인구 통계 정보를 이용한 협업 여과 추천의 유사도 개선 기법

(Improvement on Similarity Calculation in Collaborative Filtering Recommendation using Demographic Information)

이 용 준 † 이 세 훈 †† 왕 창 종 †††
(Yong-Jun Lee) (Se-Hoon Lee) (Chang-Jong Wang)

요 약 본 논문에서는 추천시스템에서 많이 활용되는 협업 여과 방법의 문제점으로 지적되고 있는 희소성(sparsity)으로 인한 유사도의 부정확한 문제를 개선하기 위하여, 인구 통계 정보를 이용한 기법을 제안하였다. 두 사용자간의 유사도는 같은 항목에 동시에 평가된 점수를 기반으로 결정되며, 두 사용자가 동시에 평가하지 않은 항목은 유사도 계산에서 제외된다. 제안된 기법은 이러한 평가 점수 부족으로 인하여 유사도 계산이 정확치 못한 단점을 보완하기 위하여, 인구 통계 정보를 이용한 가상 평가 점수를 부가하여 유사도 계산을 개선, 예측의 정확도를 향상시킨 방식으로 기존의 피어슨 상관관계를 이용한 협업여과 방식의 확장이다. 실험은 Grouplens의 영화 평가 자료를 활용하였고, 평균절대오차(MAE)와 반응자 작용 특성(ROC)값을 이용하여 제안 기법과 피어슨 상관관계를 이용한 협업 여과 방식을 비교 하였다. 제안한 기법이 피어슨 상관관계를 이용한 협업 여과 추천 방식에 비하여 평균절대오차는 9%, 반응자 작용 특성의 민감도는 13% 향상되었음을 확인하였다.

키워드 : 추천시스템, 인구 통계 정보, 협업 여과

Abstract In this paper we present an improved method by using demographic information for overcoming the similarity miss-calculation from the sparsity problem in collaborative filtering recommendation systems. The similarity between a pair of users is only determined by the ratings given to co-rated items, so items that have not been rated by both users are ignored. To solve this problem, we add virtual neighbor's rating using demographic information of neighbors for improving prediction accuracy. It is one kind of extentions of traditional collaborative filtering methods using the peason correlation coefficient. We used the Grouplens movie rating data in experiment and we have compared the proposed method with the collaborative filtering methods by the mean absolute error and receive operating characteristic values. The results show that the proposed method is more efficient than the collaborative filtering methods using the pearson correlation coefficient about 9% in MAE and 13% in sensitivity of ROC.

Key words : Recommendation System, Demographic Information, Collaborative Filtering

1. 서 론

최근 인터넷의 사용이 점차 일반화됨에 따라 인터넷을 이용하는 통신 및 사이트의 수가 매 4개월 마다 2배씩 증가하고 있어, 여러 분야의 다양한 정보의 홍수 속에서 사용자가 원하는 정보를 찾는 것이 점차 어려워지고 있다[1]. 특히 전자상거래의 발달은 새로운 시장의 창출 및 고객에 대한 서비스 강화를 위하여, 고객들의 취향과 구매 이력을 분석하여 차별화된 정보를 자동으로 여과하여 추천하는 추천시스템을 출현하게 하였다[2]. 이러한 추천시스템에 사용되는 추천 기법은 내용기반 여과(content-based filtering), 협업 여과(collaborative filtering)가 주류를 이루고 있다[2,3]. 내용 기반 여과는 사용자의 프로화일을 기반으로 사용자의 유형을 추출하고, 이를 이용하여 사용자의 유형에 맞는 제품을 추천한다. 그러나 사용자가 처음 시도하는 항목에

† 종신회원 : 한국전기연구원 전기시험연구소 연구원
yjlee@keri.re.kr
†† 종신회원 : 인하공업전문대학 컴퓨터정보공학부 교수
seihoon@inhatc.ac.kr
††† 종신회원 : 인하대학교 컴퓨터공학부 교수
cjwang@inha.ac.kr
논문접수 : 2003년 2월 3일
심사완료 : 2003년 7월 21일

대해서는 설정된 정보가 없어 제품의 추천이 곤란하다[4,5].

협업 여과는 이러한 문제점을 해결하기 위해 다른 사용자의 평가를 기반으로 사용자에게 추천을 생성하는 기술이다. 어떤 정보를 이미 보았거나 경험한 사람들의 행동과 의견을 가지고 그 정보를 아직 보지 못한 사람들에게 그 정보의 가치를 예측하여 주는 시스템으로, 다른 사람들의 평가를 의미적으로 수집하고 분석하여 정보를 찾는 시간을 줄일 수 있다. 협업 여과는 Goldberg에 의해서 정보검색시스템에 적용하는 것을 시작으로 다양한 종류의 추천시스템에서 사용되고 있다[2,6,7,8,9]. 사무 업무그룹과 같은 폐쇄그룹 사용자간의 정보 공유를 위하여 개발된 TAPESTRY[6], 유즈넷 사용자와 영화를 위한 익명의 협업 여과 기법을 제시한 GroupLens[7,8], 음악 추천을 위한 Ringo[9]와 비디오 추천시스템[4] 등 여러 분야에서 적용되고 있다. 그러나 이러한 협업 여과 방식은 크게 두 가지 문제점을 가지고 있다[1,5]. 하나는 초기화에 대한 문제이며, 두 번째는 평가 자료의 희소성(sparsity)에 대한 문제이다.

본 논문에서는 두 번째의 문제점인 협업 여과 추천 계산의 희소성으로 발생되는 추천의 정확도를 보완하기 위하여 인구 통계 정보를 이용하여 가상평가 점수를 부가하고, 유사도 계산의 정확도를 향상시켜, 예측의 정확도를 높이는 방식을 제안하고자 한다. 실험은 Grouplens에서 제시한 영화 평가 자료를 기반으로 실시하였으며, 기존의 피어슨 상관관계(Pearson correlation coefficient)를 이용한 협업 여과 추천 방식에 비해 평가 정확도가 우수함을 검증하였다.

2. 관련 연구

이 장에서는 협업 여과와 관련된 기존의 연구동향 중 협업 여과 추천을 위한 추천 생성 기법(recommendation generation algorithm) 및 협업 여과 기법에 대한 연구와 유사도 개선 관련 기법에 대하여 살펴보고자 한다.

