사용자 기반과 아이템 기반 협업여과 추천기법에 관한 실증적 연구

김예나*, 최인복*, 박태근**, 이재동*
*단국대학교 컴퓨터과학과
**단국대학교 멀티미디어공학과
e-mail:yenakim@dankook.ac.kr

A Empirical Study on Recommendation Schemes Based on User-based and Item-based Collaborative Filtering

Ye-Na Kim*, In-Bok Choi*, Taekeun Park**, Jae-Dong Lee*
*Dept of Computer Science, Dankook University
**Dept of Multimedia Engineering, Dankook University

요 약

협업여과 추천기법에는 사용자 기반 협업여과와 아이템 기반 협업여과가 있으며, 절차는 유사도 측정, 이웃 선정, 예측값 생성 단계로 이루어진다. 유사도 측정 단계에는 유클리드 거리(Euclidean Distance), 코사인 유사도(Cosine Similarity), 피어슨 상관계수(Pearson Correlation Coefficient) 방법 등이 있고, 이웃 선정 단계에는 상관 한계치(Correlation-Threshold), 근접 N 이웃(Best-N-Neighbors) 방법 등이 있다. 마지막으로 예측값 생성 단계에는 단순평균(Simple Average), 가중합(Weighted Sum), 조정 가중합(Adjusted Weighted Sum) 등이 있다. 이처럼 협업여과 추천기법에는 다양한 기법들이 사용되고 있다. 따라서 본 논문에서는 사용자 기반 협업여과와 아이템 기반 협업여과 추천기법에 사용되는 유사도 측정 기법과 예측값 생성 기법의 최적화된 조합을 알아보기 위해 성능 실험 및 비교 분석을 하였다. 실험은 GroupLens의 MovieLens 데이터 셋을 활용하였고 MAE(Mean Absolute Error)값을 이용하여 추천기법을 비교 하였다. 실험을 통해 유사도 측정 기법과 예측값 생성 기법의 최적화된 조합을 찾을 수 있었고, 사용자 기반 협업여과와 아이템 기반 협업여과의 성능비교를 통해 아이템 기반 협업여과의 성능이 보다 우수했음을 확인 하였다.

1. 서론

인터넷이 발전함에 따라 정보과잉 현상이 발생하고, 이로 인해 사용자가 원하는 정보를 찾는 것이 어려워지고 있다. 이를 해결하기 위해 사용자의 선호도 정보를 분석하여 필요한 정보만 자동적으로 여과해주는 정보 필터링이 제안되었고, 이 같은 역할을 수행하는 추천 시스템이 등장하였다.

추천 시스템은 인터넷 상에서 많이 사용되고 있으며, 사용자들을 위한 다양한 종류의 아이템들을 추천한다. 예를 들어, 야후(Yahoo!)나 알타비스타(Alta Vista) 같은 검색엔진의 경우 사용자가 입력한 키워드를 기반으로 적절한 문서를 사용자에게 추천한다. 또한 Amazon.com과 Barnesandnoble.com의 경우 다른 소비자의 선호도를 기반으로 책과 영화를 추천한다[3]. 이처럼 추천시스템은 문서, 책, 영화 등 다양한 분야에 걸쳐 적용되고 있고 점차 그 분야를 넓혀가고 있다.

이러한 추천 시스템에서 널리 사용되는 추천기법으로 협업여과(Collaborative Filtering)가 있다[1][2][12]. 협업여과 추천기법은 다른 사용자의 평가를 기반으로 사용자에게 추천을 하는 기법으로 사용자 기반 협업여과(User-based Collaborative Filtering)와 아이템 기반 협업여과(Item-based Collaborative Filtering)가 있다. 사용자 기반 협업여과는 특정 사용자에게 아이템 추천을 하기 위해 다른 사용자의 평가를 이용하는 추천기법이다[11][7]. 아이템 기반 협업여과는 사용자 기반 협업여과의 단점을 보완하기 위해 나온 기법으로 사용자가 기존에 평가한 아이템들과 선호도를 예측하고자 하는 아이템간의 상관관계를 이용하여 선호도를 예측하는 방법이다[9].

