고객 맞춤 서비스를 위한 HPPS(Hybrid Preference Prediction System) 설계

정은희^{*}, 이병관^{**}

요 약

본 논문에서는 고객 맞춤 서비스의 선호도를 정확하게 예측하기 위하여 사용자 프로파일 분석, 사용자간 유사도 분석을 이용한 HPPS(Hybrid Preference Prediction System) 설계를 제안한다. 기존의 NBCFA(Neighborhood Based Collaborative Filtering Algorithm)과 달리, 본 논문은 첫째, 선호도 예측식에서 이웃의 상품 평가가 없을 경우 상품에 대한 평균값을 이용하도록 하였고, 둘째, 선호도 예측식에서 사용자의 특성을 분석한 가중치를 반영하도록 하였고, 끝으로, 인접 이웃을 선정할 때 유사도, 상품 평가 여부, 평가 횟수를 반영하여 HPPS에 선호도의 정확도를 향상시켰다. 따라서 첫째와 둘째의 선호도 예측식을 이용하면 HPPS의 정확도는 기존의 NBCFA에 비해 97.24% 향상되었고, 인접이웃 선정방식에서도 HPPS 시스템의 정확도가 75% 향상되었다.

A Design of HPPS(Hybrid Preference Prediction System) for Customer-Tailored Service

Eun Hee Jeong^{*}, Byung Kwan Lee^{**}

ABSTRACT

This paper proposes a HPPS(Hybrid Preference Prediction System) design using the analysis of user profile and of the similarity among users precisely to predict the preference for custom-tailored service. Contrary to the existing NBCFA(Neighborhood Based Collaborative Filtering Algorithm), this paper is designed using these following rules. First, if there is no neighbor's commodity rating value in a preference prediction formula, this formula uses the rating average value for a commodity. Second, this formula reflects the weighting value through the analysis of a user's characteristics. Finally, when the nearest neighbor is selected, we consider the similarity, the commodity rating, and the rating frequency. Therefore, the first and second preference prediction formula made HPPS improve the precision by 97.24%, and the nearest neighbor selection method made HPPS improve the precision by 75%, compared with the existing NBCFA.

Key words: HPPS(혼합선호예측시스템), Customer-Tailored Service(고객맞춤서비스), Nearest Neighbor Selection(인접이웃선정), Similarity(유사도), Weight(가중치)

1. 서 론

초고속 인터넷 인프라의 확산과 인터넷의 대중화

와 더불어 전자상거래 또한 빠르게 성장하여 왔으며, 영화나 음악 서비스 같은 다양한 형태의 인터넷 서비스가 급성장하고 있다. 특히, Web2.0, UCC(User

※ 교신저자(Corresponding Author): 정은희, 주소: 강원도 삼척시 중앙로 1 (245-711), 전화: 033)570-6646, FAX : 033)574-6640, E-mail: jeongeh@kangwon.ac.kr
접수일: 2011년 8월 23일, 수정일: 2011년 9월 26일

완료일: 2011년 10월 24일

^{*} 정회원, 강원대학교 지역경제학과 부교수

^{**} 정회원, 관동대학교 컴퓨터학과 교수
(E-mail: bklee@kwandong.ac.kr)

Create Contents) 등은 집단 지성 특성을 이용하여 사용자들의 다양한 의견과 지식을 생성, 공유, 소비하는 지적 흐름을 나타내고 있으며, Web2.0 서비스가 늘어나면서 전자상거래 기업들은 더욱 치열해진 경쟁에서 생존하기 위해 다양한 마케팅 전략을 구사하여야 하고, 고객들도 기존의 서비스와 차별화된 서비스를 원하고 있다[1,2]. 따라서 고객의 취향이나 관심에 초점을 맞춘 서비스를 제공 해주는 고객 맞춤 서비스에 대한 관심과 중요성이 날로 증가하고 있다.

하지만, 이러한 고객 맞춤 서비스를 제공하기 위해서는 사용자가 명백히 드러나는 사용자 질의 히스토리, 브라우징 히스토리, 웹 커뮤니티, 그리고 클라이언트 측에서의 상호작용이나 클릭 정보 분석 등이 필요하며, 더 높은 질적 수준을 가진 고객 맞춤 서비스를 제공하기 위해서는 사용자의 숨겨진 의도를 추론해내는 알고리즘이 개발되어야 실질적인 고객 맞춤 서비스가 가능하다고 볼 수 있다[3].

본 논문에서는 이러한 실질적인 고객 맞춤 서비스가 가능하도록 고객의 선호도를 파악하는데 필요한 HPPS(Hybrid Preference Prediction System)를 설계한다. 이 HPPS는 사용자 선호도와 특성을 분석하여 사용자에게 대한 프로파일을 작성하고, 사용자와 이웃들 간의 유사도, 상품 평가여부, 상품 평가 횟수를 기준으로 인접이웃(Nearest Neighbor)을 선정한다. 그리고 이웃들의 제품 선호 성향을 분석하여, 이 정보로 선호도 예측값의 가중치로 활용하여 특정 상품에 대한 사용자의 선호도 예측값을 좀 더 정확하게 추정하여 추천 리스트를 작성한다. 따라서 일반적인 추천 시스템이 사용자와 추천 제품 간의 상호 관계만을 표현하는 반면, 본 논문에서 제안하는 HPPS는 사용자의 개인별 특성을 분석하여, 그 특성 정보를 바탕으로 추천 리스트를 작성하였으므로, 이때 디스

플레이 된 정보는 사용자의 선호도에 맞고 신뢰성이 높은 최적의 정보라 할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 추천시스템에 관련된 관련연구를 살펴보고 3장에서는 본 논문에서 제안하는 HPPS 설계에 대해 설명한다. 그리고 4장에서는 본 논문에서 제안한 시스템을 실험 및 평가하고 5장에서 결론을 맺는다.