2.1 추천 생성 기법

Breese는 다양한 협업 여과 추천 기법을 모델 기반 기법과 메모리 기반 기법으로 크게 2가지로 구분하였다[10]. 모델 기반 기법은 수학적 모델에 기반을 두고 각 항목에 관한 사용자의 평가를 예측한다. 비록 모델 기반 기법에 많은 변수가 있어도 두 가지 그룹 즉 클러스터 모델과 베이시안 네트워크 모델로 분류할 수 있다. 추천 시스템 초기에는 베이시안 네트웍이나 클러스터 모델과 같은 추천 기법이 주로 사용되었다[2,11]. 클러스터 모델은 사용자들을 어떤 수의 클래스들로 분류하고 클래스 멤버가 항목 j에 평가 선호 점수 k를 줄 확률을 계산한다. 여기서 확률은 표준적인 베이시안 확률 공식을 이용하여 계산할 수 있다.

Bayesian probability: (1)

$$\Pr(C=c, e_1, \cdots e_n) = \Pr(C=c) \prod_{n=i}^{n} \Pr(e_i \mid C=c),$$

where e_i is evaluation for item i and

C is class membership.

클래스 멤버쉽 확률 ($\Pr(C=c)$)과 주어진 클래스 평가의 조건적 확률($\Pr(e_i \mid C=c)$)은 사용자 평가의 집합으로부터 계산된다. 클러스터 모델은 그룹에 중점이 두어져 있어 다른 방법보다 개인화 추천과는 거리가 있으며 근접 이웃 알고리즘보다 정확도가 떨어지기도 한다[10,12]. 근접 이웃 알고리즘[12]이란 사용자와 유사한 사용자를 이웃으로 선정하는 추천시스템에서 많이 사용되는 방법이다.

베이시안 네트워크 모델은 노드로서 항목을 나타내고 노드의 상태로서 평가를 나타낸다. 이 학습 알고리즘은 각 항목의 의존성에 대한 여러 가지 다양한 모델 구조를 찾을 수 있다. 결과 네트워크에서 각 항목은 평가의 최고 예측 값인 상위 항목의 집합을 갖는다. 베이시안 네트워크 모델은 오프라인에서 구성되므로 결과 모델은 작고, 빠르며, 근접 이웃 알고리즘과 같이 효과적이다. 모델 구성 시간에 비해 상대적으로 사용자의 선호도가 천천히 변하는 환경에 적합하다.

메모리 기반 알고리즘은 추천 받을 사용자에 대한 부분적 정보와 기존 사용자 데이타베이스로부터 계산된 가중치 집합에 기반하여 특정 사용자의 평가를 예측한다. 사용자 i의 항목 j에 대한 예측된 평가 (e_{ij})는 다른 사용자들의 평가에 대한 가중치 합이다.

Estimated evaluation : $e_{ij} = ef(s_{i.}, e_{.j})$ (2)

Estimated function : $ef_{ij}(s_{i.}, e_{.j}) = \sum_{t=1}^{n} e_{tj} wf(s_{it})$ $(t \neq i)$ (3)

where $wf(s_{it})$ is the weight function,

e_{ij} is the evaluation by user i on item j,

$e_{i.}$ is a evaluation by user i $(=e_{i1}, e_{i2}, \cdots, e_{im})$,

$e_{.j}$ is a users' evaluation on item j

$(=e_{1i}, e_{2j}, \cdots, e_{2n})$,

s_{ab} is the degree of similarity of user a and user b and

s_{ab} is $sf(e_{a.}, e_{b.})$.

다음 두 가지 방법 즉 상호관계성식과 벡터유사성식이 두 사용자 사이에 유사성을 계산하는데 가장 널리 사용되고 있다.

Correlation : (4)

$$sf(e_{a\cdot}, e_{b\cdot}) = \frac{\sum_{i=1}^{m}(e_{a,i}-e_{a\cdot})(e_{b,i}-e_{b\cdot})}{\sqrt{\sum_{i=1}^{m}(e_{a,i}-e_{a\cdot})^2}\sqrt{\sum_{i=1}^{m}(e_{b,i}-e_{b\cdot})^2}}$$

Vector similarity : (5)

$$sf(e_{a\cdot}, e_{b\cdot}) = | E_b - E_a |$$

where $E_a=(e_{a1},e_{a2},..,e_{am})$ and $E_b=(e_{b1},e_{b2},..,e_{bm})$.

2.2 협업 여과 기법

협업 여과의 하나의 연구 흐름은 협업 여과 시스템에서 사용된 기법에 대한 것이다. 다른 기법을 비교하거나[9,10,13], 성능을 향상키 위하여 알고리즘을 변형하였으며[14,16], 다른 방법들을 협업 여과에 조합하여 관찰하였다[2,15,17]. 두 번째 흐름은 이메일 메시지[6], 음악[9], 영화[16], usernet messages[1,7]과 같은 다양한 응용분야에 적용하여 관찰한 것이다.

Herlocker은 유사 가중치 방법(similarity weighting methods), 중요도 가중치(significance weighting), 변형 가중치(variance weighting),이웃 선택 방법(neighborhood selection methods), 예측 생성 방법(precision producing methods) 등을 비교하였다[16].

• 유사 가중치 방법 : 두 개의 유사한 방법(피어슨 상관관계, 스피어맨 순위 상관관계)을 비교하였다. 가중치 계산에 사용되는 상관관계는 피어슨 상관관계식이 많이 사용되며, 이 모델은 선형 회기 분석 모델(linear regression model)을 근간으로 구성되며, 선형이 아닌 경우 에러의 폭이 커진다. 즉 자료가 모두 선형일 수는 없으므로 유사도의 정확도가 떨어지는 요인이 된다.

$$w_{a,u} = \frac{\sum_{i=1}^{m}(r_{a,i}-\overline{r_a})*(r_{u,i}-\overline{r_u})}{\sigma_a * \sigma_u} \quad (6)$$

where $r_{a,i}$, $r_{u,i}$는 user a, u의 item i의 점수,

$\overline{r_a}$, $\overline{r_u}$는 user a, u의 점수 평균,

σ_a , σ_a는 user a, u의 점수 표준편차.