협업여과 추천기법은 유사도 측정, 이웃 선정, 예측값 생성 단계로 이루어진다. 유사도 측정 기법에는 유클리드 거리(Euclidean Distance), 코사인 유사도(Cosine Similarity), 피어슨 상관계수(Pearson Correlation Coefficient) 등이 있고, 이웃 선정 기법에는 상관 한계치(Correlation-Threshold) 방법과 근접 N 이웃(Best-N-Neighbors) 방법 등이 있다. 마지막으로 예측값 생성 기법에는 단순평균(Simple Average), 가중합(Weighted Sum), 조정 가중합(Adjusted Weighted Sum) 등이 있다. 본 논문에서는 사용자 기반과 아이템 기반 협업여과 추천기법에 사용되는 유사도 측정 기법과 예측값 생성 기법의 최적화된 조합을 알아보기 위해, 추천기법에

서 사용되는 기법들을 비교 분석한다.

본 논문의 구성은 다음과 같다. 2장 관련연구에서는 사용자 기반 협업여과, 아이템 기반 협업여과와 협업여과 추천기법의 절차에 대해 알아본다. 3장에서는 협업여과 추천기법에 사용되는 유사도 측정 기법과 예측값 생성 기법의 비교 실험과 사용자 기반과 아이템 기반 협업여과 추천기법의 비교 실험을 위한 실험 환경을 알아보고, 4장에서는 실험을 통한 결과를 분석한다. 5장에서는 결론 및 향후 연구계획에 대하여 기술한다.

2. 관련연구

2.1 사용자 기반 협업여과와 아이템 기반 협업여과

사용자 기반 협업여과는 특정 사용자의 아이템에 대한 선호도를 예측하기 위하여 유사한 선호도를 가지는 이웃들을 기반으로 값을 예측하는 기법이다. 사용자 기반 협업여과는 다른 사용자들의 평가를 의미적으로 수집하고 분석하여 특정 사용자를 위한 추천을 하기 때문에 정보를 찾는 시간을 줄일 수 있고, 사용자의 취향이나 아이템의 질에 기반을 둔 추천도 가능하다. 그러나 특정 사용자와 다른 모든 사용자와의 유사도를 계산해야하기 때문에 수행 시간이 많이 소요되고, 선호도 정보가 적은 환경에서 신규 등록된 아이템과 신규 사용자에 대한 추천이 안 된다는 단점이 있다.

아이템 기반 협업여과는 사용자 기반 협업여과의 단점을 보완하기 위해 나온 기법으로 사용자가 기존에 평가한 아이템들과 선호도를 예측하고자 하는 아이템간의 상관관계를 이용하여 선호도를 예측하는 방법이다. 아이템 기반 협업여과는 아이템들 간의 유사도를 계산하여 모델을 형성하고 이것을 기반으로 아이템을 추천하므로 사용자 기반 추천기법에 비해 효율적이다[4]. 아이템 기반 협업여과 기법은 미리 계산된 모델을 사용하므로 계산 횟수가 적어 사용자 기반 협업여과에 비해 상대적으로 적은 시간이 소모된다는 장점이 있다. 그러나 모델을 구축하는데 많은 시간이 소요되고 특정 사용자와 선호도가 비슷하지 않은 사용자들의 평가를 기반으로 예측값을 생성한다면 아이템들 간의 유사도가 정확하지 않을 수 있다.

사용자 기반 협업여과와 아이템 기반 협업여과의 차이점은 (그림 1)과 같다. 사용자5의 아이템5에 대한 평점을 예측하기 위해 사용자 기반 협업여과를 사용할 경우 사용자5와와 유사도가 높은 사용자1과 사용자3을 선정하여 예측값을 생성한다. 아이템 기반 협업여과를 사용할 경우에는 아이템5와 가장 유사한 아이템2와 아이템3을 선정하여 예측값을 생성한다.

2.2 협업여과 추천기법 절차

협업여과 추천기법의 절차는 유사도 측정, 이웃 선정, 예측값 생성 단계로 이루어진다. 본 장에서는 사용자 기반 협업여과를 기준으로 각 단계별 설명한다.

