추천 시스템을 위한 2-way 협동적 필터링 방법을 이용한 예측 알고리즘

(A Predictive Algorithm using 2-way Collaborative Filtering for Recommender Systems)

박 지 선 † 김 택 헌 †† 류 영 석 ††† 양 성 봉 ††††

(Ji-Sun Park) (Taek-Hun Kim) (Young-Suk Ryu) (Sung-Bong Yang)

요 약 최근 전자상거래에서 대부분의 개인화 된 추천 시스템들은 고객의 취향에 맞는 적절한 상품을 추천하기 위하여 협동적 필터링 기술을 적용하고 있다. 사용자 기반 협동적 필터링은 특정 고객의 선호도와 가장 유사한 선호도를 가지는 고객 그룹의 선호도를 바탕으로 그 고객의 특정 상품에 대한 선호도를 예측하는 기법이다. 그러나 이 방법은 두 고객이 모두 평가를 한 상품이 있어야 하고 오직 두 고객 사이에서만 상관 관계를 구할 수 있으므로 예측의 정확성이 떨어질 가능성이 있다.

아이템 기반 협동적 필터링은 고객이 선호도를 입력한 기존의 상품들과 예측하고자 하는 상품의 상관관계를 계산하여 선호도를 예측한다. 이 방법에서는 상품들간의 유사도를 계산하기 위하여 두 상품에 대해 선호도를 입력한 고객들의 정보를 사용한다. 그러나 고객들간의 유사도가 전혀 고려되지 않기 때문에 만약 특정 고객과 전혀 선호도가 비슷하지 않은 사용자들의 평가를 기반으로 한다면, 상품들간의 유사도가 정확하지 않고 아울러 추천 시스템의 예측 능력과 추천 능력이 저하되는 문제점이 있다.

본 논문에서는 기존의 아이템 기반 협동적 필터링 기술의 문제점을 보완하고 추천 시스템의 예측 능력을 향상시키기 위하여 유사한 선호도를 가지는 고객들의 평가에 근거하여 상품들간의 유사도를 구하여 특정 상품에 대한 고객의 선호도를 예측하여 추천해 주는 기법을 제안한다. 본 논문에서 제안한 방법의 성능을 기존의 여러 다른 협동적 필터링 방법들과의 비교실험을 통해 평가하였다. 실험 결과 본 논문에서 제안한 방법이 기존의 다른 방법들보다 우수함을 확인할 수 있었다.

키워드 : 전자상거래, 추천 시스템, 협동적 필터링

Abstract In recent years most of personalized recommender systems in electronic commerce utilize collaborative filtering algorithm in order to recommend more appropriate items. User-based collaborative filtering is based on the ratings of other users who have similar preferences to a user in order to predict the rating of an item that the user hasn't seen yet. This may decrease the accuracy of prediction because the similarity between two users is computed with respect to the two users and only when an item has been rated by the users.

In item-based collaborative filtering, the preference of an item is predicted based on the similarity between the item and each of other items that have rated by users. This method, however, uses the ratings of users who are not the neighbors of a user for computing the similarity between a pair of items. Hence item-based collaborative filtering may degrade the accuracy of a recommender system.

In this paper, we present a new approach that a user's neighborhood is used when we compute the similarity between the items in traditional item-based collaborative filtering in order to compensate the weak points of the current item-based collaborative filtering and to improve the prediction accuracy.

· 본 논문은 2000년도 연세대학교 학술연구비의 부분적인 지원에 의하여 이루어진 것임.

† 학생회원 : LG전자 CDMA단말연구소
jspark@mythos.yonsei.ac.kr

†† 학생회원 : 연세대학교 컴퓨터과학과
kimthun@mythos.yonsei.ac.kr

††† 비 회 원 : 연세대학교 컴퓨터과학과
ryu@mythos.yonsei.ac.kr

†††† 비 회 원 : 연세대학교 컴퓨터산업공학부 교수
yang@mythos.yonsei.ac.kr

논문접수 : 2001년 6월 27일
심사완료 : 2002년 6월 27일

We empirically evaluate the accuracy of our approach to compare with several different collaborative filtering approaches using the EachMovie collaborative filtering data set. The experimental results show that our approach provides better quality in prediction and recommendation list than other collaborative filtering approaches.