수식 (6)은 피어슨 상관관계 상수에 의해 정의된 사용자 a와 이웃 u 간의 항목 i에 대한 유사도 가중치를 나타낸다. 수식 (6)의 변형으로 수식 (7)과 같이 점수가 아닌 순위를 도입한 스피어맨 순위 상관관계(Spearsman rank correlation)을 비교하여 유사 가중치 방법인 두 방식이 크게 차이가 없음을 발견하였다.

$$w_{a,u} = \frac{\sum_{i=1}^{m}(rank_{a,i}-\overline{rank_a})*(rank_{u,i}-\overline{rank_u})}{\sigma_a * \sigma_u} \quad (7)$$

where $rank_{a,i}$, $rank_{u,i}$ 는 user a, u의 item i의 순위

$\overline{rank_a}$, $\overline{rank_u}$ 는 user a, u의 순위 평균

• 중요도 가중치 : 항목이 적은 경우 중요도 가중치를 추가하면 예측 결과의 정확도가 높아진다. 점수 항목이 50 이하이면 n/50의 중요도 가중치를 적용한다. n은 점수 항목의 수이며, 점수 항목이 50 이상이면 중요도 가중치는 1로 적용한다. 중요도 가중치는 협업 여과의 정확도를 향상시켰다.

• 변형 가중치 방법 : 변형 가중치는 정확한 평가를 제공한다고 입증된 사용자에게 보다 높은 가중치를 부여하는 가중치 방법이다. 이 가중치는 협업 여과 시스템의 정확도에 큰 영향이 없었다.

• 이웃 선택 방법 : 추천을 제공하기 위해 협업 여과에서는 사용자 마다 선호 이웃(참조 그룹)을 선택하여야 한다. 현재의 사용자와 높은 유사도를 갖는 사람들의 그룹을 정의하는 작업이다. 이웃은 상관-한계치(threshold), 근접 n 이웃으로 선택할 수 있다. 상관-한계치 방법은 현재 사용자와의 상관 값이 특정 한계치(예:0.5, 0.7)보다 큰 사용자를 이웃으로 선택하는 방법이다. 근접 n 이웃은 현재 사용자보다 높은 상관 값을 갖는 사용자 n 명을 선택하는 방법이다. 상관-한계치가 근접 n 이웃보다 성능이 좋은 것으로 밝혀졌으나, 범위(추천을 생성하는데 참여하는 사용자의 비율)에 따라 오차가 크게 나타났다.

• 예측 생성 방법 : 이웃이 선택되면 이 이웃들로부터의 점수를 기반으로 예측을 계산한다.

$$P_{a,i} = \overline{r_a} + \frac{\sum_{u=1}^{n}(r_{u,i}-\overline{r_u})*w_{a,u}}{\sum_{u=1}^{n}w_{a,u}} \quad (8)$$

where $w_{a,u}$는 수식 (6) 또는 (7)의 유사도 가중치

수식 (8)은 사용자 a의 항목 i에 대한 평가계산예측을 나타내며 n은 이웃의 수를 나타낸다. 사용자의 점수 분포가 다르므로 이를 표준화하는 z-score 방식(9)도 도입되었으나, 예측에는 크게 영향을 미치지 못하였다[8].

$$P_{a,i} = \overline{r_a} + \sigma_a * \frac{\sum_{u=1}^{n}\frac{(r_{u,i}-\overline{r_u})}{\sigma_u} * w_{a,u}}{\sum_{u=1}^{n}w_{a,u}} \quad (9)$$

협업 여과의 성능은 예측을 위해 사용된 참조 그룹의 수에 따라 너무 적거나, 크면 예측 평가 결과가 낮아진다는 것도 보였다[16].

Breese는 상관관계, 벡터 유사도, 베이시안 클러스터, 베이시안 네트웍등의 4가지 협업 여과 기법을 웹 페이지 방문, TV 프로그램, 영화 등 3가지 다른 분야에서 비교하여 피어슨 상관관계와 베이시안 네트웍이 대부분 동등한 성능을 나타냄을 발견하였다[10].

Shardanand과 Maes는 유사도 계산에 두 가지 다른 기법(피어슨 상관관계와 평균 제곱 차)을 비교하였다. 피어슨 상관관계가 정확도와 범위 면에서 성능이 좋은 것으로 결과를 보였다[9].

음악 추천 시스템인 Ringo[9]에서는 수식 (10)을 사용하여 유사도 가중치를 계산하였으며, 7점제의 중간 값

인 4를 수식에 사용하여, 한계치를 높여 정확도를 높이려 노력하였으나 피어슨 상관관계 방법에 비해 효율적이지 못한 것으로 발표되어 있다[16].

$$w_{a,u}=\frac{\sum_{i=1}^{m}(r_{a,i}-4)*(r_{u,i}-4)}{\sigma_a*\sigma_u} \quad (10)$$

Sarwar는 유사도 계산을 사용자가 아닌 항목을 기준으로 실험을 하였다. 사용자는 급속히 증가하나, 항목은 크게 증가하지 않는다는 가정 하에서 항목을 비교하는 것이 계산 시간을 축소할 수 있으며, 계산 결과도 좋다는 결과를 도출하였다[17].

피어슨 상관관계를 근간으로 하는 방식은 여러 분야에서 좋은 결과를 나타냈으나, 가장 큰 문제점은 희소성으로 인해서 발생하는 예측의 정확도 저하이다. 희소성이 높아지면 예측의 정확도가 크게 낮아진다.

2.3 유사도 개선 관련 기법

희소성으로 인하여 발생되는 유사도 계산의 정확도를 개선하기 위한 방법은 크게 두 가지로 구분할 수 있다. 첫째는 희소한 문제 영역 자체를 축소하여 유사도 계산에 이용되는 불분명한 자료를 제거하는 방법이며, 두 번째는 문제 영역을 가용한 정보를 이용하여 채워 유사도 계산의 정확도를 향상시키는 방법이다.