2. 관련연구

2.1 추천시스템

추천 시스템은 자동화된 정보 필터링 기술을 적용하여 고객의 취향에 맞는 상품을 추천하여 고객의 구매 결정을 도와주는 시스템으로 가장 중요한 것은 고객의 선호도를 정확하게 분석하고 정제하여 정확한 예측 능력으로 고객이 원하는 가장 적절한 상품을 추천해 줄 수 있는 능력이다. 이를 위해서는 데이터 마이닝, 패턴인식, 정보 필터링 등 다양한 기법들이 적용될 수 있으며, 추천 접근 방식에 따라 내용 기반(Content-based), 협업(Collaborative)방식, 혼합(Hybrid) 방식으로 나눌 수 있으며, 각 방식에 대한 설명은 표 1에서 하고 있다[1,4].

본 논문에 제안한 HPPS에서는 추천 접근 방식 중에서 혼합방식을 사용하고 있다.

2.2 이웃기반 협업 필터링 알고리즘

이웃기반 협업 필터링 알고리즘은 Xerox Palo Alto 연구소에서 메일의 분류를 위한 실험 시스템인 Tapestry에 의해 소개되었다[5]. 그리고 GroupLens에서 인터넷을 기반으로 형성된 토론 시스템인 유즈넷 뉴스 그룹의 기사를 추천하기 위해 최초의 자동화

표 1. 추천 접근 방식 비교

추천 접근 방식	아이디어와 관련기술	한계성 및 주요 이슈
내용기반 방식	고객의 과거 선호 아이템의 유사 속성을 이용함 - User Profiling, Clustering	새로운 사용자 문제 확장성 문제, 성능 문제
협업 방식	집단지성을 바탕으로 유사 사용자 그룹의 의견을 이용함 - Nearest neighbor, Clustering	희박성, 확장성 문제 평가치가 없는 경우 새로운 아이템에 대한 문제
혼합 방식	내용기반과 협업기반 방식을 혼합 활용하거나, 사용자기반 협업 방식과 내용기반 협업 방법을 혼합 활용함.	

된 협업 필터링 알고리즘인 이웃기반 협업 필터링 알고리즘(Neighborhood Based Collaborative Filtering Algorithm, 이하 NBCFA)을 제안하였다[6].

NBCFA에선 고객 간의 선호도 유사 정도를 나타내기 위한 유사도는 피어슨의 상관계수를 사용하였으며, 이 유사도를 이용해 사용자가 상품에 대해 어떤 선호도를 가질 수 있는지를 예측하는 식은 (1)과 같다.

$$\hat{U}_i = \bar{U} + \frac{\sum_{J \in Rates} (J_i - \bar{J}) r_{UJ}}{\sum_{J \in Rates} |r_{UJ}|} \quad (1)$$

여기서, $\hat{U}_i$ 는 상품 i 에 대한 특정 고객 U 에 대한 예측값이고, $\bar{U}$ 는 특정 고객 U 가 선호도를 평가한 모든 상품에 대한 선호도 평균이고, J_i 는 이웃 고객 J 가 상품 i 에 평가한 실제 선호도 평가치 이고, $\bar{J}$ 는 이웃 고객 J 가 평가한 모든 상품에 대한 선호도 평균이다. 그리고 r_{UJ} 는 특정 고객 U 와 이웃 고객 J 의 유사도, Rates는 고객 U 와 동일한 상품에 대해 선호도를 표시한 고객들을 의미한다.

이웃기반 협업적 필터링 알고리즘에서는 사용자와 동일한 상품을 평가하는 이웃들만을 모아서 인접 이웃으로 선정하므로, 인접이웃이 사용자와 동일한 상품을 평가하였으므로 반드시 사용자와의 유사도가 높다고 단정 지을 수 없을 뿐만 아니라, 우연히 사용자와 동일한 상품을 평가한 사용자일 수도 있고, 선정된 인접이웃이 평가한 상품이 매우 적을 수도 있다는 단점을 가진다. 또한, 선정된 인접이웃이 상품 i 에 대해 반드시 평가를 하였다고 단정 지을 수 없다는 단점을 가진다.

본 논문에서는 인접이웃 선정의 단점을 해결하고자 유사도, 상품평가 여부, 평가 참여 횟수를 기준으로 선정하도록 설계하였고, 이때 선정된 인접 이웃 J 가 상품 i 에 평가하지 않은 경우에 상품 i 의 평가 평균을 이용하도록 설계하여 선호도 예측값을 좀 더 정확하게 측정하도록 하였다.

2.3 유사도 측정

NBCFA를 적용하기 위한 첫 번째 단계는 특정 상품에 대한 선호도 예측을 위해 사용자와 동일한 상품에 대한 평점을 내린 다른 사용자들의 프로파일과 사용자의 아이템을 비교하여 유사도를 구하는 것이다.

