베이지안 추정치가 부여된 유사도 가중치와 연관 사용자 군집을 이용한 선호도 예측 시스템

(Preference Prediction System using Similarity Weight granted Bayesian estimated value and Associative User Clustering)

정 경 용 † 최 성 용 † 임 기 욱 †† 이 정 현 †††

(Kyung-Yong Jung) (Seong-Yong Choi) (Kee-Wook Rim) (Jung-Hyun Lee)

요 약 기존의 협력적 필터링 기술을 이용한 사용자 선호도 예측 방법에서는 피어슨 상관 계수에 의해 사용자의 유사도를 구하고, 아이템에 대한 사용자의 선호도를 기반으로 이웃 선정 방법을 사용하므로 아이템에 대한 내용을 반영하지 못할 뿐만 아니라 희박성 문제를 해결하지 못하였다. 본 논문에서는 기존의 사용자 선호도 예측 방법의 문제점을 보완하기 위하여 베이지안 추정치가 부여된 유사도 가중치와 연관 사용자 군집을 이용한 선호도 예측 시스템을 제안한다. 제안한 방법에서는 협력적 필터링 시스템에서의 희박성 문제를 해결하기 위하여 Association Rule Hypergraph Partitioning 알고리즘을 사용하여 사용자를 장르별로 군집하며 새로운 사용자는 Naive Bayes 분류자에 의해 이들 장르 중 하나로 분류된다. 또한, 분류된 장르 내에 속한 사용자들과 새로운 사용자의 유사도를 구하기 위해 Naive Bayes 학습을 통해 사용자가 평가한 아이템에 추정치를 달리 부여한다. 추정치가 부여된 선호도를 기존의 피어슨 상관 관계에 적용할 경우 결측치(Missing Value)로 인한 예측의 오류를 적게하여 예측의 정확도를 높일 수 있다. 제안된 방법의 성능을 평가하기 위해서 기존의 협력적 필터링 기술과 비교 평가하였다. 그 결과 기존의 협력적 필터링 기술의 문제점을 해결하여 예측의 정확도를 높이는데 효과적임을 확인하였다.

키워드 : 협력적 여과, 내용기반 여과, CRM, 네이브 베이지안, 추천 시스템, 인지공학

Abstract ABSTRACT A user preference prediction method using an exiting collaborative filtering technique has used the nearest-neighborhood method based on the user preference about items and has sought the user's similarity from the Pearson correlation coefficient. Therefore, it does not reflect any contents about items and also solve the problem of the sparsity. This study suggests the preference prediction system using the similarity weight granted Bayesian estimated value and the associative user clustering to complement problems of an exiting collaborative preference prediction method. This method suggested in this paper groups the user according to the Genre by using Association Rule Hypergraph Partitioning Algorithm and the new user is classified into one of these Genres by Naive Bayes classifier to slove the problem of sparsity in the collaborative filtering system. Besides, for get the similarity between users belonged to the classified genre and new users, this study allows the different estimated value to item which user vote through Naive Bayes learning. If the preference with estimated value is applied to the exiting Pearson correlation coefficient, it is able to promote the precision of the prediction by reducing the error of the prediction because of missing value. To estimate the performance of suggested method, the suggested method is compared with existing collaborative

† 비 회 원 : 인하대학교 전자계산공학과
kyjung@gcgc.ac.kr
sychoi@nlsun.inha.ac.kr
†† 종신회원 : 선문대학교 산업공학과 교수
rim@omega.sunmoon.ac.kr
††† 종신회원 : 인하대학교 컴퓨터공학부 교수
jhlee@inha.ac.kr
논문접수 : 2002년 2월 15일
심사완료 : 2003년 1월 28일

filtering techniques. As a result, the proposed method is efficient for improving the accuracy of prediction through solving problems of existing collaborative filtering techniques.

