논문 2005-42CI-6-8

개인화 추천 시스템의 예측 정확도 향상을 위한 사용자 유사도 가중치에 대한 비교 평가

(Comparative Evaluation of User Similarity Weight for Improving Prediction Accuracy in Personalized Recommender System)

정 경 용*, 이 정 현**

(Kyung-Yong Jung and Jung-Hyun Lee)

요 약

전자상거래에서 최근 대부분의 개인화된 추천 시스템들은 협력적 필터링 기술을 적용하고 있다. 이 방법은 사용자의 성향에 맞는 아이템을 예측하고 추천하기 위하여 비슷한 선호도를 가지는 사용자들간의 유사도 가중치를 계산한다. 이때 일반적으로 피어슨 상관계수를 많이 사용한다. 그러나 이 방법은 두 사용자가 공통으로 선호도를 평가한 아이템들이 있을 때만 상관관계를 계산할 수 있으므로 예측의 정확도는 떨어진다. 사용자 유사도 가중치는 사용자의 성향에 맞는 아이템을 예측하는 경우 뿐만 아니라 개인화된 추천 시스템의 성능에 영향을 미칠 수 있다. 본 논문에서는 정보검색 분야의 벡터 유사도, 엔트로피, 역사용자 빈도, 기본 선호도 평가를 적용하여 유사도 가중치 공식에 대해서 살펴보고, 추천 시스템의 예측 정확도 향상에 대해서도 실험을 통해 확인해 보았다. 실험 결과는 엔트로피를 이용한 유사도 가중치에 기본 선호도 평가를 결합하는 방법이 가장 성능이 우수함을 알 수 있다.

Abstract

In Electronic Commerce, the latest most of the personalized recommender systems have applied to the collaborative filtering technique. This method calculates the weight of similarity among users who have a similar preference degree in order to predict and recommend the item which hits to propensity of users. In this case, we commonly use Pearson Correlation Coefficient. However, this method is feasible to calculate a correlation if only there are the items that two users evaluated a preference degree in common. Accordingly, the accuracy of prediction falls. The weight of similarity can affect not only the case which predicts the item which hits to propensity of users, but also the performance of the personalized recommender system. In this study, we verify the improvement of the prediction accuracy through an experiment after observing the rule of the weight of similarity applying Vector similarity, Entropy, Inverse user frequency, and Default voting of Information Retrieval field. The result shows that the method combining the weight of similarity using the Entropy with Default voting got the most efficient performance.

Keywords: 협력적 필터링(Collaborative Filtering), 정보검색(Information Retrieval), 데이터마이닝(Data Mining)

I. 서 론

최근에는 사용자 개개인의 취향이나 특성에 맞는 정보를 자동으로 찾거나 추천해 주는 개인화 기술이 크게 요구되고 있다[1,2]. 개인화 추천 시스템은 자동화된 정보 필터링 기술을 적용하여 사용자의 취향에 맞는 상품을

* 정회원, 한세대학교 IT학부
(Division of Information Technology, Hansei Univ.)
** 평생회원, 인하대학교 컴퓨터공학부
(School of Computer Sci. & Eng., Inha Univ.)
※ 이 논문은 2005학년도 한세대학교 연구비 지원에 의하여 연구되었습니다.
접수일자: 2005년9월15일, 수정완료일: 2005년11월1일

추천해 주는 시스템이다. 여기서 가장 중요한 것은 사용자의 선호도를 정확하게 분석하고 정제하여 정확한 예측력으로 고객이 원하는 가장 적절한 상품을 추천해 줄 수 있는 능력이다. 이를 위해 데이터 마이닝, 패턴 인식 기술, 정보 필터링 기술 등의 다양한 기법들이 적용될 수 있으나 대부분의 추천 시스템들은 정보 필터링 기술을 적용한다. 정보 필터링 기술에서 가장 대표적인 것은 협력적 필터링 기술이다[3]. 협력적 필터링 기술은 추천 시스템에서 가장 많이 사용되는 방법으로 Amazon*, Moviecritic**, Jester***, Firefly**** 등의 유명한 전자 상거래 사이트에서 성공적인 결과를 보여주고 있다.

Tapestry[1]는 협력적 필터링 기술을 가장 먼저 적용한 문서 필터링 시스템으로 워크 그룹과 같은 공동체 구성원들의 의견에 기반하여 추천을 해주므로 개인화된 추천 서비스는 제공해 주지 못한다. 이런 Tapestry의 문제점을 보안하면서 성능을 인정 받은 GroupLens은 유즈넷 뉴스와 영화에 대한 추천을 수행한다[4,5]. 그리고 협력적 필터링 개인화 기술을 응용한 패션 디자인 추천 에이전트 시스템(FDRAS)[6,7,8]이 있다. 이 방법은 사용자가 좋아할만한 아이템을 예측하기 위해 비슷한 선호도를 보이는 다른 사용자들의 아이템에 대한 평가 데이터를 근거하여 추천하는 방법이다. 그러므로 높은 예측력과 추천 능력을 가지고 있는 장점이 있다. 그러나 협력적 필터링 기술에서 지적되는 문제점은 사용자간의 유사도 가중치를 계산하기 위해 사용하는 피어슨 상관계수 기반의 예측 기법에서부터 야기된다[1,9,10,11].

첫째, 사용자간의 상관관계는 선호도를 평가한 아이템에 대해서만 계산된다. 만약 아이템이 많으면 일반적으로 같은 아이템에 대하여 사용자가 선호도를 평가할 확률은 적어지게 된다. 둘째, 비록 사용자가 선호도에 따른 상관관계가 높지 않더라도 다른 사용자의 선호도 예측에 좋은 자료가 될 수 있다. 그러나 상관관계가 높지 않다는 이유로 이 정보는 활용되지 못한다. 마지막으로 두 사용자 사이에서만 상관관계를 계산할 수 있으므로 예측의 정확도는 낮아진다.

본 논문에서는 위에서 언급한 기존의 협력적 필터링 기술의 문제점을 보안하기 위해 정보검색 분야를 적용하여 유사도 가중치를 계산하는 방법을 제안하고 그 성능을 기존의 협력적 필터링 기술에서 가장 많이 쓰이는 피어슨 상관계수와 비교 평가하였다.

* Amazon (http://www.amazon.com/)
** Moviecritic (http://www.moviecritic.com)
*** Jester (http://shadow.ieor.berkeley.edu/humor/)
**** Firefly (http://www.firefly.com/)

Ⅱ. 기존의 협력적 필터링 기술

협력적 필터링 기술은 선호도에 대한 데이터를 기반으로 사용자의 관심을 갖게 할 아이템을 추천해주는 기법이다[12]. 협력적 필터링 기술에서 우선적으로 필요한 것은 특정 사용자와 유사한 선호도를 가지는 이웃을 찾아 내는 것이다. 유사도 가중치에 관계없이 가중치가 구해진 모든 이웃들을 사용해서 선호도를 예측할 수 있지만 이는 성능이나 정확도면에서 그리 좋은 방법은 아니다. 반면 유사도가 높은 이웃들만을 예측해서 고려할 경우 다른 사용자들과 유사도가 높지 않은 사용자의 아이템에 대해서 예측할 수 없는 경우가 발생한다. 그러므로 시스템이 예측할 수 있는 적절한 이웃의 수를 결정하는 것이 무엇보다도 중요하다. 예측에 사용될 이웃의 수를 결정하기 위해서 임계값과 가장 좋은 이웃을 선정해야 한다[13].