최초의 GroupLens 시스템에서는 피어슨 상관계수를 유사도로 이용하였다[6]. 아래의 식 (2)은 고객의 선호도 유사 정도를 나타내는 피어슨 상관계수를 설명한 것이다.

$$r_{UJ} = \frac{\sum_{i=1}^m (R_{Ui} - \bar{R}_U)(R_{Ji} - \bar{R}_J)}{\sqrt{\sum_{i=1}^m (R_{Ui} - \bar{R}_U)^2 \cdot \sum_{i=1}^m (R_{Ji} - \bar{R}_J)^2}} \quad (2)$$

여기서 r_{UJ} 는 추천을 받고자 하는 고객 U 와 그룹 내의 이웃 고객 J 가 선호도를 평가한 상품에 대한 피어슨 상관계수이며, 최대 1과 최소 -1 사이의 값을 갖는다. 그리고 m 은 고객 U 와 J 가 모두 평가한 상품의 전체의 수를 의미하고, R_{Ui} 는 고객 U 가 선호도를 평가한 상품 i 의 선호도 평가치 이고, $\bar{R}_U$ 는 고객 U 가 선호도를 평가한 상품들의 선호도 평가치 평균이며, R_{Ji} 는 이웃 고객 J 가 선호도를 평가한 상품 i 의 선호도 평가치 이고, $\bar{R}_J$ 는 이웃 고객 J 가 선호도를 평가한 상품들의 선호도 평가치의 평균을 의미한다.

본 논문에서도 피어슨의 상관계수를 이용해 고객 간의 선호도를 나타내는 유사도 측정에 이용하였다.

2.4 선호도 예측 정확도 평가

MAE(Mean Absolute Error)는 협업 필터링에 의한 예측값의 성능을 평가하기 위해 가장 일반적으로 적용되는 평가척도이다. MAE는 계산된 선호도 예측값과 실제 선호도 평가치의 절대 편차의 평균으로 계산되는데, MAE가 낮을수록 전체 예측 알고리즘의 정확도가 높다고 볼 수 있다[7,8].

$$MAE = \frac{1}{N} \sum_{j=1}^N |R_{Uj} - \hat{R}_{Uj}| \quad (3)$$

여기서 N 은 추천을 받을 모든 고객에 대한 예측의 총 개수를 나타내며, R_{Uj} 는 실제 선호도 평가치 이고, $\hat{R}_{Uj}$ 는 R_{Uj} 에 대응하는 예측값이다.

3. HPPS 설계

본 논문에서 설계한 고객 맞춤 서비스를 위한 HPPS(Hybrid Preference Prediction System)의 혼합 추천 선호도 예측 절차는 다음과 같다.

1단계 : 사용자의 컴퓨터를 통해 사용자의 기본적인

그림 1. HPPS의 전체적인 흐름도

인 정보를 입력받는다.

2단계: 입력된 사용자의 기본정보를 이용해 그동안 사용자가 작성한 추천 목록과 사용자의 특성으로 프로파일을 작성한다.

3단계: 사용자와 이웃들 간의 유사도를 피어슨의 상관계수를 이용해 계산하여 유사도를 계산하고, 유사도, 동일 상품 평가 여부, 평가 횟수를 기준으로 인접 이웃을 선정하고, 인접 이웃 크기를 설정한다.

4단계: 사용자의 특성과 이웃들의 특성을 분석하여 선호도 예측식의 가중치를 계산한다..

5단계: 본 논문에서 제안하는 선호도 예측식을 이용하여 사용자의 선호도 예측값을 계산하여 추천 목록을 작성한다.

6단계: MAE 평가를 통해 작성한 추천 목록의 적합성을 평가한다.

7단계: 사용자의 개인 특성에 맞는 추천 목록을 사용자에게 제공한다.

그림 1은 본 논문에서 제안하는 고객 맞춤 서비스를 위한 HPPS의 전체적인 흐름을 설명한 것이다.

3.1 사용자 프로파일 작성

사용자가 입력한 기본 정보를 데이터베이스에 저장하고, 사용자가 사용한 검색단어, 자주 찾는 사이트, 구매 제품 등을 사용자가 사용할 때마다 데이터베이스에 기록한다. 이렇게 기록된 정보를 이용하여 사용자의 행동을 분석해 사용자의 선호도를 측정하

고, 관심분야를 분류하여 사용자의 행동 패턴에 대한 프로파일을 작성하는데, 이 정보는 사용자의 개인 특성에 맞는 제품이나 정보 등을 사용자의 관심분야에 맞는 상품을 추천하고자 할 때, 유용하게 이용된다.

그림 2는 사용자 프로파일 작성 알고리즘의 한 예로 제품 구매 현황을 분석하는 것을 설명한 것으로 사용자의 동일한 제품 구매횟수가 HPPS가 정한 임계치보다 크다면 카운트를 증가시키고 선호 리스트에 등록하고, 그렇지 않으면 사용자의 후보배열에 등록한다. 이렇게 후보 배열에 등록된 상품들은 사용자가 동일한 제품을 구매할 때마다 제품에 대한 카운트

```

user_profile(item){
  item : 제품
  p_item : 선호하는 제품들이 저장된 배열
  d_p_item : HPPS가 정한 임계치보다
  작은 제품이 저장된 선호 후보 배열
  if(search_p_item(p_item, item)){
    // 선호제품이면
    update_count_p_item(item);
  }
  else{ // 선호제품이 아니면,
    if(search_p_item(d_p_item, item)){
      update_d_p_item(item);
      if(d_p_item(item,count)>= critical_value)
        add_p_item(item);
    }
    add_d_p_item(item);
  }
}

```

그림 2. 사용자 프로파일 작성 알고리즘 - 선호하는 제품 등록

가 증가되어 HPPS가 정한 임계치보다 커지면 선호 리스트에 등록되어 사용자의 선호도 예측값 계산식의 가중치(w_2)에 사용된다.