문제 영역 자체를 축소하는 방식으로는 SVD(Singular Value Decomposition)가 많이 사용되고 있다[18]. SVD는 m x n 크기의 R 행렬을 m x r, r x n의 크기를 가지는 2개의 직교(orthogonal) 행렬 U, V를 포함하는 3개의 행렬로 분해하는 기술이다.

$$R=U\cdot S\cdot V' \quad (11)$$

S는 r x r의 크기를 갖는 대각 행렬이다. r x r 대각행렬을 k < r인 k x k 대각행렬을 얻기 위하여 가장 큰 대각선상의 값을 갖도록 행렬 S를 축소한다. 이를 근간으로 행렬 R에 가장 근접한 rank-k 행렬인

$$R_k=U_k\cdot S_k\cdot V_k' \quad (12)$$

을 구할 수 있다. SVD는 대상 행렬을 축소하여 계산 공간이 축소되므로, 계산 시간을 줄일 수 있는 장점이 있다. 대상 자료중의 학습자료(train data)와 실험자료(test data)의 비율이 0.5보다 작은 경우 협업 여과 추천 방식보다 효과적이나, 학습 자료와 실험 자료의 비율이 0.5보다 높아지면, 즉 희소성이 개선되면 유사도를 이용하는 협업 여과 추천 방식이 효과적인 것으로 발표되었다[18].

Good는 문제 영역의 희소공간을 축소하기 위하여 자동으로 점수를 생성하는 내용기반 소프트웨어 에이전트의 사용을 제안하였다. 새로운 항목이 추가되면 이 항목에 대한 평가를 시스템에 저장된 내용을 기반으로 에이전트가 자동으로 점수를 부여하는 방식이다. 에이전트는 개인의 프로화일을 기반으로 학습을 통하여 개인별로 또는 소그룹별로 새로운 항목에 대한 점수를 부여한다. 에이전트가 하나보다는 여러 개를 배치하는 것이 효과적이었으며, 사용자와 함께 에이전트를 같이 사용하는 것이 보다 효과적이라는 결론을 도출하였다. 그러나 에이전트의 수에 따라 결과가 달라질 수 있으며, 유사도가 커질수록 상대적으로 효과가 작아지는 단점이 있다[19].

Melville은 희소공간을 내용기반 자료를 활용하여 모두 채우고, 이를 기반으로 예측을 생성하는 방식을 제안하였다. web crawler를 이용하여, 영화제목, 제작자, 배우, 장르, 영화 줄거리, 주제어, 사용자 비평, 외부 관전평, 뉴스그룹 견해등의 정보를 Naive Baysian Text 분류를 이용하여, 군집화하고, 이를 기반으로 희소공간을 가상 점수로 모두 채운다. 실 평가 점수에 가중치를 주어, 사용자가 실제 부여한 점수에 우선권을 주었으며, 기존의 예측 식을 일부 변형하였다. 피어슨 상관관계를 기반으로 하는 협업 여과 추천에 비하여 평균 절대 오차는 4%, 반응자 작용 특성의 민감도는 4.6% 향상되었음을 발표하였다[14]. 이 경우 내용기반 추천과 협업여과 추천을 상호보완하는 장점은 있으나, 두 가지 방식이 다른 형태로 구성되어 있어 web crawler의 도입 등 시스템 구성이 복잡해진다.

사용자에게 점수 부여를 요청함에 따라 발생되는 희소성에 대한 문제의 다른 해결 방법으로 사용자의 접속 시간, 접속빈도, 클릭 횟수, 출력 여부 등 시스템 측면에서의 취득 가능한 정보를 활용하는 방안이 다양하게 연구되고 있다. 정보 취득에 시간이 오래 걸리며, 점수 부여 방식보다 정확도가 낮은 단점이 있다[20].

3. 인구 통계 정보를 이용한 협업 여과 추천의 유사도 개선 기법

여러 다양한 방법을 이용하여 추천의 정확도를 높이기 위하여 노력하였으나 크게 개선하지는 못하였다. 학습 자료가 적은 경우, SVD 방식이 피어슨 상관관계 방식보다 우수하나, 학습 자료가 많은 경우에는 피어슨 상관관계 방식이 보다 우수하다. SVD는 계산속도 향상의 부가적인 효과를 거둘 수는 있었으나 결과적으로 정확도가 크게 향상되지 못하였다[18]. 따라서 이 논문에서는 자료의 많고 적음에 모두 적용할 수 있으며, 희소성으로 인한 유사도 계산의 문제를 극복할 수 있는 새로운 추천시스템을 제안한다. 일부 에이전트 등을 이용하여 희소성을 줄여 유사도 계산을 향상시키려는 노력이 있었으나, 에이전트를 이용하는 경우는 희소성보다는 초기화 문제를 해결키 위한 방편으로 이용된 경우가 많다[21]. 희소 공간을 내용기반 정보를 이용하여 모두 채

우는 경우, 추천을 원하는 사용자의 유사도 계산시, 사용자가 평가하지 않은 항목까지 유사도 계산에 적용되며, 내용기반 추천과 협업 여과 추천 두 가지를 모두 수행하여야 하므로 계산시간이 상대적으로 많이 소요된다.

본 논문에서는 melville[14]이 내용기반 정보를 이용하여 행렬 전체를 채워 유사도 계산을 하는 방식과 다르게 인구통계정보를 활용하였으며, 협업 여과 추천의 유사도 계산시 추천을 요청한 사용자의 평가는 모두 반영하고, 상대되는 사용자의 점수가 없는 경우에 한하여 가상 평가 점수를 반영하여 계산을 줄이는 방식을 제안하고자 한다.

그림 1의 행은 항목의 종류(m)를, 열은 사용자(n)을 나타내며, 이는 m x n 행렬을 구성하게 된다. 유사도 계산은 상대성이 있어서, 상대되는 항목이 없으면 계산에서 제외가 된다. i번째 사용자의 m번째 항목의 추천을 하는 경우 j번째 사용자와의 유사도 계산에서 u번째, m-1번째 항목은 j번째 사용자의 정보가 없어 제외되게 된다. 이 경우 2번째 항목의 평가만으로 유사도를 계산하게 되어 정확하게 유사한 이웃인지가 확인되지 않는다. 따라서 정확한 계산을 위해서는 사용자와 이웃간의 상호성을 보장하여 주어야 한다. 즉 사용자 자료가 계산에 이용되는 경우 이웃의 자료도 함께 존재하여야 계산 결과는 보다 정확해질 것이다. 사용자의 자료는 있으나 이웃의 자료가 없는 경우 가상 평가 값을 보완하고 계산을 수행하면 계산의 정확도가 높아질 것이므로 이웃의 자료를 보완하는 방법을 통하여 추천 성능을 향상시키고자 한다. 빈 자리를 채우는 가상 평가 값을 어떤 항목을 기반으로 하느냐에 따라 계산 결과는 달라질 수 있다.