임계값은 사용자간의 유사도 가중치가 어느 정도의 값 이상인 이웃들만 사용해서 예측하도록 제안하는 방법이고, 가장 좋은 이웃은 특정 사용자와 유사한 n명의 이웃을 사용하여 예측할 수 있게 제안하는 방법이다. 위의 두 방법을 조합하여 유사도 가중치가 어느 정도 값 이상인 이웃들 중 n명을 사용할 수 있도록 하는 것도 좋은 방법이다. 여기서 사용자간의 유사도 가중치를

[단계 1] 사용자 유사도 가중치
새로운 사용자와 이웃들과의 유사도 가중치를 정의하고 계산
[단계 2] 이웃선택
새로운 사용자의 특정 아이템에 대한 선호도를 예측하기 위해서 유사도가 높은 이웃을 어떤 기준으로 몇 명 선택할지를 결정
[단계 3] 예측
유사한 선호도를 가지는 이웃들의 아이템에 대한 선호도를 기반으로 새로운 사용자의 선호도가 평가되지 않는 아이템의 선호도를 예측

그림 1. 협력적 필터링 기술의 3단계
Fig. 1. 3 Steps for Collaborative Filtering Technique.

계산하기 위해 사용되는 대표적인 유사도 기준 값으로 피어슨 상관계수가 있다. 기본적으로 협력적 필터링 기술은 3단계를 통하여 구현되며 각 단계가 진행된 후 새로운 아이템에 대한 선호도를 발견할 수 있다. 협력적 필터링 기술의 각 단계는 그림 1과 같다[12].

단계 1에서 피어슨 상관계수를 사용하여 사용자 a와 사용자 i의 유사도 가중치는 식 (1)과 같이 정의된다.

$$w(a,i)=\frac{\sum_{j=1}^{m}(v_{a,j}-\overline{v_a})\times(v_{i,j}-\overline{v_i})}{\sqrt{\sum_{j=1}^{m}(v_{a,j}-\overline{v_a})^2\times\sum_{j=1}^{m}(v_{i,j}-\overline{v_i})^2}} \quad (1)$$

$v_{a,j}$는 새로운 사용자 a가 아이템 j에 대해서 평가한 선호도이고, $v_{i,j}$는 사용자 i가 아이템 j에 대해서 평가한 선호도이다. $\overline{v_a}$는 새로운 사용자 a가 선호도를 평가한 아이템들에 대한 선호도 평균값이다. j는 새로운 사용자 a와 사용자 i가 공통으로 선호도를 평가한 아이템들이고, m은 아이템의 총 개수이다.

단계 3에서 선호도를 예측하기 위해서 가능한 모든 이웃들의 아이템에 대한 선호도를 사용하는 것이 기본적인 방법이지만 이웃들의 유사도를 가중치로 보고 선호도를 유사도 기준의 가중 평균값으로 계산하는 방법이 많이 사용된다. 따라서 모든 사용자들의 선호도는 근사적으로 같은 분포를 가진다고 가정하게 된다.

1. 기존의 개인화 추천 시스템에서의 예측 방법

특정 사용자와 가까운 이웃들이 선택이 되면, 예측을 위해서 이웃들의 아이템에 대한 선호도를 같은 분포의 척도로 변환하여 조합한다. 가능한 모든 이웃들의 아이템에 대한 선호도를 사용하는 것이 기본적인 방법이지만, 이웃들의 유사도를 가중치로 보고 선호도를 유사도 기준의 가중 평균값으로 계산하는 방법이 사용된다.

GroupLens는 이웃들의 선호도와 이웃들의 선호도 평균과의 거리를 가중 평균함으로써, 특정 사용자의 아이템에 대한 선호도를 예측한다. 이러한 접근 방법을 평균 편차라고 정의한다[4,5].

GroupLens의 확장된 접근 방식은 사용자의 선호도가 동일한 분산이 아니라는 가정하에 사용자의 선호도의 분산을 사용자별로 다르게 적용하여 예측하는 것인 데, 이러한 접근 방법을 z값 가중치 평균이라고 정의한다[14,15].

가. 평균 편차(Deviation from mean)

평균 편차는 특정 아이템에 대한 이웃들의 선호도와 각 이웃들의 선호도 평균과의 거리를 이웃들과의 유사도로 가중 평균하고, 이를 이용하여 특정 사용자의 아이템에 대한 선호도를 예측하는 것이다. 이 경우, 모든 사용자의 선호도는 근사적으로 같은 분포, 동일한 분산을 가진다고 가정한다. 이를 수식으로 표현하면 식 (2)와 같이 정의할 수 있다.

$$P_{a,k}=\overline{v_a}+\frac{\sum_{i=1}^{n}w(a,i)\times(v_{i,k}-\overline{v_i})}{\sum_{i=1}^{n}w(a,i)} \quad (2)$$

$P_{a,k}$는 사용자 a의 아이템 k에 대해서 예측한 선호도이고, $\overline{v_a}$는 사용자 a의 평균 선호도이다. n은 단계 2에서 결정된 이웃들 안에서 사용자 수이다. w(a,i)는 사용자 a와 사용자 i의 유사도 가중치이다.

나. z값 가중치 평균(Weighted average of z-score)

표준 편차 방법은 모든 사용자의 선호도가 근사적으로 같은 분포를 가지며, 동일한 분산을 가진다는 가정을 하지만 현실적으로 각 사용자 별로 선호도가 동일한 분산을 가질 것이라고 확신할 수 없다. 따라서 선호도가 사용자 별로 다른 분산을 가질 것으로 간주하고 선호도의 분산을 가중치로 주어 사용자간의 유사도 가중치 외에 더 상세한 정보를 고려하는데 이러한 방법을 z값 가중치 평균이라고 부른다. 수식으로 표현하면 식 (3)과 같이 정의한다.

$$P_{a,k}=\overline{v_a}+\sigma_a\frac{\sum_{i=1}^{n}w(a,i)\times(\frac{v_{i,k}-\overline{v_i}}{\sigma_a})}{\sum_{i=1}^{n}w(a,i)} \quad (3)$$

$P_{a,k}$는 사용자 a의 아이템 k에 대해서 예측한 선호도이고, $\overline{v_a}$는 사용자 a의 평균 선호도이다. n은 단계 2에서 결정된 이웃들 안에서 사용자의 수이다. w(a,i)는 협력적 필터링의 단계에서 사용자 유사도 가중치를 계산하게 된다. σ_a는 사용자 a의 선호도의 표준 편차이고, σ_i는 사용자 i의 선호도의 표준 편차이다. 표준화된 z 값으로 치환된 $(v_{i,k}-\overline{v_i})/\sigma_i$는 z 값의 가중 평균으로 예측하게 된다.