Key words : Electronic Commerce, Recommender Systems, Collaborative Filtering

1. 서 론

전자상거래에서 개인화 된 추천 시스템은 자동화된 정보 필터링 기술을 적용하여 고객의 취향에 맞는 상품을 추천하여 고객의 구매 결정을 도와 주는 시스템이다. 추천 시스템에서 가장 중요한 것은 고객의 선호도를 정확하게 분석하고 정제하여 정확한 예측 능력으로 고객이 원하는 가장 적절한 상품을 추천해줄 수 있는 능력이다. 이를 위해서는 데이타 마이닝, 패턴 인식, 정보 필터링 등 다양한 기법들이 적용될 수 있으나 대부분의 추천 시스템들은 정보 필터링을 적용한다. 정보 필터링의 대표적인 것으로는 협동적 필터링(collaborative filtering)이 있다.

협동적 필터링은 추천 시스템에 가장 많이 사용되는 방법으로 Amazon.com, CDnow.com 등 상업적으로 성공한 여러 전자상거래 사이트에서 적용하고 있다[1]. 협동적 필터링에는 사용자 기반 협동적 필터링(user-based collaborative filtering)과 아이템 기반 협동적 필터링(item-based collaborative filtering)이 있다.

첫 번째로 사용자 기반 협동적 필터링[1][2][3][4]은 고객이 좋아할 만한 상품을 예측하기 위하여 비슷한 선호도를 가지는 다른 고객들의 상품에 대한 평가에 근거하여 추천하는 방법이므로 높은 예측능력과 추천능력을 가지는 장점이 있다. 특정 고객과 비슷한 선호도를 가지는 이웃들(neighbors)을 선정하는 기법에는 클러스터링(clustering), K-최대근접 이웃(k-nearest neighbor), 베이지안 네트워크(bayesian networks)와 같은 여러 가지 방법이 있으나 대부분의 경우 K-최대근접 이웃 방법을 이용한다[5][6]. 사용자 기반 협동적 필터링에서 지적되는 가장 큰 문제점은 고객 선호도간의 유사성을 평가하기 위해 사용하는 피어슨 상관 계수(Pearson correlation coefficient)로부터 야기된다[1][2][4]. 두 고객이 모두 평가를 한 상품이 있어야 하고 오직 두 고객 사이에서만 상관 관계를 구할 수 있으므로 예측의 정확성이 떨어질 가능성이 있다.

두 번째로 아이템 기반 협동적 필터링[5]은 대부분의 사람들이 과거에 자신이 좋아했던 상품과 비슷한 상품이면 좋아하는 경향이 있고 반대로 싫어했었던 상품과 비슷한 상품이면 싫어하는 경향이 있다는 점을 기반으로 하고 있다. 이 필터링 방법은 고객이 선호도를 입력한 기존의 상품들과 예측하고자 하는 상품과의 유사도(similarity)를 계산하여 고객의 선호도를 예측하는 방법이다. 즉, 예측하고자 하는 상품과 비슷한 상품들에 대하여 고객이 높은 평가를 하였다면 그 상품도 높게 평가를 할 것이라고 예측하고, 낮은 평가를 하였다면 그 상품도 낮게 평가를 할 것이라고 예측하는 것이다. 아이템 기반 협동적 필터링 방법은 상품들간의 유사도를 계산하기 위하여 두 상품에 모두 선호도를 입력한 고객들의 선호도를 사용한다. 그러나 고객들간의 유사도가 전혀 고려되지 않기 때문에 만약 특정 고객과 전혀 선호도가 비슷하지 않은 사용자들의 평가를 기반으로 한다면 상품들간의 상관 관계의 정확도가 떨어지고 아울러 추천 시스템의 예측 능력과 추천 능력이 저하될 수 있다.

본 논문에서는 위에서 언급한 기존의 협동적 필터링 기술의 문제점을 보완하기 위하여 사용자 기반 협동적 필터링과 아이템 기반 협동적 필터링을 결합한 새로운 방법을 제안한다. 이 방법은 K-최대근접 이웃 방법과 K-means 클러스터링 알고리즘을 사용하여 선호도를 예측하고자 하는 고객과 유사한 선호도를 가지는 고객들을 선별하여 그 고객들의 평가를 기반으로 상품들간의 유사도를 계산하여 상품에 대한 고객의 선호도를 예측하는 방법이다. 본 논문에서는 그 성능을 기존의 협동적 필터링 기술들과의 비교 실험을 통해 평가하였다.