이웃의 인구 통계 정보를 이용하여, 빈 자리에 가상 평가 값을 추가하여, 유사도 계산의 정확도를 향상시키

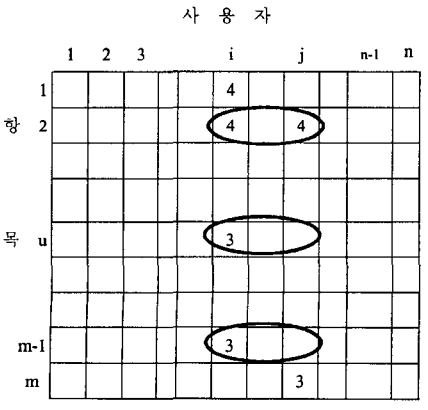

그림 1 사용자-항목 행렬 구성

도록 하였으며, 이는 예측 계산의 정확도를 향상시키는 기반이 된다. 인구 통계 정보를 이용한 추천은 사용자의 성별, 나이, 직업과 같은 인구 통계학적 요소를 이용하여 사용자의 유형을 분석하여 추천을 하는 방식이다. 단순한 방식이나 초기 구축 단계나 처음 방문자의 경우에도 적용이 가능한 장점이 있다.

먼저 평가를 마친 사용자의 자료를 기초로 Train-Matrix를 생성하고, 사용자가 원하는 대상에 대한 유사도 계산을 위하여 Train-Matrix를 사용자 및 이웃의 정보를 근간으로 인구 통계 정보를 추가로 반영한다. 인구 통계 정보가 반영된 정보를 기반으로 각 사용자의 유사도를 계산한다. 계산된 유사도를 기반으로 사용자가 원하는 대상의 평가 예측치를 계산한다. 계산된 예측치와 실측치와의 차이를 계산하여 평가의 정확도를 비교한다.

문제 영역이 커지면 유사도 계산은 실시간으로 처리하기에는 적합지 않으므로 유사도 계산까지의 과정은

알고리즘 1 인구 통계 정보 기반 추천 정확도 향상

```
//Create environment (off-line-통계정보 생성등)
BEGIN
open Train_data_file
  while (end-of-Train_data_file)
  {
    read 학습자료(Train_data)
    train[user][movie] = rating;
  }
close Train_data_file
calculate V_Pa;
for(int i = 1;maxuser;i++)
{
 for(int j = 1;maxmovie;j++)
 {
   If (사용자 점수 ≠ null) and (이웃 점수 = null)
     then
       calculate S_u,i reference V_Pa;
       r_u,i = s_u,i;
   End if;
 }
 r_a,r_u 평균 계산
}
for(int i = 1;maxuser;i++)
{
for(int j = 1;maxmovie;j++)
  calculate W_a,u // similarity using pearson correlation
  coefficient;
}
END
//Calculation estimate(on-line)
BEGIN
for(int k = 1;test_data수;k++)
 {
  While (유사도 >= 한계치 or 근접 이웃 수<= 설정치)
  Calculate P_a,i
 }
Calculate mean absolute error
END
```

오프라인에서 계산을 수행하는 것이 효율적이며, 실험 자료가 반영되는 예측계산 부분은 실시간으로 처리가 가능한 부분이다.

본 논문에서는 인구 통계 정보를 일종의 군집화를 통하여, 이웃군의 평균을 형성하고, 유사도 계산 시 필요한 이웃의 빈 평가자리를 채우기 위해 가상 평가값에 이 이웃군의 평균값을 이용하는 방식이다. 그림 1의 j번째 이웃의 인구 통계 정보를 기반으로 u번째, m-1번째 항목의 j번째 이웃의 빈자리를 채우고 유사도를 계산하여, 보다 정확도 높은 유사도 계산을 유도하였다. 다음과 같이 이웃군을 이용하여 가상 평가값을 구성한다.

인구 통계 정보가 k개의 속성을 가진다면,

$P = \{P1, P2, \cdots, Pk\}$

예를 들어 k = 3인 경우

$P = \{P1, P2, P3\}$이다. 여기에서 속성 Pj가 Sj개의 다른 속성값을 가진다면,

$Pj = \{q1, q2, \cdots, qs_j\}$로 표현된다.

속성 Pj를 n개씩 군집화 할 경우 다음과 같은 형태를 갖는다.

$Pj_1 = \{q_1, q_2, \cdots, q_n\}$

$Pj_2 = \{q_{n+1}, q_{n+2}, q_{2n}\}$

.

.

$Pj_m = \{q_{m+1}, q_{m+2}, q_{mn}\}$

...

이 군집의 대표값(평균) V_{P_n}는 첫 번째 항목을 선택한 경우

$$V_{P1_1} = \frac{\sum_{u=1}^{N} r_{u,1}}{n_{P1_1}}$$

where $r_{u,1} \in P1_1$의 속성을 가진 평가자의 평가점수,

n_{P1_1}은 $P1_1$의 속성을 가진 평가자의 수

로 나타난다.

예를 들어 P = {age, gender, occupation}라면 P2 = gender이고, P2 = {남,여}로 구성된다. 이 경우 $P2_1 = \{남\}$, $P2_2 = \{여\}$이고, 첫 번째 항목의 대표값 V_{P2_1}은 첫 번째 항목을 평가한 남자들의 평균값을 의미한다.

사용자 u가 P1의 첫 번째 군에 속하고, P2의 두 번째 속성에 속하며, P3의 첫 번째 속성에 속하는 경우 사용자 u의 속성값인 C_u는 $\{V_{P1_1}, V_{P2_2}, V_{P3_1}\}$로 구성된다. 여러 개의 속성 중 각 개인의 특성에 따라 평가에 반영되는 정도의 차이를 고려하여, 빈 자리에 적용하는 가상 평가값은 다음과 같이 정의하였다.

$$S_{u,i} = E_1 * V_{P1_1} + E_2 * V_{P2_2} + E_3 * V_{P3_1} \quad (13)$$

where $\sum_{j=1}^{k} E_j = 1, \quad 0 \le \forall E_j \le 1$

u는 대상이웃, i는 대상 항목, E_j는 가중치

이 경우에도 학습 자료가 적으면 속성별 군집 대푯값 V_{Pj}에서 희소성이 발생하게 된다. 이 경우 $\overline{r_{.,i}}$를 $S_{u,i}$로 대체하였다.

