 구성하는 방법을 표현한 것이다. 장르 G_1 의 선호도는 영화 A, 영화 B, 그리고 영화 C 의 평점을 평균한 것이다. 그리고 영화 A 는 장르 G_1, G_3 의 선호도를 결정하는데 영향을 미친다. 이와 같이 구성된 유저 벡터를 Pearson's correlation coefficient [2]에 적용하여 타겟 유저와 가장 유사한 *k* 명의 유저를 찾고, 이를 유사 유저 집단이라 부른다.



(그림 1) 유저 벡터의 구성.

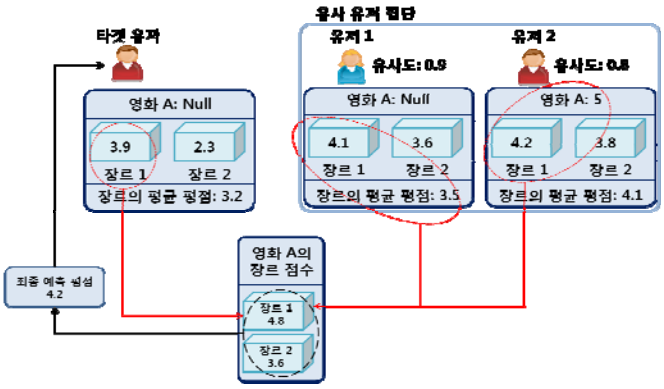
타겟 유저 *u* 의 영화 *m* 에 대한 예측 평점은 *m* 이 속한

[1] www.movielens.org

각 장르별 점수의 평균으로 계산된다. 영화 *m* 이 속한 장르 *g* 에 대한 점수는 *u* 가 *g* 를 선호하는 정도와 *u* 의 유사 유저 집단에 속한 유저의 영화에 대한 의견 또는 장르에 대한 의견을 종합하여 결정된다.

유사 유저 집단에 속한 유저들 중 영화 *m* 에 평점을 부여한 유저에게는 영화에 대한 의견, 평점을 부여하지 않은 유저에게는 장르에 대한 의견을 파악한다. 영화에 대한 의견은 그 유저가 *m* 에 부여한 평점과 장르 *g* 의 선호도의 차이를 통해 계산된다. 즉, 유저가 *g* 에 속하는 다른 영화들과 비교하여 *m* 을 얼마나 좋아하는지를 나타낸다. 이와 달리, 장르에 대한 의견은 그 유저의 모든 장르에 대한 선호도의 평균과 장르 *g* 의 선호도의 차이를 통해 계산된다. 이는 그 유저가 다른 장르들과 비교하여 장르 *g* 를 얼마나 좋아하는지 나타낸다.

그림 2 는 장르기반 추천 방법에서 타겟 유저의 영화 A 에 대한 점수 예측 과정을 표현한 것이다. 영화 A 의 점수는 A 가 속한 두 장르에 대한 점수의 평균으로 계산된다. 그림 2 에서 장르 1 에 대한 점수가 계산되는 과정을 설명한다. 장르 1 의 점수는 타겟 유저의 장르 1 에 대한 선호도(3.9)와 유저 1 의 장르에 대한 의견(4.1 - 3.5), 유저 2 의 영화에 대한 의견(5 - 4.2)을 종합하여 가중치 합을 통해 계산된다.



(그림 2) 장르기반 추천 방법.

3. 실험

실험을 위하여 MovieLens 데이터를 사용하였다. MovieLens 데이터는 유저는 943 명, 영화는 1,682 개, 장르는 19 개를 포함하고 있다. 모든 유저가 부여한 평점의 수는 총 100,000 건이고, 각 유저는 최소 20 개의 영화에 대하여 평점을 부여하였다. 평점은 1 점에서 5 점 사이의 정수로 부여되었다. 검증을 위하여 100,000 개의 평점을 훈련군과 실험군의 비율을 4:1 로 나누어서 5 번의 교차 검증(cross validation) 을 수행하였다.

실험에서는 *k*-NN 추천 방법과 장르기반 추천 방법을 비교하였다. 두 방법에서 타겟 유저와 유사한 유저의 수를 결정하는 파라미터 *k* 는 942 로 두고, 모든 유저로부터 추천을 받도록 하였다.

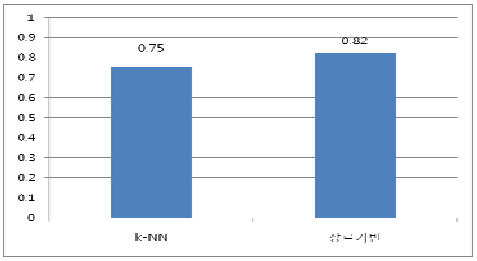
표 1 은 기존의 추천 방법과 제안하는 방법이 실험군의 평점에 대하여 예측하지 못한 유저-영화의 수를 보이고 있다. *k*-NN 추천 방법은 방법의 한계 때문에 유사 유저 전체를 대상으로 실험을 하였어도 추천을 하지 못하는 경우가 생기는 것을 확인할 수 있다. 그러나 제안하는 방법에서는 예측 못한 평점이 없다는 것을 확인할 수 있다.

<표 1> 각 방법이 추천하지 못한 유저-영화의 수

k-NN	장르기반
167 개	0 개

표 2 는 *k*-NN 추천 방법과 장르기반 추천 방법의 Mean Absolute Error (MAE) 를 측정한 것이다. MAE 는 실험 군에 있는 정답과 추천 시스템이 예측한 값의 차이의 평균을 나타낸다. 표 2 에서 볼 수 있듯이 장르기반 추천 방법이 *k*-NN 추천 방법보다 오차가 조금 더 크다는 것을 확인할 수 있었다.

<표 2> 각 추천 방법의 MAE



두 실험 결과를 통해 *k*-NN 추천 방법이 평점 예측에서 미세하게 정확했지만, 장르기반 추천 방법 또한 정확하다는 것을 파악할 수 있었다. 특히, 장르기반 추천 방법은 모든 유저-영화에 대해 예측을 할 수 있다는 점에서 기존 추천 방법의 문제점을 해결할 수 있음을 보였다.

4. 결론

본 논문에서는 *k*-NN 추천 방법에서 발생하는 문제를 해결하는 장르기반 추천 방법을 제안하였다. 또한, 실험을 통하여 장르기반 추천 방법이 대다수의 아이템에 대하여 추천하는 것이 가능하고, *k*-NN 추천 방법에 비하여 정확성이 많이 떨어지지 않는 것을 검증하였다.

감사의 글

본 연구는 지식경제부 및 정보통신산업진흥원의 'IT 융합 고급인력과정 지원사업' (NIPA-2011-C6150-1101-0001) 및 IT/SW 창의연구과정의 연구결과로 지식경제부와 삼성전자 주식회사에 의해 지원된 과제 (NIPA-2010-(C1810-1003-0007))로 수행되었음

참고문헌

[1] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl, “An algorithmic framework for performing collaborative filtering,” in *Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR ’99)*, pp. 230-237, 1999.

[2] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “Grouplens: an open architecture for collaborative filtering of netnews,” in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pp. 175-186, 1994.

[3] J.S. Breese, D. Heckerman, and C. Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering,” in *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pp. 43-52, 1998.

[4] J.J. Sandvig, B. Mobasher, and R. Burke, “A Survey of Collaborative Recommendation and the Robustness of Model-Based Algorithms,” *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, pp. 3-13, 2008.