본 논문의 구성은 다음과 같다. 2장에서 협동적 필터링 기술에 대한 관련연구를 설명하고, 3장은 본 논문에서 제안하는 알고리즘에 대해서 설명한다. 4장에서 실험을 통한 성능을 분석하며 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

2.1 사용자 기반 협동적 필터링

사용자 기반 협동적 필터링은 특정 고객의 상품에 대한 선호도를 예측하기 위하여 대부분의 경우 식(2)에 나타나 있는 피어슨 상관 계수를 이용하여 유사한 선호도를 가지는 이웃들을 정하고 식(1)에 의해 예측 선호도 값을 계산한다[2].

$$U_x = \overline{U} + \frac{\sum_{J \in Raters}(J_x - \overline{J})r_{UJ}}{\sum_{J \in Raters}|r_{UJ}|} \quad (1)$$

여기서

$$r_{UJ} = \frac{\sum(U-\overline{U})(J-\overline{J})}{\sqrt{\sum(U-\overline{U})^2 \cdot \sum(J-\overline{J})^2}}, -1 \le r_{UJ} \le 1 \quad (2)$$

U_x는 상품 x에 대한 고객 u의 예측된 선호도이고, r_{UJ}는 고객 u와 j의 상관관계를 나타내며 두 고객 모두 선호도를 표시한 상품에 대해서만 계산된다. 여기서 J_x는 상품 x에 대한 고객 j의 선호도를 나타내며 $\overline{J}$는 고객 j의 평균 선호도를 의미한다. r_{UJ}가 1에 가까울수록 두 고객의 선호도 경향이 매우 유사함을 나타내고 -1에 가까울수록 반대의 선호 경향을 나타낸다. *Raters*는 테스트 상품에 대해 선호도를 표시한 고객들을 의미한다.

협동적 필터링 방법을 적용한 Tapestry[7]는 협동적 필터링 방법을 가장 먼저 적용한 문서 필터링 시스템으로 워크그룹과 같은 공동체 구성원들의 의견에 기반 하여 추천을 해주므로 개인화 된 추천 서비스는 제공해 주지 못한다[1][4]. 최근 몇 년 동안에는 특히 자동화된 협동적 필터링 시스템이 많이 개발되었는데 그 중 GroupLens research system[3]은 고객과 유사 선호도를 가지는 이웃들의 의견에 기반 하여 유즈넷 뉴스와 영화에 대한 추천을 수행함으로서 Tapestry의 문제점을 보완하면서 성능을 인정받은 시스템이다 . GroupLens를 포함한 대부분의 협동적 필터링 기법을 사용하는 추천 시스템들로 Ringo[1], Video Recommender[1] 등이 있으며 이들은 모두 피어슨 상관 계수를 사용하여 유사 선호도를 가지는 이웃들을 결정한다[1][2].

2.2 K-최대 근접 이웃 기반 협동적 필터링

K-최대근접 이웃 방법은 과거 구매 기록을 통하여 특정 고객과 선호도가 가장 비슷한 k명의 고객들을 선택하는 것이다[1]. 전통적인 사용자 기반 협동적 필터링은 피어슨 상관 계수를 이용하여 특정고객과 다른 고객들의 선호도간의 유사도를 계산하여 이 고객들을 모두 특정고객의 이웃으로 인정한다. 반대로 K-최대근접 이웃 기반 협동적 필터링은 이 유사도가 높은 순서대로 k명을 선정하여 이웃으로 인정한다. 그러므로 K-최대근접 이웃 기반 협동적 필터링은 전통적인 사용자 기반 협동적 필터링보다 향상된 예측 능력을 가진다[8].

2.3 사용자 기반 협동적 필터링 기술의 한계점

GroupLens와 같은 사용자 기반 협동적 필터링 기술에서 사용한 피어슨 상관 계수 기반 예측 기법의 단점은 다음 세 가지로 요약할 수 있다[1][2][4]. 첫째, 두 고객사이의 상관관계는 오직 두 고객 모두 선호도를 표시한 상품에 대해서만 계산되므로 만약 상품의 수가 많으면 일반적으로 같은 상품에 대하여 두 고객 모두 선호도를 표시할 확률은 매우 적게 된다. 둘째, 비록 두 고객이 선호도에 따른 상관관계가 높지 않더라도 다른 고객의 선호도 예측에 좋은 자료가 될 수 있으나 상관관계가 높지 않다는 이유로 이 정보는 활용되지 못한다. 마지막으로, 상관관계가 오직 두 고객 사이에서만 계산된다는 것이다. 예를 들어 사용자 갑과 을이 아주 높은 상관관계에 있고, 을과 병도 그렇다고 가정하면 갑과 병도 상관관계가 높다고 할 수 있다. 그러나 만약에 갑과 병이 공통된 상품 어느 것에도 선호도를 표시하지 않았다면 상관관계를 구할 수 없게 된다.