2.2.1 유사도 측정 단계

유사도 측정 단계에서는 사용자의 과거 구매 기록이나 아이템에 대한 선호도를 토대로 유사도를 측정한다. 유사도 측정 기법에는 유클리드 거리, 코사인 유사도, 피어슨 상관계수 등이 있다.

유클리드 거리 측정 기법은 두 사용자를 m차원 공간에서 두 벡터로 표시하여 유클리드 거리를 측정하는 방법이다. 아이템 i에 대한 두 사용자 a와 b의 유클리드 거리를 구하는 식은 식(1)과 같다.

$$Sim(a,b) = \sqrt{\sum_{i=1}^{N}(r_{a,i} - r_{b,i})^2} \quad (1)$$

코사인 유사도 측정 방식은 두 데이터를 n차원 공간에서 두 벡터로 표시하여 코사인 각도를 측정하는 방법이며, 두 벡터의 상대적인 크기 차이에 영향을 받지 않는다. 식(2)는 사용자 a와 b 사이의 코사인 각도를 구하는 식이다.

$$Sim(a,b) = \frac{\vec{a} \bullet \vec{b}}{\| \vec{a} \|_2 \times \| \vec{b} \|_2} = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2} \sqrt{\sum_{i=1}^{N} b_i^2}} \quad (2)$$

피어슨 상관계수는 두 데이터 간의 관련성을 구하기 위해 보편적으로 이용되는 척도이며 식으로 표현하며 식(3)과 같다[5].

$$Sim(a,b) = \frac{\sum_{i=1}^{N}(r_{a,i} - \overline{r_a}) \bullet (r_{b,i} - \overline{r_b})}{\sigma_a \bullet \sigma_b} \quad (3)$$

(그림 1) 사용자 기반 협업여과와 아이템 기반 협업여과

2.2.2 이웃 선정 단계

이웃 선정 단계에서는 사용자와 높은 유사도를 갖는 이웃을 선택한다. 선택된 이웃들의 수는 그 수가 커질수록 성능이 향상되지만, 어느 수 이상 늘어나면 성능이 저하되므로 적정 크기의 이웃을 선택해야 한다[8]. 이를 위한 방법에는 특정 사용자와의 상관 값이 특정 한계치보다 큰 사용자 모두를 이웃으로 선택하는 상관 한계치 방법과 특정 사용자에 대하여 유사도가 높은 순서대로 앞의 N명을 이웃으로 선택하는 근접 N 이웃 방법이 있다.

2.2.3 예측값 생성 단계

마지막 단계인 예측값 생성 단계에서는 선택된 이웃들의 평점을 기반으로 사용자의 평점을 계산한다. 예측값을 생성하는 방법에는 단순평균, 가중합, 조정 가중합 등이 있다[6].

단순평균은 단순하게 합계만을 항목수로 나누거나 서로 곱한 것을 항목수로 풀어 근을 구하는 방법이다. 가중합은 가장 기초적이고 일반적인 방법으로서 특정 아이템에 대한 이웃의 평점과 유사도를 곱한 후 유사도의 합으로 나누는 방법이다. 식으로 표현하면 식(4)와 같다.

$$P_{a,i} = \frac{\sum_{b=1}^{N} Sim(a,b) \times r_{b,i}}{\sum_{b=1}^{N} (|Sim(a,b)|)} \quad (4)$$

조정 가중합은 초기 GroupLens 시스템에서 사용된 것으로 기존의 가중합 방법이 유사도에 바로 평점을 곱했다면 조정 가중합 방법에서는 평점에서 평균을 뺀 값을 유사도에 곱하는 방법이다[11]. 이를 식으로 표현하면 식(5)와 같다.

$$P_{a,i} = \overline{r_a} + \frac{\sum_{b=1}^{N} (r_{b,i} - \overline{r_b}) \times Sim(a,b)}{\sum_{b=1}^{N} (|Sim(a,b)|)} \quad (5)$$

3. 사용자 기반과 아이템 기반 협업여과의 비교 실험

3.1 실험데이터 셋

본 논문에서는 미국 미네소타 대학에서 개인화 추천을 연구하기 위해 영화에 대해서 선호도를 평가한 MovieLens 데이터 셋[10]을 사용하였다. 이 데이터 셋은 1997년 9월에서 1998년 4월까지의 총 조사기간 동안 수집된 943명의 사용자가 1,682편의 영화에 대한 총 100,000개의 평가치로 구성 되어 있다.