유사도 계산을 위한 피어슨 상관관계식 (6)에서 $r_{a,i}$는 있으나, $r_{u,i}$가 없는 경우 $r_{u,i}$는 가상 평가값인 $S_{u,i}$ 값으로 대체된다.

E_j는 오프라인에서 계산되며, 사용자의 특성이 어떤 속성에 근접한지를 실험을 통하여 계산하였다. 즉 사용자의 속성 각각을 반영하여 계산한 예측치와 실 평가치와의 차이를 비교하여 가중치에 반영하였다. 예를 들어 어떤 사용자가 10회 평가한 결과와 예측치의 결과가 특성 P1을 적용한 경우 다른 특성을 고려한 경우 보다 3회 정확하였고, P2를 적용한 경우 2회, P3를 적용한 경우 5회 정확하였다면 가중치는 $E_1 = 0.3$, $E_2 = 0.2$, $E_3 = 0.5$로 정의하였다.

4. 실험 및 평가

실험은 객관성과 공정성을 위해, 널리 이용되고 있는 GroupLens Research Project[22]에서 제공한 MovieLens 데이타 집합을 이용하여 실험을 하였다. 제공된 데이타 집합은 1~5 사이의 점수로 평가된 100,000개의 영화 평점으로 최소 20개 이상의 영화를 평가한 943명의 사용자가 본 1,682개 영화를 대상으로 구성되어 있다. 사용자의 인구 통계 정보로는 나이, 성별, 직업, 우편번호 등이 포함되어 있으며, 영화는 19개의 장르로 구분되어 있고, 중복 장르 선택이 가능토록 되어있다. 총 100,000개의 데이타 집합 중에서 학습 자료로 80,000개의 정보가 실험 자료로 20,000개의 자료로 구분되어 있다.

인구 통계 정보는 GroupLens에서 제시한 user-id, age, gender, occupation, zip code 중에서 일부를 활용하였다. P = {age, gender, occupation}

사용자의 특성에 따른 가중치 계산은 학습자료 80,000개중 64,000개를 학습 자료로 사용하고 16,000개를 실험 자료로 사용하여 개인별 가중치를 오프라인에서 계산하였다.

그림 1에서 비어 있는 j번째 이웃의 u번째 영화 항목을 가상 평가값으로 대체하는 방법은 다음과 같다. j번째 이웃의 인구 통계 정보가 11살이고, 남자이고, 학생이라면, u번째 영화에 대한 11살 나이군의 평균값과 남자군의 평균값, 학생의 평균값을 (13)식을 이용하여 계산하고, 이 결과를 가상 평가값으로 사용한다. 각 영화

의 각 군집에 대한 평균값은 Train-Matrix 구성시 오프라인 작업으로 미리 구성한다.

본 논문에서는 실제 부여 점수와 예측 간의 차이 분석에 많이 사용되고 있는 평균오차(mean absolute error, MAE)방법을 사용하여 제안한 방법의 타당성을 검증하였다[17].

$$\text{평균오차}(MAE) = \frac{|R_1 - P_1| + \cdots + |R_n - P_n|}{n} \quad (14)$$

표 1에서 나타난 바와 같이 유사도의 한계치가 작거나 모든 이웃에 대해 수행한 결과는 피어슨 상관관계 방식보다 좋으나 유사도의 한계치가 큰 경우는 좋지 않은 것을 볼 수 있다. 이는 자료 빈곤으로 인해 유사도가 높게 평가되었던 사용자에게 가상 평가 자료를 반영시켜 나타난 결과이다. 예를 들자면 하나의 항목만을 비교하여 그 값이 같은 경우와 여러 항목을 비교하여 그 값이 같은 경우 유사도는 같으나 신뢰도는 크게 차이가 날 것이다.

표 2는 이웃의 수와 유사도를 기준으로 살펴본 것으로 제안된 방식이 피어슨 상관관계 방식보다 평균오차가 적음을 알 수 있다. 표 2에서 나타난 바와 같이 제안 시스템은 그 특성상 유사도 한계치보다는 이웃의 수에 반응 결과가 좋게 나타남을 알 수 있다. 이는 희소성 해소를 위하여 대상이 되는 이웃 사용자의 빈자리를 가상 평가값으로 채워주어 신뢰할 수 있는 이웃 사용자 수가 증가함에 따라 나타나는 현상이다.

그림 2는 근접 이웃의 수에 따른 변화 추이를 나타낸 곡선이다.

Grouplens에서 제시한 5개 군의 자료에 대하여 각각을 평가하고 평가결과를 평균하여 나타내었다. 피어슨 상관관계 방식의 경우 근접 이웃의 수 N = 60인 경우 평균오차가 0.765로 가장 좋은 것으로 나와 있으나[16], 본 실험에서는 인구 통계 정보를 반영하여, N= 50인 경우 평균오차가 0.693로 나타나 제안된 방법이 효율적임을 확인하였다. 20,000건의 학습 자료를 근간으로 한 실험에서도 80,000건의 학습 자료인 경우의 제안된 결과보다는 높으나, 상대적으로 결과가 낮아져 학습 자료가 적은 경우에도 활용이 가능할 것으로 보인다.

희소성 문제 영역 자체를 축소시키는 SVD를 이용한 경우와 비교하여 보면 표 3에 나타난 바와 같다. 학습 자료와 실험 자료의 비율에 따라 비율이 0.5보다 작은 경우(Train Data 20,000건) 피어슨 상관관계 방식보다 SVD방식의 결과가 좋으며, 비율이 0.5보다 큰 경우 피어슨 상관관계 방식이 우수하나[18], 제안된 방식은 학습 자료의 크기에 관계없이 모든 경우 결과가 좋았다.