Ⅲ. 협력적 필터링에서 사용자 유사도 가중치

협력적 필터링 기술을 이용한 개인화 추천 시스템에서 사용자간의 유사도 가중치를 계산하기 위해서는 적절한 유사도 가중치를 부여할 필요가 있다. 기존에 사용된 유사도 기준값으로는 대표적으로 피어슨 상관계수가 사용된다. 본 논문에서는 개인화 추천 시스템의 예측 정확도 향상을 위한 사용자 유사도 가중치를 구하는 방법을 정보검색 분야에서의 벡터 유사도, 기본 선호도 평가, 엔트로피, 역 사용자 빈도수를 적용[16,17]하여 기존의 피어슨 상관계수와 실험을 통해 비교 평가하였다. 예측의 정확도 향상을 위한 사용자 유사도 가중치에 대한 비교, 평가는 다음과 같다.

<table>
<tr><th>협력적 필터링 기술</th><th>비교평가</th></tr>
<tr><td colspan="2">[단계1] 사용자 유사도 가중치</td></tr>
<tr><td>피어슨 상관계수</td><td>역 사용자 빈도수</td></tr>
<tr><td>벡터 유사도</td><td>엔트로피</td></tr>
<tr><td colspan="2">기본선호도 평가</td></tr>
<tr><td rowspan="4"></td><td>피어슨 상관계수 + 기본 선호도 평가</td></tr>
<tr><td>역 사용자 빈도수 + 기본 선호도 평가</td></tr>
<tr><td>벡터 유사도 + 기본 선호도 평가</td></tr>
<tr><td>엔트로피 + 기본 선호도 평가</td></tr>
<tr><td colspan="2">[단계2] 이웃선택</td></tr>
<tr><td colspan="2">임계값</td></tr>
<tr><td colspan="2">가장 좋은 이웃 선택(best-n-neighborhood)</td></tr>
<tr><td colspan="2">사용자의 대표속성을 이용한 이웃 선택
(Representative Attribute-Neighborhood)</td></tr>
<tr><td colspan="2">[단계3] 예측</td></tr>
<tr><td colspan="2">평균 편차</td></tr>
<tr><td colspan="2">z값 가중치 평균</td></tr>
</table>

1. 벡터 유사도를 이용한 유사도 가중치

정보검색 분야에서 문서 사이의 유사도를 측정하기 위해서 문서 안에 포함된 특정 단어의 발생 빈도 벡터를 사용하여 코사인 벡터 유사도를 구한다. 이러한 형태를 협력적 필터링에 적용하여 정보검색 분야의 문서를 인터넷 사용자로 보고 포함된 특정 단어의 발생 횟수를 선호도로 보아 두 사용자간의 유사도 가중치를 구할 수 있다. 벡터 유사도를 사용했을 경우 개념적으로 명시적 선호도 데이터를 문서에 포함된 단어 발생 빈도로 간주하므로 선호도는 항상 긍정적인 선호도만이 존재하는 것으로 보아야 하며 선호도가 표시되지 않은 아이템의 경우에는 가장 선호도가 낮은 아이템으로 보아야 한다. 이러한 개념 때문에 벡터 유사도는 명시적 선호도 데이터보다는 인터넷 사용자가 특정 아이템을 클릭하거나 본 횟수, 구매한 횟수를 이용하는 암묵적 선호도 데이터를 다룰 경우에 사용자간의 유사도 가중치로 사용되어야 할 것이다. 벡터 유사도를 사용했을 경우 사용자 a와 사용자 i의 유사도 가중치는 식 (4)와 같이 정의한다.

$$w(a,i)=\sum_{j}\frac{v_{a,i}}{M}\times\frac{v_{i,j}}{N} \tag{4}$$

$$M=\sqrt{\sum_{k\in I_a}(v_{a,k})^2},\quad N=\sqrt{\sum_{k\in I_i}(v_{i,k})^2}$$

I_a는 사용자 a가 선호도를 평가한 아이템들의 집합이고, M과 N은 많은 아이템의 선호도를 공통으로 평가한 사용자들이 다른 사용자들에 비해서 더 큰 벡터 유사도을 가지지 않게 하기 위해 정규화한다. 정규화를 위한 다른 방법으로는 선호도를 평가한 아이템의 개수를 사용하거나 선호도의 절대값을 합계한 값을 사용한다.

2. 기본 선호도 평가를 이용한 유사도 가중치

기본 선호도 평가는 피어슨 상관계수를 확장한 것으로, 기준이 되는 특정 사용자와 그 사용자와 유사도가 있는 사용자들이 공통으로 선호도를 평가한 아이템이 상대적으로 적을 경우에 사용한다. 피어슨 상관계수는 사용자 a와 사용자 i가 공통으로 선호도를 평가한 아이템들($I_a \cap I_i$)을 이용하는데 반해 선호도를 평가하지 않은 아이템에 대하여 기본 선호도 평가를 적용하여 사용자 a와 사용자 i중 한 사람이라도 선호도를 평가한 아이템들($I_a \cup I_i$)을 사용할 수 있다. 기본 선호도 평가는 더 나아가 사용자 중 어떤 사람도 선호도를 평가하지 않은 새로운 아이템에 대해서 기본값을 우선적으로 적용함으로써 개인화 추천이 가능하도록 할 수 있다. 대부분의 경우 기본 선호도 평가 값 d는 중립적이거나 다소간 비선호을 사용하는 경우가 많다.

피어슨 상관계수에 기본 선호도 평가를 적용하면 사용자 a와 사용자 i의 유사도 가중치는 식 (5)와 같다.

$$\rho = \frac{M(\sum_j v_{a,j}v_{i,j} + \alpha) - UV}{\sqrt{(M(\sum_j v_{a,j}^2 + \alpha) - U^2)(M(\sum_j v_{i,j}^2 + \alpha) - V^2)}}$$

$$\alpha = kd^2,\ M = n + k \tag{5}$$

$$U = \sum_j v_{a,j} + kd,\ V = \sum_j v_{i,j} + kd$$

j는 사용자 a와 사용자 i중 한 사람이라도 선호도를 평가한 아이템을 의미한다. n은 한 사람이라도 선호도를 평가한 아이템의 개수($n=|I_a \cup I_i|$)를 의미한다. k는 사용자 a와 사용자 i가 모두 선호도를 평가하지 않은 아이템의 개수이다.

기본 선호도 평가는 사용자 a와 사용자 i의 평균 선호도나 전체 사용자의 평균 선호도를 사용할 수 있다. 암묵적 데이터의 경우 웹 페이지의 방문 여부나 어떤 제품의 구매 여부등과 같이 방문이나 구매를 1로 볼 수 있는 경우에는 방문하지 않거나 구매하지 않는 경우 기본 선호도를 0으로 설정할 수 있다. 또한 웹 페이지의 방문 횟수나 제품군의 제품 구매 횟수 등에서도 마찬가지로 기본 선호도 평가는 방문 횟수나 아이템 구매 횟수를 0으로 설정할 수 있다. 기본 선호도 평가는 사용자에 의해서 선호도를 평가하지 않은 아이템에 대해서 보완점을 가질 수 있다. 이는 사용자 a와 사용자 i가 공통으로 평가한 아이템의 수가 일정한 기준을 넘는 경우에 사용하는 것이 좋다.