3.2 HPPS의 필터링 설계

3.2.1 가중치 설계

사용자의 선호도에 맞는 추천 리스트를 작성하기 위해, HPPS는 사용자의 선호도에 맞는 분야를 필터링해 디스플레이 한다. 그런데 본 논문에서 제안하는 HPPS는 피어슨 상관계수를 사용하는데, 이 피어슨의 상관계수는 모든 고객을 동일하게 취급하는 문제점을 가지고 있다. 즉, 사용자와 많은 상품에 대해 공통적으로 선호도를 표시한 다른 사용자와 한 두 개의 상품에만 공통으로 선호도를 나타낸 사용자와의 상관계수를 계산할 때 다르게 취급되어야 한다.

본 논문에서는 좀 더 정확한 선호도를 예측하기 위해, 사용자의 프로파일을 분석하여 그 결과와, 사용자와 인접 이웃의 평가의 일치 정도를 가중치로 사용하도록 설계하였으며, 가중치 계산 기준은 다음과 같다.

첫째, 사용자 x 와 인접이웃 그룹이 동일한 상품에 대해 평가 값이 일치하는 정도인 $w_1 = \frac{1}{k} \sum_{j=1}^k \left(\frac{r_j}{n_j} \right)$ 이다. 여기서 r_j 는 사용자 x, y 가 동일한 상품을 동일한 값으로 평가한 개수, n_j 는 사용자 x, y 가 동시에 평가하는 상품의 수, k 는 인접이웃 크기를 의미한다.

둘째, 사용자 x 의 프로파일 분석 정보와 인접이웃의 성향이 비슷한 정도인 $w_2 = \frac{1}{k} \sum_{j=1}^k f_j$ 이다. 여기서 f_j 는 사용자 x 의 선호 성향과 인접이웃의 선호 성향을 비교하여 동일한 성향비율, k 는 인접이웃 크기를 의미한다.

$$w = \frac{1}{k} \sum_{j=1}^k \left(\frac{r_j}{n_j} \right) + \frac{1}{k} \sum_{j=1}^k f_j \quad (4)$$

본 논문에서는 식 (4)의 가중치를 사용자의 선호도 예측 식에 이용하여 좀 더 정확한 선호도 예측값을 구하고자 한다.

3.2.2 선호도 예측식 설계

본 논문에서는 상품에 대한 선호도를 나타낸 사용자 U_x 와 유사도, 상품 평가여부, 평가 횟수의 조건에 맞는 사용자들을 인접이웃 U_y 으로 선정한 후, 상품

i 에 대한 사용자 U_x 의 선호도를 예측하는 선호도 예측식을 식 (5)과 같이 제안하였다.

$$\hat{U}_{x,i} = \begin{cases} \bar{U}_x + \frac{1}{k} \sum_{j=1}^k \left(\frac{r_j}{n_j} \right) + \frac{1}{k} \sum_{j=1}^k f_j + \left(\frac{\sum_{y \in (U_x \cap U_y)} (U_{y,i} - \bar{U}_y) R_{x,y}}{\sum_{y \in (U_x \cap U_y)} |R_{x,y}|} \right), & U_{y,i} \neq 0 \\ \bar{U}_x + \frac{1}{k} \sum_{j=1}^k \left(\frac{r_j}{n_j} \right) + \frac{1}{k} \sum_{j=1}^k f_j + \left(\frac{\sum_{y \in (U_x \cap U_y)} (r_i - \bar{U}_y) R_{x,y}}{\sum_{y \in (U_x \cap U_y)} |R_{x,y}|} \right), & U_{y,i} = 0 \end{cases} \quad (5)$$

본 논문에서 제안한 식 (5)에서는 선정된 인접이웃 U_y 의 상품 i 에 대해 평가를 하였다면, 인접이웃 U_y 의 평가값 $U_{y,i}$ 을 사용하지만, 인접이웃 U_y 이 상품 i 에 대해 평가하지 않았다면 평가값 $U_{y,i} = 0$ 이므로 이 값을 인접이웃들의 상품 i 의 평가 평균값($\bar{r}_i$)으로 대체하도록 설계하였다.

또한, 사용자의 프로파일 분석에 따라 사용자의 상품 성향에 맞는 인접 이웃의 유사도를 분석해 그 값을 가중치로 활용한다. 즉, 0과 1사이의 가중치를 이용함으로써 사용자의 상품 i 에 대한 선호도를 좀 더 정확하게 예측할 수 있다. 따라서 좀 더 개인 속성을 평가한 고객 맞춤 서비스를 위한 혼합 선호도 예측 시스템이라 할 수 있다.

그림 3은 선호도를 계산하는 알고리즘을 설명한 것이다.

```
HPP(x, i, k){
  V : 사용자의 이웃집단;
  N : 전체 사용자의 수;
  r_xy : 사용자 x와 y의 유사도;
  n_x : 사용자 x의 상품 평가 수 ;
  n_y : 사용자 y의 상품 평가 수;
  V = select_Neighbor(x, i, simC, ratingC);
  // x의 인접 이웃 선택
  avg_x = sum(x, rating)/n_x; // x의 평가평균
  avg_y = sum(y, rating)/n_y; // y의 평가평균
  avg_rating_i = sum(i, rating)/N;
  // 상품 i에 대한 평가평균
  for(l=1 ; l<=k ; l++){
    if(rating_y_i == 0)
      a = a + (avg_rating_i - avg_y) * r_xy;
    else
      a = a + (rating_y_i - avg_y) * r_xy;
      b = b + r_xy;
  }
  p_x_i = avg_x + w_xy + a / b;
  return p_x_i;
}
```

그림 3. 선호도 예측 계산 알고리즘

3.2.3 인접이웃 선정 방법 설계

인접이웃의 정보에 따라 추천 결과가 달라질 수 있기 때문에 추천시스템에서 인접이웃 선정은 매우 중요하다. 본 논문에서는 동일한 상품만을 평가한 이웃들을 인접이웃으로 선정하는 방법과 유사도가 높은 이웃들을 인접이웃으로 선정하는 두 가지 방법을 혼합하여 인접이웃을 선정하도록 하여 사용자의 선호도 예측값을 향상시키고자 하였다.