2.4 아이템 기반 협동적 필터링

아이템 기반 협동적 필터링은 사용자 기반 협동적 필터링과 달리 고객이 기존에 평가한 각각의 상품들과 그 고객의 선호도를 예측하고자 하는 상품이 얼마나 비슷한가를 계산하여 k개의 가장 비슷한 상품들을 선택한다. 이렇게 가장 비슷한 k개의 상품들이 찾아지면 예측하고자 하는 상품에 대한 예측 값이 계산된다.

2.4.1 상품들간의 유사도

상품 x, y에 모두 선호도를 입력한 고객들을 추출하여, 이 고객들의 집합을 U라고 놓는다. 집합 U에 속한 각 고객들이 상품 x, y에 입력한 선호도에 의하여 식(3)과 같이 x와 y의 유사도 $Corr(x,y)$가 계산된다.

$$Corr(x,y) = \frac{\sum_{u \in U}(R_{u,x} - \overline{R_u})(R_{u,y} - \overline{R_u})}{\sqrt{\sum_{u \in U}(R_{u,x} - \overline{R_u})^2}\sqrt{\sum_{u \in U}(R_{u,y} - \overline{R_u})^2}},$$
$$-1 \le Corr(x,y) \le 1 \quad (3)$$

위의 식에서 $R_{u,x}$는 집합 U의 고객 u가 입력한 상품 x에 대한 선호도를 나타내며, $\overline{R_u}$는 고객 u의 전체 선호도의 평균을 의미한다. 식(3)에 의하여 계산된 유사도에 따라서 상품 x와 가장 비슷한 상품들을 k개 선정하여 집합 N이라 놓는다.

2.4.2 고객의 선호도 예측

특정 고객 u에 대한 상품 x의 예측 선호도 값을 x와 비슷한 상품들의 집합 N의 각 상품들에 대한 u의 선호도 값의 가중치 합계를 계산함으로써 구해진다. 즉, 상품 x와 y사이의 유사도에 따라서 각각의 선호도 값이 가중치 된다. 고객 u의 상품 x에 대한 선호도 예측 값을 계산하기 위한 식은 다음 식(4)와 같다. 여기에서 고객 u의 x와 비슷한 상품들에 대한 선호 경향에 따라 상품 x에 대한 u의 선호도 값을 예측한다.

$$P_{u,x} = \frac{\sum_{all\ similar\ items,\ N}(Corr_{x,N} \cdot R_{u,N})}{\sum_{all\ similar\ items,\ N}(|Corr_{x,N}|)} \quad (4)$$

2.5 아이템 기반 협동적 필터링의 한계점

아이템 기반 협동적 필터링은 상품들간의 유사도를 계산하기 위하여 이 두 상품에 모두 선호도를 입력한 고객들의 선호도를 사용하는데 고객들간의 유사도가 전혀 고려되지 않는 한계점이 있다. 따라서 만약 특정 고객과 전혀 선호도가 비슷하지 않은 사용자들의 평가를 기반으로 예측을 수행한다면 상품들간의 유사도의 정확도가 떨어지고 아울러 추천 시스템의 예측 능력과 추천 능력이 저하될 수 있다.

3. 사용자 기반과 아이템 기반 협동적 필터링의 결합

본 논문에서는 기존의 아이템 기반 협동적 필터링의 문제점을 보완하기 위하여 사용자 기반 협동적 필터링에서와 같이 먼저 고객의 취향과 비슷한 이웃들을 선별한 후, 그 이웃들의 평가를 기반으로 아이템 기반 협동적 필터링에서의 상품들간의 유사도를 계산하여 고객의 상품에 대한 선호도를 예측하여 추천해 주는 hybrid 기법을 제안한다. 이 기법은 특정 고객과 취향이 비슷한 이웃들의 평가에 의하여 상품들간의 유사도를 계산하게 되므로 추천 시스템의 예측 능력이 향상 될 수 있다. 따라서 식(3)에서 집합 U는 먼저 특정 고객과 선호도 경향이 비슷한 다른 고객들 중 선호도를 예측하고자 하는 상품 x와 이미 고객에 의하여 선호도가 표시된 상품들 $\{y_1, y_2, y_3, \ldots, y_n\}$에 선호도를 표시한 고객들의 집합이 된다. 본 논문에서는 특정 고객과 비슷한 선호도를 가지는 이웃들을 선택하기 위한 방법으로 K-최대 근접 이웃 기법과 K-means 클러스터링 알고리즘을 사용한다.