3. 역 사용자 빈도를 이용한 유사도 가중치

정보검색 분야에서 벡터 유사도를 이용하는데 있어 문서에 포함된 단어의 수는 단어 빈도의 역수에 의해 수정된다. 공통적으로 많이 발생하는 단어의 경우 문서를 분류해내는데 유용한 역할을 할 수 없다는 판단에 따라 공통적으로 많이 발생하는 단어에 대하여 가중치를 줄이는 것이다. 협력적 필터링 기술에 이러한 개념을 적용해서 일반적으로 사용자에 의해 많이 선호도가 평가되는 아이템은 적게 선호도가 평가되는 아이템에 비해서 사용자간의 차이를 보여주는데 덜 기여한다고 판단하여 가중치를 줄인다. 이를 위해 f_j를 정의하는데, f_j는 $\log(n/n_j)$로 정의된다. n_j는 아이템 j에 선호도를 평가한 사용자 수이며, n은 전체 사용자의 수를 나타낸다. 따라서 모든 사용자에 의해 선호도를 평가한 아이템 j는 $f_j = log1 = 0$이 되어 모든 사용자에 의해서 선호도가 평가된 아이템 j는 사용자의 유사도 가중치를 계산하는데 사용되지 않게 된다.

역 사용자 빈도 f_j를 피어슨 상관계수에 적용하여, 새로운 사용자 a와 사용자 i의 유사도 가중치는 식 (6)과 같다.

$$w(a,i) = \frac{\sum_j f_j(\sum_j v_{a,i}v_{i,j} - (\sum_j f_j v_{a,i})(\sum_j f_i v_{i,j}))}{\sqrt{M \times N}}$$

$$M = \sum_j f_i(\sum_j f_j v_{a,j}^2 - (\sum_j f_j v_{a,j})^2) \tag{6}$$

$$N = \sum_j f_i(\sum_j f_j v_{i,j}^2 - (\sum_j f_j v_{i,j})^2)$$

역 사용자 빈도 f_j를 벡터 유사도에 적용하여 구한 새로운 사용자 a와 사용자 i의 유사도 가중치는 식 (7)과 같다.

$$w(a,i) = \sum_j \frac{f_j v_{a,j}}{\sqrt{\sum_{k \in I_a}(f_k v_{a,k})^2}} \frac{f_j v_{i,j}}{\sqrt{\sum_{k \in I_a}(f_k v_{i,k})^2}} \tag{7}$$

4. 엔트로피를 이용한 유사도 가중치

사용자 유사도 가중치를 계산하기 위해서 아이템들이 갖는 평균 정보량인 엔트로피를 이용하여 확률 벡터 연산을 사용한다. 군집 내에 각 속성 가중치는 일반적으로 샤논의 정보 이론[18]을 사용한다.

가. 확률 벡터의 정보량

확률 벡터의 정보량은 샤논의 정보이론에 근거하여 계산하였다. 샤논은 불확실성의 크기를 엔트로피로 측정하였는데, 이것이 사용자가 갖는 평균 정보량이 된다. 식 (8)은 평균 정보량을 구하기 위한 식이다.

$$E(p) = -\sum_{i=1}^{n} p_i log_2 p_i \tag{8}$$

사용자 갖는 정보량 $\log_2 p_i$는 선택될 확률벡터 p_i에 의해 결정된다. 따라서 식 (8)에서 E(p)는 n개의 사용자가 갖는 평균 정보량이며, p_i는 i번째 아이템이 선택될 확률을 나타낸다. 엔트로피는 확률 벡터에 대하여 유일한 값을 가지며 보통 확률 분포에 대한 평균치와 유사하다. 일반적으로 확률 벡터에 대한 엔트로피는 유일한 숫자로 결정되며 어떤 의미에서 그 확률 벡터를 요약하는 대표값이라 볼 수 있다. 이러한 엔트로피의 특징으

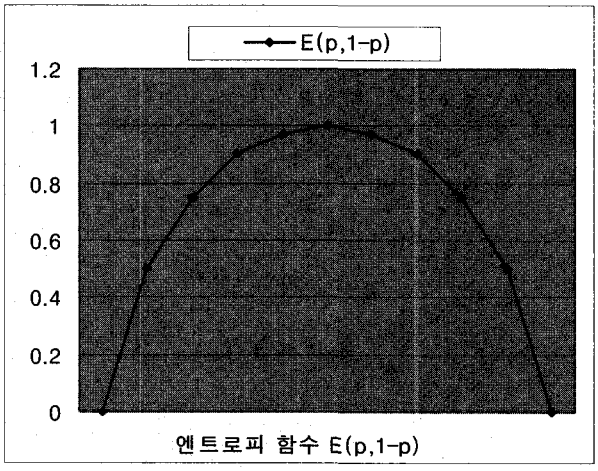

그림 2. E(p,1-p)의 엔트로피 함수
Fig. 2. Entropy Fuction of E(p,1-p).

로는 다음과 같다.

첫째는 확률 벡터에서 하나의 확률만이 극값을 갖고, 나머지 다른 확률들은 모두 0의 값을 가질 때, 엔트로피 E(p)는 최소값을 가지며 이때 얻어지는 엔트로피 정보량은 0이다.

둘째로 확률 벡터의 각 확률이 모두 일정한 값을 가질 때 엔트로피 E(p)의 최대값을 갖게 되며 그 값은 $H(1/q,1/q,\cdots,1/q) = \log_2 q$와 같다.

그림 2는 엔트로피 함수 E(p)=plogp+(1-p)log(1-p)를 설명하고 있다. p가 균일한 분포일 때, 즉 p=1/2일 때, E(p)는 최대값을 갖는다. 엔트로피 값이 높다는 것은 확률 벡터에 대해 확률 분포간의 변별력이 낮음을 의미하고 엔트로피 값이 작다는 것은 사용되는 확률 벡터간에 서로 변별력이 크다는 것을 의미한다[19].

나. 샤논의 정보 이론

샤논으로부터 시작된 정보이론에 기반하여 연관성 측정에 쓰이는 것으로 상호정보량과 상대 엔트로피가 있다. 상호정보량이란 두 독립 사건의 확률 변수 X와 Y사이의 의존관계를 정량적으로 나타낸 것이다. 상호정보량은 두 확률이 완전히 독립적일 경우 0이 되고 의존관계가 깊을수록 높은 값을 가진다. 상호정보량은 대칭성을 만족하는 함수이므로 $MI(x,y)=MI(y,x)$가 성립한다. 상대 엔트로피는 KL거리(Kullback-Leiber: D_{KL}), 교차 엔트로피라고도 불리며 두 확률 분포 p(x)와 q(x)사이의 평균적인 차이를 측정하는 것으로서 연관성 측정을 위해서 주로 사용된다. 상대 엔트로피 값은 항상 0보다 크거나 같으며 두 확률이 일치할 경우에만 0이 되고, 대칭성을 만족하지 않는다.