샘플링에 따른 Bias를 없애기 위해 10-fold Cross Validation을 수행하였다. 각 사용자의 1,682편의 영화에 대한 평가치들을 무작위로 90%와 10%로 나눈 후 90%는 학습데이터 셋(Training Data Set)로 사용하고, 10%는 평가데이터 셋(Test Data Set)으로 사용하였다.

3.2 평가 기준

본 논문에서는 예측값의 정확성을 평가하기 위해 MAE(Mean Absolute Error)를 사용하였다. MAE는 사용자의 실제값과 예측값의 차이에 대한 절대값의 평균을 나타내는 것이며, 절대적으로 알고리즘이 얼마나 정확하게 예측을 했는지를 알 수 있으며 식(6)과 같이 정의된다.

$$MAE = \frac{\sum_{i=1}^{N} |p_i - r_i|}{N} \quad (6)$$

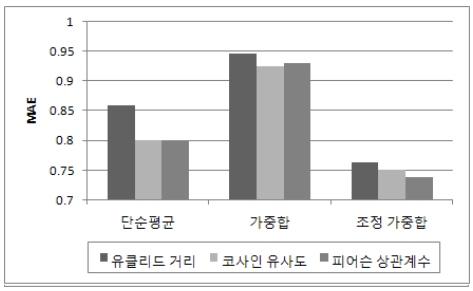
식(6)에서 p_i는 예측된 선호도이며 r_i는 실제로 사용자가 평가한 선호도이다. 또한 N은 새로운 사용자에 의해 평가된 아이템의 수를 의미한다.

3.3 비교 평가 방법

본 논문에서는 협업여과 추천기법에 사용되는 기법들을 비교 평가하기 위해 유사도 측정의 유클리드 거리, 코사인 유사도, 피어슨 상관계수와 예측값 생성의 단순평균, 가중합, 조정 가중합 방법을 비교 분석하였다. 실험을 통하여 최적화된 조합을 알아낸 후 사용자 기반과 아이템 기반 협업여과에서의 성능을 비교해 보았다.

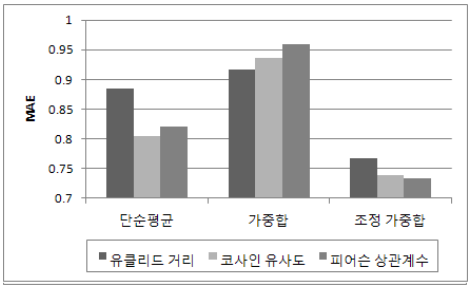
4. 실험 결과 및 분석

협업여과 추천기법에서의 유사도 측정과 예측값 생성 기법의 조합을 비교 분석하기 위해 이웃 선정은 근접 N 이웃을 사용하였고 N은 많은 연구에서 사용된 50으로 설정하였다[7][8]. 실험 결과는 (그림 2)와 같으며 유사도 측정 기법에서는 피어슨 상관계수의 MAE가 가장 낮은 것을 알 수 있다. 또한 예측값 생성 기법에서는 조정 가중합의 성능이 가장 우수함을 알 수 있다.



(그림 2) 사용자 기반 협업여과 추천기법 성능 비교

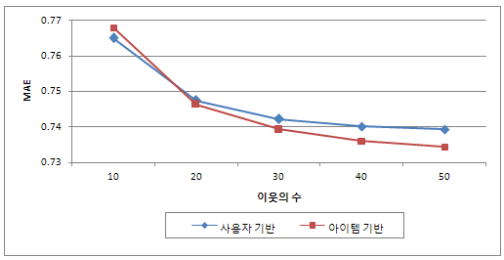
아이템 기반 협업여과에서 성능 실험 결과는 (그림 3)과 같다. (그림 3)를 보면 유사도 측정 기법에서는 피어슨 상관계수가, 예측값 생성 기법에서는 조정 가중합의 성능이 우수함을 알 수 있다.