1단계: HPPS에서 선정한 유사도 임계치보다 크고 상품평가 횟수 임계치 범위 내에 포함되는 이웃들을 1차 인접이웃으로 선정한다.

2단계: 동일한 제품을 평가하고 상품평가 횟수 임계치 범위 내에 포함되는 이웃들을 2차 인접이웃으로 선정한다.

3단계: 두 개의 자료를 병합하여 최종 인접이웃을 선정하는데, 이때 중복되는 이웃들은 하나만 남기고 나머지는 삭제하여 최종 인접이웃을 선정한다.

그림 4는 인접이웃 선정 알고리즘을 설명한 것이다.

3.3 HPPS의 데이터베이스 설계

본 논문에서 제안하는 개인 맞춤 서비스를 위한 HPPS의 데이터베이스는 기본적인 사용자의 정보를 저장하는 UserInfo 테이블, 상품 정보를 저장하는 ItemInfo 테이블, 실험 데이터인 Training 테이블이 있다. 그리고 사용자의 정보를 분석하여 저장하는 UserProfile 테이블과 사용자간의 유사도 분석한 결과, 평가여부, 평가횟수, 동일 상품 평가 비율, 상품

```

Select_Neighbor(x, i, simC, ratingC){
    simC : 유사도 임계치;
    ratingC : 상품평가 횟수 임계치;
    r_xy : 사용자 x와 y의 유사도;
    N : 전체 사용자의 수;
    rating_i : 상품 i에 대한 평가;
    rating_cnt : 사용자의 상품평가 횟수;
    for(l=0 ; l<=N ; l++){
        if(r_xy >=simC &&
           rating_cnt >= ratingC)
            select_neb_1(l); // 1차 인접이웃
        if(rating_i != 0 &&
           rating_cnt >= ratingC)
            select_neb_2(l); // 2차 인접이웃
    }
    nn = merge_del(nn_1, nn_2); //병합하고
    // 중복 데이터 삭제하여 최종 인접이웃선정
    return nn;
}

```

그림 4. 인접 이웃 선정 알고리즘

선호 비율을 저장하는 Neighbor 테이블, 추천목록을 저장하는 Recommendation 테이블로 구성된다.

그림 5는 HPPS의 데이터베이스의 테이블간의 관계를 설명한 그림이다.

3.3.1 사용자 프로파일 테이블 설계

Training 테이블에 저장된 사용자의 상품 평가 데이터와 사용자의 기본 정보가 저장된 UserInfo 테이블, 상품 정보가 저장된 ItemInfo 테이블을 조인(join)하여 사용자가 평가한 상품들, 선호하는 성향 등을 분석해 UserProfile 테이블에 저장한다. 사용자

그림 5. HPPS의 데이터베이스

프로파일에는 사용자의 평가평균값(avg_rating), 사용자의 평가횟수(rating_cnt), 선호 상품 순위를 5순위까지 평가하여 저장한다. 이때 저장된 사용자의 평가평균값(avg_rating)은 본 논문에서 제안한 선호도 예측식 (5)에 사용되고, 평가 횟수(rating_cnt)는 인접이웃을 선정할 때 사용한다. 표 2는 UserProfile 테이블의 형식을 설명한 것이다.

표 2. UserProfile 테이블 형식

필드명	유형	길이	설 명
userID	VARCHAR	20	사용자ID, 기본키
avg_rating	FLOAT		사용자 평가평균값
rating_cnt	INT		평가횟수
p_item1	VARCHAR	30	선호상품 1순위
p_item2	VARCHAR	30	선호상품 2순위
p_item3	VARCHAR	30	선호상품 3순위
p_item4	VARCHAR	30	선호상품 4순위
p_item5	VARCHAR	30	선호상품 5순위

3.3.2 상품별 평가평균값 테이블 형식

Training 테이블에는 사용자가 각 상품에 대해 평가한 평가값이 저장되어 있다. 이 Training 테이블의 각 상품에 대한 평가평균값을 구해 ItemAvg 테이블에 저장한다. 각 상품에 대한 평가평균값(avg_rating_i)는 본 논문에서 제안한 선호도 예측식 (5)에서 상품에 대한 사용자의 평가값이 없을 경우에 사용자의 상품 평가값 대신에 사용한다. ItemAvg 테이블의 형식은 표 3과 같다.

표 3. ItemAvg 테이블 형식

필드이름	유형	길이	설 명
itemID	VARCHAR	20	상품ID, 기본키
avg_rating_i	FLOAT		상품 ID에 대한 평가평균값

3.3.3 인접이웃 테이블 설계

HPPS는 UserProfile 테이블, ItemInfo 테이블, Training 테이블을 참조하여 사용자 X와 이웃 Y의 유사도, 평가여부, 평가횟수에 대한 조건 값이 맞는 사용자들을 HPPS의 인접 이웃을 선정해 Neighbor 테이블에 저장한다. 이때, 선정된 인접이웃과 사용자

의 동일한 평가값 비율(s_rates), 상품 선호 비율(p_rates)을 측정하여 본 논문에서 제안한 선호도 예측식 (5)의 가중치로 이용하여 좀 더 정확한 고객의 선호도를 예측하고자 한다.