다. 확률 벡터 연산 과정

속성의 엔트로피을 이용한 유사도 가중치 방법은 속성의 빈도수를 기반으로 군집 내에 응집도가 높은 속성을 발견하여 가중치를 높게 부여하는 방법이다. 기본 구조는 정보검색의 가중치 기법인 TF·IDF의 형태로 지역적 빈도 정보(TF)와 엔트로피를 사용한 전역적 빈도 정보(IDF)와의 결합을 통해서 군집간에 공통으로 분포하는 속성 가중치를 낮추고, 특정 군집에만 포함된 속성의 가중치를 높이는 효과가 있다. 이러한 효과로 군집 특징 선택에 있어서 중요한 필터링 방식을 제공한다. 즉, 전체 군집에 균등하게 분포된 속성의 가중치를 낮게 책정함으로써 최종 군집 특징 선택에서 발견 가능성을 최소화한다.

엔트로피를 이용한 유사도 가중치를 계산하는 과정은 다음과 같다. 사용자가 선호도를 평가한 아이템들을 확률 벡터로 표현할 수 있다. 사용자 i의 확률 벡터 U_i는 식 (9)와 같이 정의한다.

$$U_i = (P_{U_{i,1}}, P_{U_{i,2}}, P_{U_{i,3}}, \ldots, P_{U_{i,m}}) \tag{9}$$

$P_{U_{i,m}}$는 사용자 i가 평가한 모든 선호도의 값에 대한 아이템 m의 선호도 비율이다. 평균 정보량을 계산하기 전에 우선, 아이템들의 유한 집합 $T=\{t_1, t_2, t_3, t_4, \cdots, tm\}$이 주어졌을 때, 확률 함수는 식 (10)과 같이 정의한다.

$$p_i = P(U_i|t) = p_{U_i,t} / \sum_{j=1}^{n} p_{p_{U_i,t}} \tag{10}$$

$$\text{for } i = 1,2,\ldots,n, p_i \geq 0, \sum_{i=1}^{n} p_i = 1$$

식 (10)에서 확률 벡터 p_i는 아이템별 확률 벡터이다. 여기서 n은 사용자의 수를 나타낸다. 이러한 경우 아이템 t에 대한 엔트로피는 식 (11)과 같이 정의한다.

$$E(t) = E(p_1^t, p_2^t, \ldots, p_n^t) = -\sum_{i=1}^{n} p_i^t \log_2 p_i^t \tag{11}$$

이와 같은 방법으로 측정된 E(t)값은 확률 벡터들과의 곱으로서 다음과 같은 식 (12)가 성립되고, 새로운 테이블을 생성할 수 있다.

$$H(a,t) = E(t) \times P_{a,t} \tag{12}$$

식 (12)의 H(a,t)는 사용자 a가 아이템 t에 대해 평가한 선호도 $P_{a,t}$와 아이템에 대한 엔트로피 E(t)를 곱한 값으로 계산한다. $P_{a,t}$는 사용자 a에 대한 아이템 t의 확률 벡터 값이며, E(t)는 아이템별 엔트로피 값이다.

모든 사용자의 선호도 값은 근사적으로 같은 분포, 동일한 분산을 가진다고 가정하지만 현실적으로 같은 분포, 동일한 분산을 가질 것이라는 확신할 수 없다. 따라서 이 사용자 선호도 값은 식 (10)을 이용하여 확률 벡터 P_t로 표현한다. 사용자 확률 벡터는 식 (11)을 이용하여 평균 정보량을 구하며 식 (12)에 의해 성립된 H(a,t)를 이용하여 기존의 테이블을 새로운 테이블로 변경할 수 있다. 따라서 사용자 유사도 가중치는 식 (13)과 같이 정의한다.

$$w(a,i)=\frac{\sum_j (H_{a,i}-\overline{H_a})(H_{i,j}-\overline{H_i})}{\sqrt{\sum_j (H_{a,i}-\overline{H_a})^2 \sum_j (H_{i,j}-\overline{H_i})^2}} \quad (13)$$

$H_{a,i}$는 사용자 a가 아이템 i에 대해서 평가한 선호도이고, $H_{i,t}$는 사용자 i가 아이템 t에 대해서 평가한 선호도이다. $\overline{H_i}$는 사용자 i가 선호도를 평가한 아이템들에 대한 선호도의 평균값이며, $\overline{H_a}$는 사용자 a가 선호도를 평가한 아이템들에 대한 선호도의 평균값이다. t는 사용자 a와 사용자 i가 공통으로 평가한 아이템들이다.

라. 엔트로피를 이용한 유사도 가중치 단계별 예제

본 논문에서 제안한 엔트로피를 이용한 유사도 가중치를 기본 선호도 평가에 결합하는 방법을 단계별 예제를 통해서 다음과 같이 설명한다.

데이터베이스의 평가 테이블과 사용자 테이블에서 샘플링하여 사용자 트랜젝션을 구성한다. 그 중에 하나의 사용자 트랜잭션을 가지고 이웃 선정 방식을 이용해서 10개의 이웃들로 나누었다. 10개의 이웃들 중 하나인 데이터(1개의 이웃)만을 추출하여 사용자들이 선호도를 평가하지 않은 아이템들에 대하여 식 (5)의 기본 선호도 평가를 이용한 유사도 가중치에 의해 값을 부여하고 식 (10)에 의해서 아이템에 대한 사용자 확률 벡터로 표현하여 계산한 결과는 표 1과 같다.

표 1. 기본 선호도 평가가 부여된 테이블
Table 1. Table applying Default Voting.

아이템	이웃선정				
	U_1	U_2	U_3	U_4	U_5
3	0.025	0.1	0.05	0.05	0.075
11	0.125	0.05	0.1	0.1	0.075
29	0.075	0.075	0.075	0.075	0.1
35	0.05	0.125	0.075	0.125	0.05
43	0.1	0.025	0.075	0.025	0.05

표 1의 기본 선호도 값이 부여된 테이블은 본 논문에서 제안한 엔트로피를 이용한 유사도 가중치를 설명하기 위해서 EachMovie 데이터[20]의 일부이다. 흑색으로 표시된 부분은 제공된 선호도에 기본 선호도 평가를 부여한 값이다.

식 (11)에 의해 사용자에 대한 아이템 별 엔트로피 값은 표 2와 같으며 구해진 엔트로피 값(E(pi))을 아이템별 사용자에게 할당함으로써 표 3과 같은 테이블을 얻어낼 수 있다.

표 3은 식 (12)를 이용하여 각각의 확률 벡터 값과 엔트로피를 곱한 결과값으로 갱신된 테이블이다.

{아이템-사용자} 행렬에서 아이템에 대해서 평가한 데이터가 존재한다. 그러나 모든 아이템에 대해서 평가를 하는 것은 불가능하다. 본 논문에서는 평가하지 않은 아이템에 기본 선호도 평가를 적용한다. 기본 선호도 평가는 사용자별 아이템에 대한 평균 값으로 정의한다. 이렇게 함으로써 {아이템-사용자} 행렬의 희박성 문제를 해결한다. 여기서 아이템별 확률 벡터를 기반으로 엔트로피를 계산할 수 있다. 계산된 엔트로피는 다

표 2. 아이템에 대한 엔트로피
Table 2. Entropy ablut the items.