3.1 K-최대근접 이웃 기반 예측 알고리즘

K-최대근접 이웃 기반 예측 알고리즘은 임의의 상품에 대한 고객의 선호도를 예측하고 추천 리스트를 생성하기 위해서 다음과 같은 단계를 거친다.

[1] 식(2)를 이용하여 테스트 고객과 기존 고객들 각각에 대하여 선호도간의 유사도를 계산한다. 이 유사도가 높은 순서대로 테스트 고객과 취향이 비슷한 이웃들을 k명 선정한다.

[2] 식(3)을 이용하여 [1]에서 선정된 이웃들의 평가에 의한 상품들간의 유사도를 계산한다. 테스트 상품 x와 테스트 고객이 기존에 선호도를 표시한 상품들의 집합 $\{y_1, y_2, y_3, \ldots, y_n\}$에서 각 상품들과의 유사도 $Corr(x, y_n)$을 [1]에서 선정된 이웃들 중 이 두 상품 모두에 선호도를 표시한 이웃들의 평가에 의해 계산한다.

[3] [2]에서 계산된 유사도를 기반으로 테스트 고객의 테스트 상품 x에 대한 예측 선호도 값을 식(4)를 이용하여 계산한다.

[3]에서 계산된 예측 선호도 값이 높은 순서대로 n개의 상품을 선정하여 테스트 고객에 대한 추천 리스트를 생성한다. 이 방법은 기존의 아이템 기반 협동적 필터링에서 고객들간의 유사도가 전혀 고려되지 않는 한계점을 보완한 방법으로, 테스트 고객과 취향이 비슷한 이웃들 k명에 의하여 상품들간의 유사도가 계산되므로 추천 시스템의 예측의 정확성이 향상 될 수 있다.

3.2 K-means 클러스터링 알고리즘 기반 예측 알고리즘

K-means 클러스터링 알고리즘은 사전에 결정 된 군집 수 k에 기초하여 전체 데이타를 상대적으로 유사한 k개의 군집으로 구분하는 방법이며 각 데이타는 좌표평면의 점으로 표현된다[9]. 본 논문에서는 유사한 선호도를 가지는 고객들을 몇 개의 의미 있는 군집으로 나누기 위하여 이 방법을 사용한다. 상품의 각 속성에 대한 선호도를 각기 다른 차원으로 하여 좌표평면의 점으로 표현하고, K-means 클러스터링 알고리즘을 적용하여 기존 고객들을 k개로 군집화 한다. 이미 K-means 클러스터링 알고리즘을 통하여 나누어진 k개의 각 군집의 대표값들과 테스트 고객의 각 속성의 선호도값에 대하여 Euclidean distance를 계산하여 가장 최소의 값을 가지는 군집을 선택한다. 결정된 군집에 속하는 다른 고객들은 테스트 고객에 대한 새롭게 구성된 이웃들이며 따라서 이 이웃들에 대해서만 3.1절의 [2]와 [3] 단계를 적용한다. K-means 클러스터링 알고리즘은 다음과 같은 단계로 구성된다.

[1] 군집의 수 k를 정한 후, k개의 초기 군집 중심을 선택한다. 일반적으로 주어진 표본집합의 처음 k개의 표본을 임의로 선택한다.

[2] 각 고객들을 각 군집의 중심과 가장 가까운 거리에 있는 군집영역에 분배한다. 여기서 거리란 Euclidean distance에 의하여 계산된 값을 의미한다.

[3] [2]의 결과로부터, 모든 군집에 대하여 해당 군집에 포함된 모든 고객들의 선호도들로부터 새로운 군집 중심을 계산한다.

[4] 모든 군집에 대하여 기존의 중심과 새로운 중심의 차이가 없을 때까지 [2]부터 반복하고 그렇지 않으면 알고리즘은 수렴하며 종료된다.