그림 3은 5개 군의 자료 중 1개 군에 대하여 인구 통계 정보인 나이, 성별, 직업을 각각 반영한 경우와 3항

표 1 유사도 한계치에 따른 평균오차

구분	평균오차(MAE)	
	피어슨 방식	제안 방식
모든이웃	0.77606	0.70791
유사도=0.1	0.77046	0.75197
유사도=0.2	0.79202	0.8182
유사도=0.3	0.75777	0.82564
유사도=0.4	0.78100	0.82491
유사도=0.5	0.76242	0.8249

표 2 유사도 한계치와 이웃의 수에 따른 평균오차

구 분	평균오차(MAE)	
	피어슨 방식	제안 방식
유사도=0.1 이웃=20	0.763836	0.75383
유사도=0.1 이웃=40	0.765492	0.75187
이웃=20	0.765686	0.70951

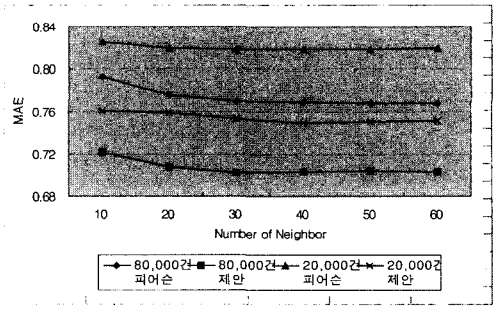

그림 2 학습 자료 수에 따른 실험 결과 비교

표 3 학습 자료 수에 따른 비교

구분	MAE	
	Train Data 20,000 건	Train Data 80,000 건
피어슨 방식	0.8188	0.7678
SVD[18]	0.7501	0.79
제안 방식	0.749	0.7010

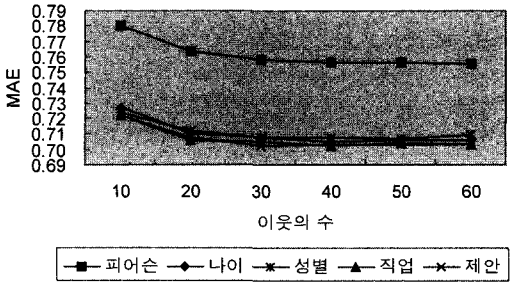

그림 3 인구 통계 정보의 반영에 따른 실험 결과

표 4 인구 통계 정보를 반영한 실험 결과(이웃의 수 n = 50인 경우)

구분	MAE	ROC-4			
		Sensitivity	Specificity	Accuracy	Error rate
피어슨방식	0.7678	0.7002	0.296	0.523	0.477
나이반영	0.7060	0.769	0.344	0.583	0.417
성별반영	0.7073	0.79	0.372	0.607	0.393
직업반영	0.70465	0.7671	0.339	0.58	0.42
제안	0.701	0.8096	0.392	0.627	0.373

목을 함께 반영한 제안 시스템의 경우를 나타낸 것이다. 어느 경우에도 결과치가 피어슨 상관관계 방식보다 좋음을 알 수 있다. 즉 희소성을 감소시켜 주면 효과가 좋아짐을 알 수 있다.

표 4는 ROC(Receiver Operating Character)[19] 측정에 대한 비교이다. ROC-4에서는 사용자의 rating 정보 중 4, 5의 값은 좋은 값(positive)으로, 1, 2, 3의 값은 나쁜 값(negative)으로 정의한다. Sensitivity는 임의로 선택된 평가값이 좋은 값으로 추천될 확률로, 그 값이 1인 경우 완벽한 경우이며, 0.5인 경우 무작위(random)로 판별한다[19]. Specificity는 임의로 선택된 평가값이 나쁜 값이 추천되지 않을 확률이다. Accuracy는 전체 실험 자료 중 예측이 맞은 경우의 확률이며, Error rate는 전체 실험 자료 중 예측이 틀린 경우의 확률이다.

인구 통계 정보를 반영한 경우 기대 했던 바와 같이 정확도가 좋아짐을 확인할 수 있었다. 실험 환경은 다르나 Melville[14]의 경우 평균 절대 오차의 경우 4%, 반응자 작용 특성의 민감도의 경우 4.3% 정도 향상된 것으로 발표하였으나, 본 논문에서 제안한 방식은 평균 절대 오차의 경우 9%, 반응자 작용 특성의 민감도의 경우 13% 정도 향상된 것으로 나타났다. 다른 속성에 비해 성별을 반영한 경우 ROC-4에서 좋은 결과를 보인 것은 ROC-4의 계산 근거인 true-positive와 false-negative가 다른 특성에 비해 높게 나타났기 때문이다. 표 4의 결과에서도 알 수 있듯이 어떤 인구 통계 정보를 적용할 것인가를 선택하는 작업은 문제 영역에 종속적이어서 매우 어렵다. 인구 통계 정보와 사용자 점수와의 상관관계를 사전에 정의할 수 있다면 해당 영역의 인구 통계 정보 구성시 반영하여, 보다 빠르게 사용자의 유사도 계산이 가능할 것이다. 인구 통계 정보는 예측에 영향을 미칠 수 있는 좋은 자료이므로 이 부문에 대한 보다 심도 있는 연구가 필요하다.

5. 결 론

제안된 추천기법은 협업 여과 추천 기법의 희소성으로 인한 유사도 계산의 정확도를 향상시키기 위하여 기존에 주로 이용된 문제 영역의 축소 방식이나, 계산식의 수정이 아닌 희소성의 근본적인 문제점인 빈자리를 인구 통계 정보를 이용하여 가상 평균값을 채워줌으로써 유사도 계산의 정확성을 보완하여 평가 예측 계산 결과를 향상시킬 수 있는 방법을 제안하였다. 상관관계를 이용하는 유사도 계산식은 비교 대상의 상호성에 대한 자료가 확보되어 있는 경우 보다 정확한 결과를 얻을 수 있다. 따라서 사용자의 특성을 반영한 가중치를 적용하고, 상호성이 보장될 수 있도록 가상 평균값을 활용하여 유사도 계산 정확도를 높여, 결과적으로 예측 결과를 높일 수 있음을 확인하였다. 평균 절대 오차의 경우 9%, 반응자 작용 특성의 민감도의 경우 13% 정도 향상된 것으로 나타났다. 또한 첫 번의 추천을 이용하는 사용자의 평점 결과가 없는 상태에서도 인구 통계 정보를 참조하여 직접적으로 평가값으로 활용하여 추천이 가능하므로 처음 사용자에 대한 초기화에 대한 문제도 해결이 가능하다. 제안한 기법의 타당성을 보이기 위해 Grouplens 프로젝트에서 제공한 100,000개의 영화 데이타 집합을 이용하여 실험하였으며 보다 좋은 실험 결과를 도출하여, 인구 통계 정보를 보완하면 보다 높은 평가 예측의 가능성이 있음을 확인하였다.