아이템	3	11	29	35	43
E(p_i)	1.836	1.934	2	1.908	1.7

표 3. 확률 벡터와 엔트로피에 의한 테이블
Table 3. Table according to Probability Vector and Entropy.

아이템	이웃선정				
	U_1	U_2	U_3	U_4	U_5
3	0.046	0.184	0.092	0.092	0.138
11	0.242	0.097	0.193	0.193	0.145
29	0.15	0.15	0.15	0.15	0.2
35	0.095	0.239	0.143	0.239	0.095
43	0.17	0.043	0.128	0.043	0.085

표 4. 기본 선호도 평가를 배제한 엔트로피
Table 4. Entropy applying Not-Default Voting.

아이템	3	11	29	35	43
E(p_i)	1.836	1.24	0.945	1.182	0.945

표 5. 확률 벡터에 의한 테이블
Table 5. Table according to Probability Vector.

아이템	이웃 선호				
	U_1	U_2	U_3	U_4	U_5
3	0.046	0.184	0.092	0.092	0.138
11	0.155	0.062	0.124	0.124	
29			0.071		
35	0.059	0.148		0.148	0.059
43	0.095	0.024		0.024	0.047

시 사용자 별 아이템에 표현된 확률 벡터들과의 곱함으로써 표 3과 같은 새로운 테이블을 얻어낼 수 있다.

표 4의 테이블은 표 1에서 기본 선호도를 부여하지 않고, 확률 벡터에 의해 구해진 테이블에서 엔트로피를 측정했을 때이다. 기본 선호도 평가를 배제한 엔트로피는 표 4와 같다.

표 5는 기본 선호도 평가를 부여하지 않고 구해진 엔트로피와 사용자 별 아이템 확률 벡터를 곱해서 구해진 결과 테이블이다. 빈 공간은 사용자가 선호도를 표시하지 않은 부분이다.

Ⅳ. 성능평가

1. 실험 환경 및 실험 데이터

본 논문에서 제안한 협력적 필터링 기술에서 정보검색 분야의 벡터 유사도, 엔트로피, 역 사용자 빈도, 기본 선호도 평가를 적용하여 유사도 가중치 알고리즘을 MS-Visual Studio C++ 6.0으로 구현되었으며, 실제 실험 환경은 PentiumIII 450MHz, 256MB RAM 환경에서 수행되었다. 실험 데이터로는 컴팩 연구소에서 18개월 동안 협력적 필터링 알고리즘을 연구하기 위해서 영화(아이템)에 대한 사용자의 선호도를 조사한 EachMovie 데이터[20]를 사용한다. 이 데이터에는 아이템에 관한 사용자의 평가 데이터가 0, 0.2, 0.4, 0.6, 0.8, 1의 6단계의 수치로 표현되어 있다. 여기서 6단계는 입력 허용 범위를 나타낸다. 즉, 0은 주어진 아이템에 대해서 매우 부정적인 선호도(Negative rating)를 의미하고, 1은 매우 긍정적인 선호도(Positive rating)를 의미한다. 본 논문에서는 개인화 추천 시스템의 예측 정확도 향상을 위한 사용자 유사도 가중치에 대한 실험하기 위해서 EachMovie 데이터를 전처리[20]하여 사용한다. 계산상의 편의와 메모리 절약을 위해 0-1까지 0.2간격으로 표현된 데이터에 5를 곱하여 0-5까지 1간격으로 변환하여 사용하였다. 최소 100회 이상 선호도를 평가한 사용자 4,798명을 추출하여 이 가운데 1,000명을 기존 사용자 군으로 두고 나머지 사용자들 중에 무작위로 테스트 사용자 100명을 선택하여 총 1,628개의 아이템 중 테스트 사용자가 선호도를 평가한 임의의 10개 아이템에 대해서 선호도를 예측하고 실제 선호도와 비교, 평가하였다.

2. 성능 평가 기준

추천의 성능을 평가하기 위한 방법으로 본 논문에서는 MAE(Mean Absolute Error)와 순위 스코어 측정(Rank scoring metric)을 사용한다[10,11,12]. MAE는 단일 아이템에 대한 추천을 평가하는데 사용하며 순위 스코어 측정은 순위가 있는 아이템의 목록을 추천하는 시스템의 성능을 평가하는데 사용한다. MAE에서 예측의 정확도는 실제로 사용자가 평가한 값과 예측된 값의 차이에 대한 절대값의 평균을 나타낸다. MAE는 절대적으로 알고리즘이 얼마나 정확하게 예측을 했는지를 알 수 있으며 식 (14)에 의해 정의된다.

$$S_a = \frac{\sum_{j \in p_a} |P_{a,j} - v_{a,j}|}{m_a} \tag{14}$$

식 (14)에서 P_{aj}는 예측된 선호도이며 v_{aj}는 실제로 사용자가 평가한 선호도이다. 또한 m_a는 새로운 사용자에 의해 평가된 아이템의 수를 의미한다.

순위 스코어 측정은 순위가 있는 아이템의 목록을 사용자가 평가하는 가의 측정이다. 순위 스코어 측정은 아이템을 선택할 확률이 목록의 하단으로 갈수록 지수적으로 감소한다는 전제에서 측정된다. 각 아이템은 사용자 선호도 가중치에 따라 j에 의해 내림차순 정렬되어 있다고 가정한다. 식 (15)는 순위가 부여된 아이템의 목록에 대하여 사용자 U_a의 순위 스코어 측정에 대한 기대 이용도(Expected utility)를 계산하기 위한 식이다.

$$R_a = \sum_j \frac{max(v_{a,j} - d, 0)}{2^{(j-1)/(\alpha-1)}} \tag{15}$$

식 (15)에서 d는 아이템에 대한 중간 평가 값이며 α는 반감기(halflife)이다. 반감기는 사용자가 평가하거나 방문할 50-50의 기회가 있는 목록에 있는 아이템의 수이다. 본 논문의 평가에서 반감기를 5로 사용한다[15].

$$R=\frac{\sum_{i=1}^{u}R_i}{\sum_{i=1}^{u}max(R_i)}\times 100 \qquad (16)$$

식 (16)에서 max(R_i)는 사용자가 평가한 아이템이 순위가 있는 목록상에서 상위에 나타났을 경우에 측정된 순위 스코어 측정에 대한 기대 이용도의 최대값이다[9].

3. 분석 및 성능 평가

본 논문에서는 평가를 위해 정보검색 분야에서의 벡터 유사도를 이용한 유사도 가중치(Vec_sim), 엔트로피를 이용한 유사도 가중치(Ent_sim), 역 사용자 빈도를 이용한 유사도 가중치(IUF_sim)와 기존의 협력적 필터링 알고리즘에서 많이 쓰이는 피어슨 상관계수를 이용한 유사도 가중치(P_Corr)의 관계를 사용자의 수를 변화시키면서 성능을 비교하였다. 또한 벡터 유사도, 엔트로피, 역 사용자 빈도를 이용한 유사도 가중치에 기본 선호도 평가를 이용한 유사도 가중치를 적용하여 성능 평가(Vec_sim_def, Ent_sim_def, IUF_sim_def, P_Corr_def)를 하였다. 이때 2.1절의 z 값 가중치 평균을 이용하여 사용자가 아이템에 대해 평가한 횟수를 변화시켜가면서 비교하였다.