이 방법은 K-means 클러스터링 알고리즘을 이용하여 군집화 한 후 테스트 고객에게 적절한 군집 내의 고객들만 이웃으로 결정하도록 함으로써 2장에서 언급한

피어슨 상관 계수에 의해 야기되는 문제점을 보완할 수 있다. 따라서 식(2)에서 J는 테스트 상품에 대해 선호도를 입력한 고객들 중 테스트 고객과 같은 군집 내에 있는 고객들을 의미한다. 같은 군집 내에 속한다는 것은 상품에 대해 유사한 선호도를 가진다는 것이므로 이러한 고객들만 협동적 필터링에 적용함으로써 기존의 사용자 기반 협동적 필터링과 아이템 기반 협동적 필터링 기법에 비하여 선호도 예측의 정확성의 향상과 질 높은 추천 리스트를 생성할 수 있다.

4. 실험 환경

4.1 테스트 데이타 셋 (data sets)

본 논문에서는 1997년 Compaq Computer Corporation에 의해서 공개된 EachMovie[10] 데이타 셋을 사용하였다. 이 데이타 셋은 총 72,916명의 사용자가 1,628개의 영화에 대해 0.0부터 최대 1.0까지 0.2의 차이를 두고 명시적으로 평가한 선호도들로 구성되어 있다. 영화의 장르는 액션, 애니메이션, 외국 예술, 고전, 코미디, 드라마, 가족, 공포, 로맨스, 스릴러의 10가지로 구분되어 있다. 먼저 총 72,916명의 사용자 중 최소 100회 이상 선호도를 입력한 사용자 4,788명을 추출한 후 모든 장르의 영화에 대해 선호도를 입력한 사용자 3,763명을 최종적으로 추출하였다. 이 중에서 테스트 고객 10명을 무작위로 선택하고 나머지 3,753명을 기존 고객으로 선택하였다. 선택된 테스트 고객 10명이 선호도를 입력한 영화들 중에서 각 고객마다 무작위로 5개의 영화를 테스트 영화로 선택하고 나머지 영화들을 기존 영화들로 선택하여 실험을 진행하였다.

4.2 평가 기준

4.2.1 예측의 정확성에 대한 평가

본 논문에서는 예측 값의 정확성을 평가하기 위해 MAE(Mean Absolute Error)를 사용하였으며 식(5)에 나타낸 것과 같이 구할 수 있다[11].

$$|E| = \frac{\sum_{i=0}^{N} |\varepsilon_i|}{N} \tag{5}$$

위의 식에서 N은 총 예측 회수이며 ε_i는 예측 값과 실제 값의 오차를 나타내고, i는 각 예측 단계를 나타낸다.

4.2.2 추천 리스트에 대한 평가

추천 리스트를 평가하기 위한 방법으로는 Precision, Recall, F-measure가 있다[2][12]. Precision은 추천 리스트 중에서 몇 개의 영화를 고객이 실제로 좋아했는지를 나타내는 평가 방법이며 Recall은 고객이 좋아하는 영화 중에서 얼마나 많은 영화가 추천이 되었는지 나타내는 평가 방법이다. 본 논문에서는 고객들의 실제 선호도 값이 0.7이상인 영화들을 고객이 좋아하는 것으로 가정하였으며, 예측된 선호도 값이 0.7이상인 영화들에 대해서만 추천 리스트를 생성하였다. F-measure는 Precision과 Recall에 동등한 중요도를 부여하여 하나의 평가방법으로 사용하는 것으로 식(6)과 같이 구할 수 있다.

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{6}$$

4.3 실험의 확장

K-means 클러스터링 알고리즘으로 이웃을 선정하는 방법을 피어슨 상관 계수로 이웃을 선정하는 방법과 비교하기 위하여 K-means 클러스터링 알고리즘을 사용하는 사용자 기반 협동적 필터링방법을 제안한다. 이 방법은 기존의 사용자 기반 협동적 필터링에서 피어슨 상관 계수가 아닌 K-means 클러스터링 알고리즘을 사용하여 유사한 선호도를 가지는 고객들을 적절히 군집화하여 특정 상품에 대한 고객의 선호도를 그 고객이 속한 군집내의 다른 고객들의 평가를 기반으로 예측하여 추천해 주는 방법이다.

4.4 파라미터

표1에서 Method_1은 본 논문에서 제안한 K-최대근접 이웃 기반 예측 알고리즘이며 Method_2는 K-means 클러스터링 알고리즘 기반 예측 알고리즘을 나타내고, KMC_CF는 K-means 클러스터링 알고리즘을 이용한 사용자 기반 협동적 필터링 방법이다. 또한 KNN_CF는 기존의 K-최대근접 이웃 기법을 이용한 사용자 기반 협동적 필터링 방법을 의미하며, PI_CF는 기존의 아이템 기반 협동적 필터링 방법을 나타낸다.