참 고 문 헌

[1] Miller,B., Riedl,J. and Konstan,J., "Experiences with GroupLens:Making Usenet useful again," Proc. of the 1997 Usenix Winter Technical Conference, pp.219-231, 1997.

[2] Ansari,A., Essegaier,S. and RKohli,R., "Internet Recommendation Systems," Journal of Marketing Reserch Vol.37, pp. 363-375, 2000.

[3] Il Im, "Augmenting Knowledge Reuse Using Collaborative Filtering Systems," A Dissertation Presented to the faculty of the graduate school USC (Information Systems), p.191, 2001.

[4] Basu,C., Hirsh,H. and Cohen,W, "Recommendation as Classification : Using Social and Content-based Information in Recommendation," Proc. of the Fifteenth National Conference on Artificial Intelligence(AAAI-98), pp.714-720, 1998.

[5] Pazzani,M., "A Framework for Collaborative, Content-Based and Demographic Filtering," Artificial Intelligent Review13(5-6), pp.393-408, 1999.

[6] 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.
[7] Konstan,J., Miller,B., Maltz,D., Herlocker,J., Gordon,K. and Riedl,J. "GroupLens :Applying Collaborative Filtering to Usenet News," Communications of the ACM, Vol.40 No.3, pp.77-87, 1997.
[8] Rensnick,P., Iacovou,N., Suchak,M., Nergstorm,P. and Riedl.,J. "GroupLens : An Open Architecture for Collaborative Filtering of Netnews," Proc. of CSCW '94, pp. 175-186, 1994.
[9] Shardanand,U. and Maes,P., "Social information filtering : Algorithms for automating 'word of mouth'," Proc. of ACM CHI '95 Conference on Human Factors in Computing Systems, pp.210-217, 1995.
[10] Breese,J., Heckerman,D. and Kadie,C., "Empirical Analysis of Prediction Algorithms for Collaborative Filtering," Proc. of the 14th Conference on Uncertainly in Artificial Intelligence, pp.43-52, 1998.
[11] O'conner,M. and Herlocker.J, "Clustering Items for Collaborative Filtering," ACM SIGIR '99, http://www.csee.umbc.edu/~ian/sigir99-rec/, 1999.
[12] Hill,W., Stead,L., Rosenstein,M. and Furnas,G., "Recommending and Evaluating Choices in a Virtual Community of Use," Proc. of CHI '95 Conference on Human Factors in Computing Systems, pp.194-201, 1995.
[13] Goldberg,K., Roeder,T, Gupta,D. and Perkins,C. "Eigentaste: A Constant Time Collaborative Filtering Algorithms," Information Retrieval Vol.4, No.2, pp.133-151, 2001.
[14] Melville,P., Mooney,R. and Nagarajan,R., "Content-Boosted Collaborative Filtering for Improved Recommendations," Proc. of the 8th National Coference on Artificial Intelligence(AAAI-2002), pp.187-192, 2002.
[15] Andrew,I., Popescul,A. and Ungar,L., "Methods and Metrics for Cold-Start Recommendations," Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.253-260, 2002.
[16] Herlocker,j., Konstan,J., Borchers,A. and Riedl,J., "An Algorithmic Framework for Performing Collaboraive Filtering," Proc. of the 1999 Conference on Research and Development in Information Retrieval. ACM Press, NY, pp.203-237, 1999.
[17] Sarwar,B., Karypis,G., Konstan,J. and Riedl,J., "Item based collaborative filtering recommendation algorithms," Proc. of the 10th International World Wide Conference, pp.285-295, 2001.
[18] Sarwar,B., Karypis,G., Konstan,J. and Riedl,J. "Application of Dimensionality Reduction in Recommendation System-A Case Study," ACM WebKDD 2000 Web Mining for E-Commerce Workshop, http://robotics.stanford.edu/~ronnyk/WEBKDD2000/papers/, 2000.
[19] Good,N., Schafer,B., Konstan,J., Borchers,A. Sarwar,B., Herlocker,J. and Riedle,J., "Combining Collaborative Filtering with Personal Agents for Better Recommendation," Proc. of the AAAI conference. pp.439-446, 1999.
[20] Claypool,M., Brown,D., Phong,L. and Waseda,M. "Infering User Interest," Computer Science Technical Report Series WPI-CS-TR-01-07, p.23, Worcester Polytechnic Institute, 2001.
[21] Sarwar,B., Karypis,G., Konstan,J. and Riedl,J., "Getting to Know you :Learning New User Preferences in Recommender System for Groups of Users," Proc. of the 7th International conference on Intelligent user interfaces, pp.127-134, 2002.
[22] http://www.grouplens.org

이 용 준
1982년 인하대학교 전자계산학과(이학사) 1995년 인하대학교 전자계산학과(공학석사). 2001년 인하대학교 컴퓨터공학부 박사 수료. 1982년~2000년 한국전기연구원 선임연구원. 1996년~1999년 한국전기연구원 전산실장. 2000년~현재 한국전기연구원 전기시험연구소 선임기술원. 관심분야는 e-Learning, 소프트웨어공학, 분산객체컴퓨팅, DAS

이 세 훈
1985년 인하대학교 전자계산학과(이학사) 1987년 인하대학교 전자계산학과(이학석사). 1996년 인하대학교 전자계산공학과(공학박사). 1987년~1990년 해병대 분석장교. 1990년~1993년 (주)비트컴퓨터 기술연구소 선임연구원. 1999년 5월 멀티미디어기술사. 2001년~2002년 미국 뉴저지 공과대학(NJIT) 교환교수. 1993년~현재 인하공업전문대학 컴퓨터정보공학부 교수. 관심분야는 e-Learning, 모바일컴퓨팅, 소프트웨어공학, XML/JAVA

왕 창 종
1964년 고려대학교 물리학과(이학사) 1975년 성균관대학교 대학원. 1985년~1992년 한국정보과학회 이사회 임원 1993년 한국정보과학회 부회장. 1979년~2003년 2월 인하대학교 공과대학 컴퓨터공학부 교수. 관심분야는 소프트웨어공학, 분산객체기술, 컴퓨터기반교육