(그림 3) 아이템 기반 협업여과 추천기법 성능 비교

이상의 두 가지 성능 분석 결과, 사용자 기반과 아이템 기반 협업여과 추천기법에서 모두 피어슨 상관계수와 조정 가중합 기법의 조합이 가장 우수함을 알 수 있다.

마지막으로 이웃 선정시 사용되는 근접 N 이웃 기법의 N의 변화에 따른 사용자 기반과 아이템 기반 협업여과 추천기법에서의 피어슨 상관계수와 조정 가중합 조합의 성능을 비교해보았다. 성능 실험 결과는 (그림 4)와 같다. 이웃의 수 N이 30 미만일 경우에는 사용자 기반 협업여과의 예측 성능이 우수했으나, 30 이상일 경우에는 아이템 기반 협업여과의 예측 성능이 보다 우수함을 알 수 있다.



(그림 4) 사용자 기반과 아이템 기반 협업여과 추천기법 비교

5. 결론 및 향후 연구

본 논문에서는 사용자 기반 협업여과와 아이템 기반 협업여과 추천기법에서 사용되는 기법들의 성능을 실험하고 분석하였다. 실험결과 사용자 기반과 아이템 기반 협업여과 모두 유사도 측정에서는 피어슨 상관계수가, 예측값 생성에서는 조정 가중합이 가장 성능이 우수했음을 알 수 있었다. 또한 피어슨 상관계수와 조정 가중합을 사용해 사용자 기반 협업여과와 아이템 기반 협업여과의 추천기법 비교 결과, 이웃의 수가 30 미만일 경우에는 사용자 기반 협업여과 기법이, 30 이상일 경우에는 아이템 기반의 예측 성능이 우수했음을 알 수 있었다.

향후 과제로는 사용자 기반과 아이템 기반 협업여과를 결합하여 보다 신뢰성 있고 예측 정확성이 향상되는 기법의 연구를 계획 중이다.

참고문헌

[1] 박지선, 김택헌, 류영석, 양성봉, “추천 시스템을 위한 2-way 협동적 필터링 방법을 이용한 예측 알고리즘,” 정보과학회지 논문지, 29권, 9·10호 pp.669-675, 2002.10.

[2] 이형동, 김형주, “협업 필터링 추천시스템에서의 취향 공간을 이용한 평가 예측 기법,” 한국정보과학회논문지, 제34권, 제5호, pp.389-395, 2007.10.

[3] Ansari A., Essegaier S., and Kohli R., "Internet Recommendation System," Journal of Marketing Research Vol 37, pp.363-375, 2000.

[4] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, “Item-Based Collaborative Filtering Recommendation Algorithms,” Proc. of the 10th International World Wide Web Conference, 2001.05.

[5] Daniel Billsus and Michael J. Pazzani, "Learning Collaborative Information Filters," Proc. of ICML, 1998.

[6] Gediminas Adomavicius and YoungOk Kwon, "New Recommendation Techniques for Multicriteria Rating Systems," IEEE Intelligent Systems, v.22, n.3, 2007.05.

[7] John S. Breese, David Heckerman, and Carl Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. of the Fourteenth Annual Conference on uncertainty in Artificial Intelligence, 1998.

[8] Jonathan L. H., Joseph A. K., Al B. and John R., "An Algorithmic Framework for Performing Collaborative Filtering," Research and Development in Information Retrieval, 1999.08.

[9] Jun Wang, Arjen P. de Vries, and Marcel J.T. Reinders, "Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion," Proceedings of SIGIR06, pp.501-508, 2006.

[10] http://www.cs.umn.edu/research/GroupLens/

[11] Resnick P., Iakovou N., Sushak M., Bergstrom P., and Riedl J., "GroupLens: An open architecture for collaborative filtering of netnews," Proc. of CSCW94, pp.175-186, 1994.

[12] Yu Chuan, Xu Jieping, and Du Xiaoyong, "Recommendation Algorithm combining the User-Based Classified Regression and the Item-Based Filtering," Proceedings of the 8th International Conference on Electronic Commerce, pp.574-578, 2006.08.