그림 3와 그림 4는 식 (14)과 식 (16)를 기반으로 사용자 수를 변화시킴에 따른 P_Corr, Vec_sim, Ent_sim, IUF_sim의 MAE와 순위 스코어 측정을 나타낸 것이다.

그림 3과 그림 4은 사용자들의 수가 많아짐에 따라 Vec_sim와 Ent_sim의 성능은 높아지나 P_Corr, IUF_sim를 이용한 방법은 큰 차이가 없음을 나타낸다. 예측의 정확도는 엔트로피를 이용한 유사도 가중치(Ent_sim)가 벡터 유사도를 이용한 유사도 가중치

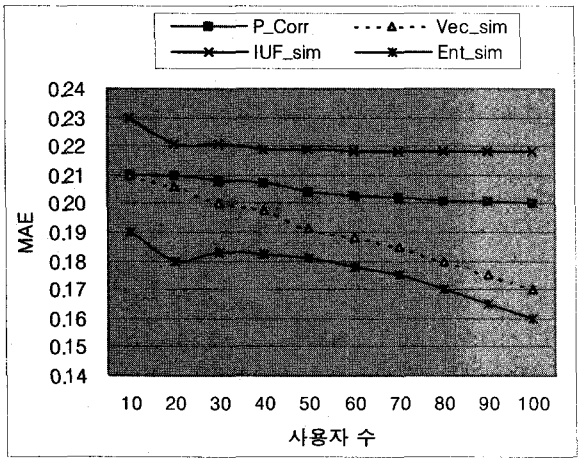

그림 3. 사용자 수의 변화에 따른 MAE
Fig. 3. MAE by varying the number of users.

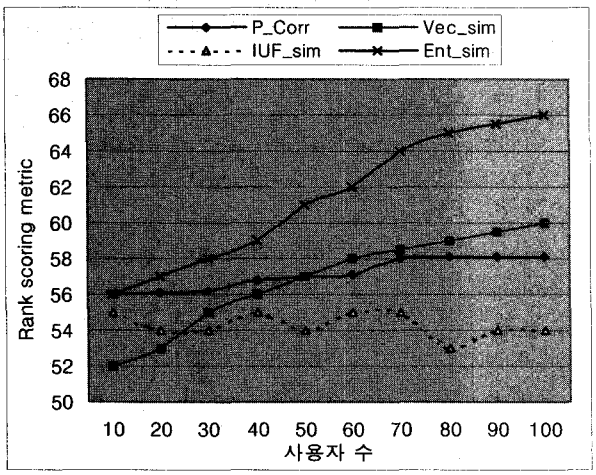

그림 4. 사용자 수의 변화에 따른 순위 스코어 측정
Fig. 4. RSM by varying the number of users.

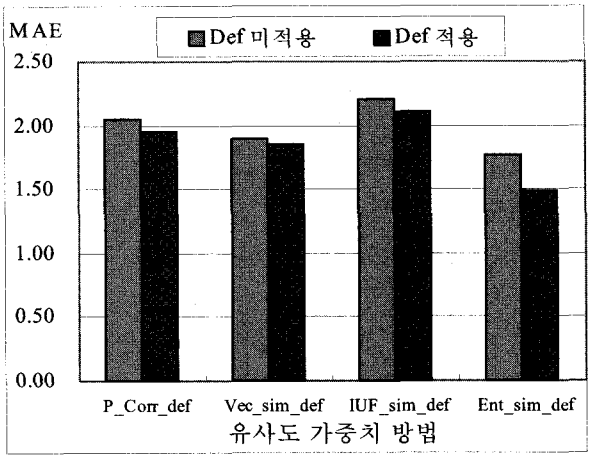

그림 5. 사용자 수의 변화에 따른 MAE
Fig. 5. MAE by varying the number of users.

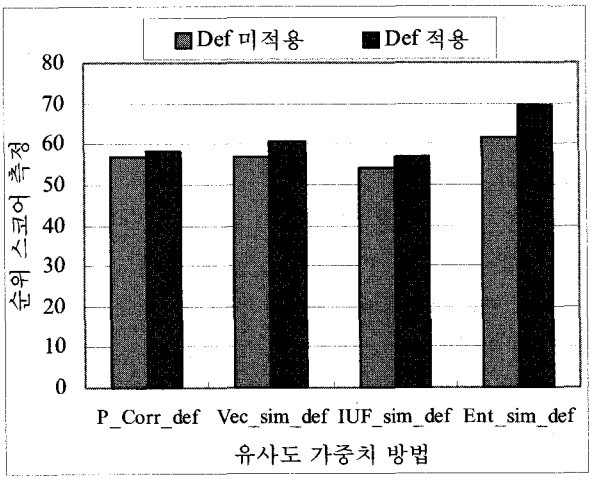

그림 6. 사용자 수의 변화에 따른 순위 스코어 측정
Fig. 6. RSM by varying the number of users.

(Vec_sim)보다 우세함을 알 수 있다.

그림 5와 그림 6는 식 (14)과 식 (16)를 기반으로 아이템에 대해 평가한 횟수를 증가시킴에 따른 GroupLens의 확장된 접근 방법인 z 값 가중치 평균을 사용한 개인화 추천 시스템에서 MAE와 순위 스코어 측정을 나타낸다. 이는 피어슨 상관계수, 벡터 유사도, 엔트로피, 역 사용자 빈도를 이용한 유사도 가중치에 기본 선호도 평가를 적용한 유사도 가중치의 성능을 나타낸다.

그림 5와 그림 6에서 피어슨 상관계수, 벡터 유사도, 역 사용자 빈도, 엔트로피를 이용한 유사도 가중치에 기본 선호도 평가를 적용하면 전체적으로 성능이 향상된 것을 볼 수 있다. 성능 향상의 차이를 보면 엔트로피를 이용한 유사도 가중치에 기본 선호도 평가를 적용한 방법 (Ent_sim_def)이 15.5%의 성능 향상으로 다른 방법들{P_Corr_def (4.49%), Vec_sim_def (2.31%), IUF_sim_def (4.13%)}보다는 성능이 향상됨을 알 수 있다. 그러므로 협력적 필터링 기술에서 사용자들간의 유사도 가중치를 계산할 때 Ent_sim_def을 적용하는 것이 가장 바람직하다.