표 1 파라미터

방법	파라미터
KNN_CF	고객의 최대 근접 이웃 수 k
PI_CF	테스트 영화와 비슷한 영화의 개수 i
KMC_CF	클러스터의 개수 c
Method_1	고객의 최대 근접 이웃 수 k, 비슷한 영화의 개수 i, 고객_임계치 u
Method_2	클러스터의 개수 c

Method_1에서 사용한 파라미터 중 고객_임계치 u는 테스트 영화와 유사도를 계산고자 하는 기존 영화에 둘 다 선호도를 입력한 고객들의 최소 수를 나타낸다. 각각의 파라미터의 값은 실험적으로 반복된 실험결과를 통해 최적 값을 선택하였다.

5. 실험 결과 및 분석

추천 알고리즘의 예측 능력과 추천 능력을 평가하기 위하여 테스트 고객 10명에 대해서 각각 5개의 테스트 영화에 대한 예측 실험을 수행하였고, 이와 같은 테스트 셋을 총 3회 반복 실험 함으로써 전체 150회의 예측을 통한 실험 결과에 대한 평균을 본 실험의 결과로 하였다. 표2와 3은 본 논문에서 제안한 방법과 기존의 여러 다른 협동적 필터링 방법과의 비교 실험 결과를 나타낸 것이다. 표에 나타난 각 방법에서의 파라미터들은 많은 실험을 통하여 가장 좋은 MAE를 갖는 값으로 선정한 것이다.

표 2 예측의 정확성에 대한 평가

방법		MAE
기존의 방법	GroupLens	0.206294
	KNN_CF (k = 70)	0.217624
	PI_CF (i = 40)	0.195235
제안된 방법	KMC_CF (c = 15)	0.195190
	Method_1 (k = 60, i = 50, u = 2)	0.189554
	Method_2 (c = 23)	0.180665

표 3 추천 리스트에 대한 평가 (%)

방법		Precision at Top 3	Precision	Recall	F-measure
기존의 방법	GroupLens	80	76	78	77
	KNN_CF (k = 20)	62	61	64	63
	PI_CF (i = 40)	85	82	78	79
제안된 방법	KMC_CF (c = 30)	75	79	74	77
	Method_1 (k = 10, i = 50, u = 3)	83	84	67	74
	Method_2 (c = 25)	92	89	67	76

표2와 3의 비교실험 결과를 보면 아이템 기반 협동적 필터링 방법이 사용자 기반 협동적 필터링 방법보다 우수한 성능을 보임을 알 수 있다. 이웃 선정 기법의 비교 실험에서는 K-최대근접 이웃 사용자 기반 협동적 필터링 방법보다 K-means 클러스터링을 적용한 협동적 필터링 방법이 더욱 정확한 예측의 결과를 나타내고 있다. 본 논문에서 제안한 K-최대근접 이웃 기반 예측 알고리즘과 K-means 클러스터링 기반 예측 알고리즘이 기존의 아이템 기반 협동적 필터링 방법보다 우수한 성능을 보였다. 특히 K-means 클러스터링 기반 예측 알고리즘은 Precision at Top 3의 평가 항목에서는 여러 다른 협동적 필터링 방법에 비해 아주 우수한 성능을 보이고 있다.