표 4는 Neighbor 테이블의 형식을 설명한 것이다.

표 4. Neighbor 테이블 형식

필드명	유형	길이	설 명
userX	VARCHAR	20	사용자 ID
userY	VARCHAR	20	이웃 ID
itemID	VARCHAR	30	상품 ID
sim	FLOAT		userX와 userY의 유사도
rating	INT		사용자 Y의 실제 평가값
ranking	INT		유사도 순위
avg_rating	FLOAT		사용자의 평균 평가값
rating_cnt	INT		평가횟수
s_rate	FLOAT		동일 상품 평가율
p_rates	FLOAT		상품 선호율

3.3.4 추천 테이블 설계

본 논문에서 제안한 HPPS는 UserProfile 테이블과 Neighbor 테이블을 참조하여 선호도 예측식 (5)을 이용해 각 상품에 대한 고객의 선호도를 예측하고, 그 결과를 MAE 평가를 통해 선호도의 정확도를 계산한다. 고객의 선호도의 정확도가 높은 상품들을 Recommendation 테이블에 저장한 후, 고객에게 추천 리스트를 제공한다. 표 5는 Recommendation 테이블의 형식을 설명한 것이다.

표 5. Recommendation 테이블 형식

필드명	유형	길이	설 명
userID	VARCHAR	20	사용자 ID, 기본키
r_item1	VARCHAR	30	추천 상품 1순위
r_item2	VARCHAR	30	추천 상품 2순위
r_item3	VARCHAR	30	추천 상품 3순위
r_item4	VARCHAR	30	추천 상품 4순위
r_item5	VARCHAR	30	추천 상품 5순위

4. 실험결과 및 분석

본 논문에서 제안하는 개인 맞춤 서비스를 위한 HPPS 설계의 실험은 GroupLens 연구 프로젝트로 수집된 MovieLens 데이터 중에서 100K 데이터를 이용하였다. 100K 데이터는 943명의 인원이 1682편의 영화에 대한 평가 자료 100,000건으로 구성되며, 각 사용자는 적어도 20개의 영화에 대해 평가하였으며, 이 데이터는 80%의 훈련 데이터와 20%의 실험 데이터로 구성된다[9,10].

실험환경은 운영체제 Windows XP, Intel Core Duo CPU 2.20 GHz, RAM 2.0GB의 노트북으로 하였으며, 데이터베이스 MySQL를 이용하였으며, 알고리즘 구현은 웹 프로그래밍 언어인 PHP를 이용하였다. 그리고 본 논문에서 제안하는 HPPS의 정확도(precision)를 측정하여 기존의 협업적 필터링 추천 시스템과 비교하여 본 논문에서 제안하는 HPPS를 평가하였다.

4.1 인접 이웃(Nearest Neighbor) 선정

MovieLens의 훈련 데이터를 이용해 사용자 1과 상품(영화)번호 15를 기준으로 다른 사용자들과의 유사도를 계산한 후 그 유사도를 기준으로 20순위까지 출력한 결과가 그림 6이다. 그림 6의 Rating은 상

품 15에 대한 이웃들의 평가를 의미하고, Rating_cnt는 사용자 1과 이웃이 동시에 평가한 상품의 수를 의미한다.

그림 6을 살펴보면 유사성은 매우 높지만 사용자 1과 동일한 제품을 평가하는 이웃은 2명뿐이고, 사용자 9와 사용자 1의 유사도는 1로 나왔지만 실질적으로 사용자 9와 사용자 1은 단지 2개의 동일한 상품에 대해서만 평가하였다. 나머지 이웃들도 마찬가지로 사용자 1과 동일한 상품에 대해 평가가 적다. 즉, 유사성만으로 인접 이웃을 선정할 경우 이런 경우를 배제할 수 없으므로 사용자 1의 선호도를 정확한 예측하기가 어렵다.

만약, 사용자 1과 동일한 상품에 평가하는 이웃들만 선정한다면 그림 7과 같다. 그림 7을 살펴보면, 유사도가 0.8 이상인 사용자는 6명이고, 유사도 순위 20위 안에 드는 사용자는 2명이다. 그리고 이 2명의 유사도는 1로 유사도 순위 기준으로 이웃을 선정할 때 중복되는 이웃들이다. 그림 7에서 알 수 있듯이, 사용자 1과 동일한 상품에 평가한 이웃들은 유사도가 낮은 이웃들이 선정되므로 이 이웃들로 사용자 1의 선호도를 예측한다면 정확한 결과를 얻기가 어렵다.

이웃을 선정하는 실험에는 MovieLens의 훈련 데이터에서 이웃들은 전체 943명이었는데, 본 논문에서

rank	User Y	Similarity	Rating	rating_cnt
1	9	1	0	2
2	29	1	0	3
3	46	1	0	3
4	61	1	0	2
5	120	1	4	5
6	123	1	0	4
7	149	1	0	2
8	229	1	0	2
9	238	1	0	2
10	240	1	0	2
11	289	1	3	3
12	364	1	0	2
13	408	1	0	2
14	485	1	0	2
15	510	1	0	2
16	616	1	0	2
17	656	1	0	3
18	724	1	0	4
19	732	1	0	2
20	740	1	0	2