본 논문에서 제안한 방법은 엔트로피에 기본 선호도 평가 값을 적용함으로써 서론에서 제시한 피어슨 상관계수 첫번째 문제점인 두 사용자 사이의 상관관계는 오직 두 사용자 모두 선호도를 평가한 아이템에 대해서만 계산되는 문제점을 해결하였다. 기본 선호도 평가를 피어슨 상관계수, 벡터 유사도, 엔트로피, 역 사용자 빈도를 이용한 유사도 가중치에 적용한 결과 MAE와 순위 스코어 측정에서 본 논문에서 제안한 엔트로피에 기본 선호도 평가 값을 적용하는 방법이 가장 성능이 좋음을 알 수 있다. 그리고 피어슨 상관계수의 두번째 문제점과 세번째 문제점은 엔트로피를 이용한 유사도 가중치로 해결하였다. 엔트로피는 속성의 빈도수를 기반으로 군집 내의 응집도가 높은 속성을 발견하여 가중치를 높게 부여하는 방법을 사용한다. 그리고 군집간에 공통으로 분포하는 속성 가중치는 낮추고, 특정 군집에만 포함된 속성의 가중치를 높이는 효과가 있다. 이러한 효과로 피어슨 상관계수의 문제점을 해결하였다.

V. 결 론

협력적 필터링 기술에서 사용자의 성향에 맞는 아이템을 예측하고 추천하는데 사용자 유사도 가중치가 성능에 영향을 미친다. 대부분의 협력적 필터링 기술을 이용한 개인화 추천 시스템에서 피어슨 상관계수 기반의 예측 기법을 사용한다. 본 논문에서는 기존의 사용자 유사도 가중치를 계산하는 방법에 정보검색 분야의 벡터 유사도, 엔트로피, 역 사용자 빈도, 기본 선호도 평가를 적용한 유사도 가중치에 대해 예측 정확도 향상에 관한 실험을 하였다. 실험 결과는 엔트로피를 이용한 유사도 가중치에 기본 선호도 평가를 결합하는 방법이 성능이 가장 우수함을 알 수 있다. 이 방법은 아이템이 갖는 평균 정보량인 엔트로피를 확률 벡터 연산으로 사용한 것이다. 속성의 엔트로피 가중치 방법은 속성의 빈도수를 기반으로 군집 내에 응집도가 높은 속성을 발견하여 가중치를 높게 부여하는 방법이다. 이렇게 함으로써 군집간의 공통으로 분포하는 속성 가중치를 낮추고, 특정 군집에만 포함된 속성의 가중치를 높이는 효과가 있다. 그리고 선호도를 평가하지 않은 새로운 아이템에 대해서 기본 선호도 평가 값을 우선적으로 적용함으로써 개인화 추천이 가능하다.

참 고 문 헌

[1] M. J. Pazzani, "A Framework for Collaborative, Content-Based and Demographic Filtering," Artificial Intelligence Review, Vol. 13, no. 5-6, pp. 393-408, 1999.

[2] B. Sarwar, et. al., "Item-based Collaborative Filtering Recommendation Algorithms," WWW10 Conference, pp. 285-295, 2001.

[3] D. Billsus, M. J. Pazzani, "Learning Collaborative Information Filters," in Proc. of ICML, pp. 46-53, 1998.

[4] J. Konstan, et. al., "GroupLens: Applying Collaborative Filtering to Usenet News." Communication of the ACM, Vol. 40, no. 3, pp. 77-87, 1997.

[5] P. Resnick, et. al., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," in Proc. of ACM CSCW'94 Conference on Computer Supported Cooperative Work, pp. 175-186, 1994.

[6] 정경용, 김진현, 나영주, "소재 설계를 위한 감성공학적 Textile 기반의 협력적 필터링 시스템," 한국섬유공학회 · 한국의류학회 · 한국염색가공학회 공동 추계학술대회 논문집, 2002.

[7] K. Y. Jung, Y. J. Na, J. H. Lee, "FDRAS: Fashion Design Recommender Agent System using the Extraction of Representative Sensibility and the Two-Way Filtering on Textile," LNCS 2736, Springer-Verlag, pp. 631-640, 2003.

[8] 정경용, 김진현, 나영주, *소재 설계를 위한 감성 공학적 디자인 지원시스템 개발*, 최종 연구 개발 보고서, 한국학술진흥재단, 2003.

[9] B. Sarwar, et. al., "Analysis of Recommendation Agorithms for E-Commerce," in proc. of the ACM E-Commerce 2000 Conference, 2000.

[10] 정경용, 최성용, 임기욱, 이정현, "베이지안 추정치가 부여된 유사도 가중치와 연관 사용자 군집을 이용한 선호도 예측 시스템", 정보과학회논문지 : 소프트웨어 및 응용, 제30권, 제4호, pp. 316-325, 2003.4.

[11] 정경용, 류중경, 강운구, 이정현, "내용 기반 여과와 협력적 여과의 병합을 통한 추천 시스템에서 조화 평균 가중치," 정보과학회논문지: 소프트웨어 및 응용, 제30권, 제3호, pp. 239-250, 2003.

[12 정경용, 협력적 여과 시스템에서 연관 사용자 군집과 베이지안 추정치를 이용한 예측 방법, 인하대학교 대학원, 석사학위논문, 2002.

[13] 정경용, 김진수, 김태용, 이정현, "선호도 재계산을 위한 연관 사용자 군집 분석과 Representative Attribute-Neighborhood을 이용한 협력적 필터링 시스템의 성능향상," 한국정보처리학회(B), 제10-B권, 제3호, pp. 287-296, 2003.

[14] K. Y. Jung, D. H. Park, J. H. Lee, "Hybrid Collaborative Filtering and Content-based Filtering for Improved Recommender System," LNCS 3036, Springer-Verlag, pp. 295-302, 2004.

[15] K. Y. Jung, J. H. Lee, "Prediction of User Preference in Recommendation System using Association User Clustering and Bayesian Estimated Value," LNAI 2557, Springer-Verlag, pp. 284-296, 2002.

[16] T. Michael, Maching Learning, McGraq-Hill, pp. 154-200, 1997.

[17] 정영미, *정보검색론*, 구미무역 출판부, 1993.

[18] 강창언, 오용선, 이명호, *정보이론 - 토딩 이론과의 접목*, 생능사, pp. 233-285, 1987.

[19] R. E. Blahut, Principles and Practice of Information Theory, pp. 12-18, 1991.

[20] P. McJones, EachMovie, www.research.digital.com/SRC/eachmovie, 1997.

저 자 소 개

정 경 용(정회원)

1996년 인하대학교 전자계산 공학과 (공학사)
2000년 인하대학교 컴퓨터정보 공학과 (공학석사)
2002년 인하대학교 컴퓨터정보 공학과 (공학박사)
2001년~2005년 에이플러스전자 책임연구원
2002년~2005년 가천길대학 겸임교수
2005년~현재 한세대학교 IT학부 교수
<주관심분야 : 데이터마이닝, HCI, 감성공학, 임베디스 시스템, 컴퓨터구조>

이 정 현(평생회원)

1977년 인하대학교 전자공학과
1980년 인하대학교 대학원 전자공학과 (공학석사)
1988년 인하대학교 대학원 전자공학과 (공학박사)
1979년~1981년 한국전자기술 연구소 시스템연구원
1984년~1989년 경기대학교 교수
1989년~현재 인하대학교 컴퓨터공학부 교수
<주관심분야 : 자연어처리, HCI, 정보검색, 컴퓨터 구조, 음성인식, 음성합성, 임베디드시스템>