Key words : Collaborative Filtering, Content Based Filtering, CRM, Naïve Bayes, Recommender System, Aesthetic Engineering

1. 서 론

전자 상거래의 협력적 필터링 시스템이 증가하고 그 이용량이 증가하면서, 차별화된 고객 서비스를 위해 다양한 데이타 마이닝 기술과 정보 검색(IR) 기술들이 적용되고 있다. 대부분의 협력적 필터링 시스템들은 아이템의 수가 많아질수록 사용자가 아이템에 관련된 정보를 얻는데 어느 정도 한계가 있기 때문에 같은 아이템에 대해서 두 사용자간에 선호도를 표시할 확률은 적어지게 되고, 상관관계를 비교 할 아이템의 수는 증가하게 된다. 이러한 협력적 필터링 시스템은 세 가지 단점을 갖는다[1,2,3,4]. 첫째 사용자가 평가를 하지 않은 아이템들은 사용자에게 추천되지 않는 다는 초기 평가 문제(Early-Rater Problem)이다. 둘째, 대부분의 사용자들은 모든 상품에 대해 평가하지 않기 때문에 사용자-아이템의 데이타 집합은 희박한 특성(Sparsity Problem)을 보인다는 것이다. 셋째, 아이템의 속성에 대한 사용자의 선호도를 직접적으로 반영하지 못하는 문제점도 있다. [5,6,7,8]의 방법들은 초기 평가 문제를 해결하였으나 희박성 문제를 해결하지 못하였다. LSI[9], SVD[1] 분류를 사용한 방법은 데이타 차원의 수를 줄임으로써 협력적 필터링의 희박성 문제를 해결하였으나 초기 평가 문제는 해결하지 못하였다. 반면, [10]의 방법에서는 초기 평가 문제와 희박성 문제를 동시에 해결하려는 시도를 하였다.

최근에는 협력적 필터링 기술에서 고려하기 힘든 부분에 대해서 내용 기반 필터링을 이용함으로써 문제점을 해결한다[7]. 보다 좋은 성능을 얻기 위해서는 이러한 필터링 기법들을 결합하고 보완할 필요가 있다. 내용 기반 필터링과 협력적 필터링을 결합하여 더 좋은 예측 결과를 얻고자 하는 연구가 이루어지고 있다[5].

본 논문에서는 기존의 협력적 필터링 시스템에서의 희박성 문제를 해결하기 위하여 Association Rule Hypergraph Partitioning(ARHP) 알고리즘[11,12]을 이용한다. 사용자 트랜잭션에 Apriori 알고리즘[13]을 적용하여 연관규칙과 신뢰도를 구한 후, 연관규칙에 포함되는 사용자를 Vertex로, 연관관계를 Hyperedge로 매핑한다. 신뢰도를 Hypergraph Partitioning을 위한 가중치로 하여 사용자를 군집시킨다. 새로운 사용자는 Naive Bayes 분류자에 의해 이들 장르 중 하나로 분류된다. 또한, 분류된 장르 내에 속한 사용자들과 새로운 사용자의 유사도를 구하기 위해 Naive Bayes 학습을 통해 사용자가 평가한 아이템에 추정치를 달리 부여한다. 추정치가 부여된 선호도를 기존의 피어슨 상관 관계에 적용할 경우 결측치(Missing Value)로 인한 예측의 오류를 적게하여 예측의 정확도를 높일 수 있다. 제안된 방법의 성능을 평가하기 위해서 기존의 협력적 필터링 기술과 비교 평가하였다.

2. 기존의 사용자 유사도 가중치 방법

사용자와 유사한 선호도를 가지는 이웃을 찾아내고 사용자간에 선호도를 표시한 아이템의 선호도를 예측하기 위해서 사용되는 유사도 기준값으로는 대표적으로 Correlation, Vector based similarity, Default voting, Inverse User frequency 등이 사용된다[14,15].

2.1 Correlation

통계적인 협력적 필터링의 일반적인 형태로 Pearson correlation coefficient나 Spearman rank correlation coefficient를 사용한다[3].

Pearson correlation coefficient를 사용했을 경우 사용자 a와 사용자 i의 유사도 가중치는 식(1)과 같이 정의된다.

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

Pearson correlation coefficient는 모든 독립변수에 대해서 데이타가 선형의 관계이고 에러가 평균이 0이고 등분산을 가지는 확률분포를 가지며 독립이라는 가정에 의존하는 선형 회귀 모델에서 생성된다. 이러한 가정이 만족되지 못할 경우 Pearson correlation coefficient는 유사도의 척도로 정확하지 않을 수도 있다. Spearman rank correlation coefficient는 Pearson correlation coefficient와 비슷하지만 Correlation을 구할 때 실제 선호도 값을 사용하는 것이 아니라 선호도 값의 순위를 사용하므로 모델에 대한 가정에 의존하지 않는다.