6. 결론 및 향후 연구

본 논문에서는 기존의 아이템 기반 협동적 필터링의 문제점을 보완하고 추천 시스템의 예측 능력을 향상시키기 위하여 유사한 선호도를 가지는 고객들의 평가에 근거하여 상품들간의 유사도를 구하여 특정 상품에 대한 고객의 선호도를 예측하여 추천해 주는 기법을 제안하였다. 고객과 유사한 선호도를 가지는 이웃을 선정하는 기법으로 K-최대근접 이웃 방법과 K-means 클러스터링을 각각 적용하여 그 성능을 기존의 협동적 필터링 방법과 비교 실험한 결과 본 논문에서 제안한 방법이 예측의 정확성, 추천 리스트에 대한 평가 모두 성능 향상이 있었다. 특히 K-means 클러스터링 기법을 통해 적절히 군집화 하고 이 군집에 속한 고객들을 이웃으로 선정하는 것은 피어슨 상관 계수에 의하여 K-최대근접 이웃을 선정하는 방법에 비하여 우수한 성능을 보임을 알 수 있었다. K-최대근접 이웃 기반 예측 알고리즘은 기존의 아이템 기반 협동적 필터링의 문제점을 보완한 방법이며 K-means 클러스터링 기반 예측 알고리즘은 기존의 사용자 기반 협동적 필터링 방법과 아이템 기반 협동적 필터링 방법의 문제점을 모두 보완한 방법으로 그 성능을 비교 실험을 통해 확인할 수 있었다. 본 논문에서는 K-means 클러스터링에서 유사한 선호 패턴을 가지는 고객들을 군집화 할 때 영화의 장르에 대한 선호도를 가지고 수행하였으나 향후 연구에서는 영화의 여러 가지 속성에 대한 선호도를 이용한다면 보다 신뢰성 있고 향상된 결과를 기대할 수 있을 것이다.

참 고 문 헌

[1] Badrul M. Sarwar, George Karypis, Joseph A.Konstan, John T. Riedle, "Application of Dimensionality Reduction in Recommender System-A Case Study," *ACM WebKDD 2000 Web Mining for E-Commerce Workshop*, 2000.

[2] Daniel Billsus, Michael J. Pazzani, "Learning Collaborative Information Filters," *Proceedings of ICML*, pp.46-53, 1998.

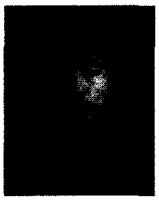
[3] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," *Proceedings of ACM CSCW94 Conference on Computer Supported Cooperative Work*, pp.175-186, 1994.

[4] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Analysis of Recommendation

Algorithms for E-Commerce," *The ACM E-Commerce 2000 Conference*, 2000.
[5] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," *Accepted for publication at the WWW10 Conference*, May, 2001.
[6] Herlocker, J., "Understanding and Improving Automated Collaborative Filtering Systems," *Ph.D. Thesis*, Computer Science Dept., University of Minnesota, 2000.
[7] Goldberg, D., Nichols, D., Oki, B.M., and Terry, D., "Using Collaborative Filtering to Weave an Information Tapestry," *Communications of the ACM*, Vol. 35, No. 12, pp.61-70, 1992.
[8] John S. Breese, David Heckerman, and Carl Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," *Proceedings of the conference on Uncertainty in Artificial Intelligence*, pp. 43-52, 1998.
[9] 이성환, 패턴인식의 원리 1권, p.96-100, 홍릉과학출판사, 1994.
[10] Steve Glassman, EachMovie collaborative filtering data set, Compaq Computer Corporation, URL: http://research.compaq.com/SRC/eachmovie/, 1997.
[11] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl, "An Algorithm Framework for Performing Collaborative Filtering," *Proceedings of the 1999 Conference on Research and Development in Information Retrieval*, 1999.
[12] Raymond J. Mooney, Loriene Roy, "Content-Based Book Recommending Using Learning for Text Categorization," *Proceedings of the fifth ACM Conference on ACM 2000 digital libraries*, pp.195-204, 2000.

박 지 선
1999년 동국대학교 컴퓨터공학 공학사. 2001년 연세대학교 컴퓨터과학 공학석사. 2001년 ~ 현재 LG전자 CDMA단말연구소 연구원. 관심분야는 전자상거래, CRM, 지능형에이전트

김 택 헌
1996년 동국대학교 컴퓨터공학 학사. 2000년 연세대학교 컴퓨터과학 석사. 2000년 ~ 현재 연세대학교 컴퓨터과학 박사과정. 1996년 ~ 2000년 삼성SDS(주) 근무. 관심분야는 전자상거래, CRM, 지능형에이전트



류 영 석
1999년 연세대학교 컴퓨터과학 공학사. 2001년 연세대학교 컴퓨터과학 석사. 2002년 ~ 현재 연세대학교 컴퓨터과학 박사과정. 관심분야는 전자상거래, CRM, 지능형에이전트

양 성 봉
1981년 연세대학교 공학사. 1984년 Univ. of Oklahoma 컴퓨터과학 석사. 1992년 Univ. of Oklahoma 컴퓨터과학 박사. 1993년 ~ 1994년 전주대학교 전자계산학과 전임강사. 1994년 ~ 현재 연세대학교 컴퓨터산업공학부 부교수. 관심분야는 전자상거래, 그래픽스, 인터넷 컴퓨팅