그림 6. 유사도 기준 이웃 선정 결과
(사용자 X=1, 상품 ID=15인 경우)

rank	User Y	Similarity	Rating	rating_cnt
5	120	1	4	5
12	289	1	3	3
29	227	0.95	4	7
32	718	0.91	5	4
42	165	0.82	5	8
43	754	0.82	5	4
50	21	0.77	4	12
53	274	0.76	5	7
63	941	0.69	4	7
66	817	0.68	3	7
67	413	0.67	4	11
71	340	0.66	5	9
79	540	0.62	3	19
80	623	0.62	4	13
81	53	0.61	5	6
90	65	0.59	5	18
101	490	0.57	1	16
109	769	0.55	3	6
110	892	0.55	4	49
111	938	0.55	2	18

그림 7. 동일한 상품을 평가한 이웃 선정 결과

서는 이 이웃들 중에 사용자 1과 유사도가 0 이하인 사용자를 제외시킨 552명의 데이터를 이용하였으며, 552명의 이웃들의 평가횟수에 따른 유사도 평균은 표 6과 같다. 이 표에서는 구간 [1, 10] 사이의 유사도 평균이 가장 높은 것을 알 수 있지만, 이것은 그림 6에서 설명했듯이 유사도가 1인 이웃이 26명으로 유사도 평균값이 커진 것이기 때문에 본 논문에서는 [1, 10] 구간에서 5미만은 제외하고, 나머지 구간 중에서 가장 높은 이웃수와 유사도 평균값을 갖는 구간 [11, 30]내에 포함되는 이웃들을 인접 이웃을 선정하였다.

본 논문에서는 유사도, 동일 제품 평가 여부, 상품 평가 횟수를 모두 반영하여 HPPS의 유사도 임계치 0.80, 상품평가 횟수 5-30 사이를 만족하는 1차 인접이웃을 선정하고, 동일 제품 평가하고, 상품평가 횟수가 5-30 사이를 만족하는 2차 인접 이웃을 선정한다. 그리고 난후, 1차와 2차 인접이웃을 병합한 후, 중복된 이웃들을 삭제하여 최종 인접이웃을 선정하여 사용자의 선호도 예측값을 향상시키고자 하였다.

표 6. 평가 횟수에 따른 유사도 분석

구간	이웃 수(도수)	유사도 평균
1-10	181	0.55
11-20	124	0.37
21-30	83	0.32
31-40	68	0.29
41-50	49	0.30
51-60	21	0.31
61-70	18	0.34
71 이상	8	0.29
합계	552	

4.2 인접이웃 크기 선정

본 논문에서는 유사도, 동일한 상품 평가 여부, 상품 평가 횟수를 모두 반영하여 사용자 1에 대한 인접이웃을 선정하고, 인접이웃 그룹의 크기를 5, 10, 15, 20, 30으로 설정해 이웃 크기별 MAE 평가를 하였다. MAE는 기존의 NBCFA의 선호도 예측식과 본 논문에서 제안하는 HPPS의 선호도 예측식을 이용하여 실험하였다.

NBCFA의 경우 인접이웃 크기를 5에서 10으로 증가시키면 선호도 예측값의 정확도가 72.14% 증가하

고, 본 논문에서 제안하는 HPPS의 선호도 예측값의 정확도는 57.14% 증가하지만, 인접 이웃 크기를 15로 증가시키면 모두 선호도 예측값의 정확도가 감소한다.

표 7과 그림 8을 살펴보면, 인접 이웃 크기는 10이 적합하다고 할 수 있으며, NBCFA 보다 본 논문에서 제안하는 HPPS의 선호도 예측값이 더 정확한 것을 알 수 있다.

나머지 실험은 본 논문에서 제안한 HPPS의 인접 이웃 크기로는 10개로 선정하여 인접 이웃 그룹을 생성해 사용자 1의 선호도를 측정하는데 사용하였으며, 사용자의 선호도 예측식은 식(5)을 사용하였다.

표 7. 이웃 크기에 대한 MAE 평가

이웃 크기(K) 선호도 예측방법	5	10	15	20	30
NBCFA	2.01	1.45 (72.14%)	1.57	1.3	1.33
HPPS	0.07	0.04 (57.14%)	0.15	0.1	0.16

그림 8. 인접이웃 크기에 따른 MAE 평가

4.3 인접 이웃 선정 방법에 따른 정확도 분석

본 논문에서 제안하는 HPPS의 인접 이웃 선정 방법은 유사도, 동일 제품 평가 여부, 평가 참여 횟수를 적용하여 인접이웃을 선정하는 혼합방식이지만, 인접 이웃 선정 방법의 정확도 분석 실험을 위해 인접 이웃 선정 방식을 아래와 같이 3가지로 구분하였다. 그리고 본 논문에 제안한 HPPS의 선호도 예측식 (5)와 기존의 NBCFA의 선호도 예측식 (1)로 선호도 예측값을 계산하였으며, 선호도 예측값의 정확도는

MAE 방법으로 분석하였다.

- 유사도 기준(Similarity) : 유사도가 0.80 이상인 이웃들만 인접이웃으로 선정
- 동일 상품 평가 여부(Rating) : 사용자 1과 동일한 상품에 대해 평가한 이웃들만을 인접이웃으로 선정
- 혼합 방식(Combination) : 유사도, 동일 제품 평가 여부, 평가 참여 횟수를 적용하여 인접이웃 선정

이러한 인접 이웃 선정 방법에 따라 인접 이웃을 선정하고 HPPS의 선호도 예측값과 NBCFA의 선호도 예측값의 정확도를 분석한 결과는 표 8과 그림 9와 같다.

그림 9에서 알 수 있듯이 인접 이웃 크기가 10일 때, 본 논문에서 제안한 HPPS의 선호도 예측식 (5)을 이용한 사용자의 선호도 예측값이 가장 정확한 것으로 나타난다. 즉, 사용자의 특성을 분석한 HPPS의 가중치를 이용한 선호도 예측값의 정확도를 측정했을 때, Similarity 선정 기법 보다 Rating 선정기법

표 8. 이웃 선정 방법에 따른 MAE 평가(이웃 크기 K=10일 경우)

선택도 이웃선정방법	선택도 예측방법	NBCFA	HPPS	평가 (NBCFA : HPPS)
유사도 기준 (Similarity)		1.68	0.16	90.48%
동일상품평가여부 (Rating)		1.38	0.1 (37.50%)	92.75%
혼합방식 (Combination)		1.45	0.04 (75%, 60%)	97.24%

그림 9. 선호도 예측식의 정확도 분석

이 37.50% 증가하였고, Similarity 선정기법과 Rating 선정기법보다 Combination 선정 기법이 각각 75%, 60% 증가하였다.

그리고 이웃 선정 방법에 따라 기존의 NBCFA와 본 논문에서 제안한 HPPS의 고객 선호도 예측값에 대한 MAE 평가 결과가 90.48%, 92.75%, 97.24%가 각각 향상된 것을 알 수 있다. 따라서 Combination 기법을 사용한 HPPS는 고객 선호도 예측값의 정확도를 향상시켰다고 할 수 있다.

5. 결 론

본 논문에서 제안한 HPPS에서는 사용자의 선호도 예측에 대한 정확도를 향상시키기 위해 다음과 같이 제안하였다.

첫째, 사용자의 선호도를 측정하기 위해 사용하는 선호도 예측식에서 이웃의 상품 평가가 없을 경우 상품에 대한 평가 평균값을 이용하도록 설계하였다.

둘째, 사용자의 선호도를 측정하기 위해 사용하는 선호도 예측식에 사용자의 특성을 분석한 가중치를 접목하여 사용자 선호도 예측의 정확도를 향상시켰다.

셋째, 인접 이웃을 설정할 때, 유사도, 상품 평가 여부, 평가 횟수를 반영하여 선정하도록 하였다.

그 결과, 인접이웃 크기를 10으로 하였을 경우에 HPPS의 정확도는 기존의 NBCFA에 비해 정확도가 97.24% 향상되었으며, HPPS는 인접이웃 선정방법에 따라 정확도가 75% 향상되었다.

본 논문에서 제안하는 HPPS는 제한적인 개념의 정보를 제공하는 시스템이 아니라, 사용자 특성 정보를 입력받아 선호도를 계산하여 제공하므로 제한적인 개념의 정보를 제공하는 단점을 해결하였으며, 사용자의 특성 정보를 프로파일링 하여 좀 더 정확한 정보를 추천하기 위한 데이터로 활용하여 사용자의 특성에 따른 정보를 고려하지 않고 서비스를 제공하는 어려움을 극복하였다고 볼 수 있다.

참 고 문 헌

- [1] 김경록, 변재희, 문남미, “장르유사도와 선호 장르를 이용한 협업 필터링 설계,” 한국컴퓨터정보학회논문지, 제16권, 제4호, pp. 161-170, 2011.

- [2] 권형준, 정동근, 홍광석, “사용자의 재생 시간을 이용한 멀티미디어 추천 시스템,” 한국인터넷정보학회논문지, 제10권, 제1호, pp. 111-121, 2009
- [3] 박정우, 이상훈, “사용자 프로파일을 이용한 개인화된 토픽맵 랭킹 알고리즘,” 정보과학회논문지 소프트웨어 및 응용, 제35권 제8호, pp. 522-528, 2008
- [4] 이미희, 우용태, “협업적 필터링과 SOM 신경망을 결합한 추천시스템 모듈,” 멀티미디어학회논문지, 제11권 제9호, pp. 1213-1226, 2008.
- [5] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry, “Using collaborative Filtering to Weave an Information Tapestry,” *Communications of the ACM*, Vol.35, Issue 12, pp. 61-70, 1992
- [6] Paul Resnick, Neophytos Iacovou, Mitest Suchak, Peter Bergstorm, and John Riedl, “GroupLens: An Open Architecture for Collaborative Filtering of Netnews,” In *Processing of ACM 1994 Conference on Computer Supported Cooperative Work*, pp. 175-186, 1994.
- [7] John S. Breese, David Heckerman, and Carl Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering,” In *Proceeding of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, Madison, Wisconsin, pp. 43-52, 1998.

- [8] Upendra Shardanand and Pattie Maes, “Social Information Filtering,” In *Processing of ACM CHI'95 Conference on Human Factors in Computing Systems*, pp. 210-217, 1995.
- [9] MovieLens Dataset, http://www.grouplens.org/system/files/README_10M100K.html
- [10] 미네소타대학 Project Website, <http://movielens.umn.edu>

정 은 희

1991년 2월 강릉대학교 통계학과
이학사
1998년 2월 관동대학교 전자계산
공학과 공학석사
2003년 2월 관동대학교 전자계산
공학과 공학박사

2003년 9월~현재 강원대학교 지역경제학과 부교수
관심분야: 네트워크 보안, 웹 프로그래밍, 전자상거래

이 병 관

1979년 2월 부산대학교 기계설계
학과 공학사
1986년 2월 중앙대학교 전자계산
공학과 공학석사
1990년 2월 중앙대학교 전자계산
학과 공학박사

1988년 3월~현재 관동대학교 컴퓨터학과 교수
관심분야: 네트워크 보안, 컴퓨터 네트워크, 전자상거래