Spearman rank correlation coefficient를 사용했을 경우 사용자 a와 사용자 i의 유사도 가중치는 식(2)와 같이 정의한다.

$$w(a,i)=\frac{\sum_j (rank_{a,j}-\overline{rank_a})(rank_{i,j}-\overline{rank_i})}{\sqrt{\sum_j (v_{a,j}-\overline{v_a})^2 \sum_j (v_{i,j}-\overline{v_i})^2}} \quad (2)$$

Spearman rank correlation coefficient인 경우 순위가 동률인 개수가 많을 경우 정확도가 감소되는 경향이 있는데 명시적인 선호도 데이타에서 순위가 동률인 선호도 값이 많이 존재할 수 있어 Pearson correlation coefficient의 경우보다 정확도가 떨어지기도 한다. 반면 구매 횟수와 같이 값의 범위가 명시적 선호도보다 넓은 암묵적인 선호도 데이타의 경우에는 좋은 결과를 보일 수 있다.

2.2 Vector similarity

정보 검색의 분야에서 두 개의 문서간의 유사도를 측정하기 위해서 두 문서 안에 포함된 특정 단어의 발생 빈도 벡터를 사용하여 Cosine vector similarity[16,17]를 구한다. 이러한 형태를 협력적 필터링에 적용하여 정보 검색 분야의 문서를 인터넷 사용자로 보고 포함된 특정 단어의 발생 횟수를 선호도로 보아 두 사용자 간의 유사도 가중치를 구할 수 있다[17]. Vector similarity를 사용했을 경우 개념적으로 명시적 선호도 데이타를 문서에 포함된 단어 발생 빈도로 간주하므로 선호도는 항상 긍정적인 선호도만이 존재하는 것으로 보아야 하며 선호도가 표시되지 않은 아이템의 경우에는 가정 선호도가 낮은 아이템으로 보아야 한다. 이러한 개념 때문에 Vector similarity[17,18]는 암묵적인 선호도 데이타를 다룰 경우에 사용자간의 유사도 가중치로 사용되어야 할 것이다.

Vector similarity를 사용했을 경우 사용자 a와 사용자 i의 유사도 가중치는 식(3)과 같이 정의한다.

$$w(a,i)=\sum_j \frac{v_{a,j}}{\sqrt{\sum_{k\in I_a} v_{a,k}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k\in I_i} v_{i,k}^2}} \quad (3)$$

2.3 Default voting

Default voting은 Correlation기반의 알고리즘을 확장한 것으로 기준이 되는 특정 사용자와 그 사용자의 유사도가 있는 사용자들이 공통으로 선호도를 보이는 아이템이 상대적으로 적을 경우 사용한다. Correlation 알고리즘이 사용자 a와 사용자 i가 공통으로 선호도를 보인 아이템들($I_a \cap I_i$)을 이용하는데 반해 Default voting처럼 선호도를 받지 못한 아이템에 대해서 기본값을 적용하면 사용자 a와 사용자 i중 한사람이라도 선호도를 보인 아이템들($I_a \cup I_i$)을 사용할 수 있다[19].

Default voting은 더 나아가 사용자중 어떤 사람도 선호도를 입력하지 않은 새로운 아이템들에 대해서 기본값을 우선적으로 적용함으로써 추천이 가능하도록 할 수 있다. 대부분의 경우 Default voting값 d는 중립적이거나 다소간 비선호의 값을 사용하는 경우가 많다.

Correlation에 대해서 Default voting을 적용하면 사용자 a와 사용자 i의 유사도 가중치는 식(4)와 같다.

$$w(a,i)=\frac{(n+k)(\sum_j v_{a,j}v_{i,j}+kd^2)-(\sum_j v_{a,j}+kd)(\sum_j v_{i,j}+kd)}{\sqrt{((n+k)(\sum_j v_{a,j}^2+kd^2)-(\sum_j v_{a,j}+kd)^2)((n+k)(\sum_j v_{i,j}^2+kd^2)-(\sum_j v_{i,j}+kd)^2)}} \quad (4)$$

2.4 Inverse user frequency

정보 검색 분야에서 Vector similarity[21]를 이용하는데 있어 문서에 포함된 단어의 수는 단어의 빈도의 역수[20]에 의해 수정된다. 공통적으로 많이 발생하는 단어의 경우 문서를 분류해내는데 유용한 역할을 할 수 없다는 판단에 따라 공통적으로 많이 발생하는 단어에 대한 가중치를 줄이는 것이다. 협력적 필터링에 이러한 개념을 적용해서 일반적으로 사용자에 의해서 많이 선호도가 입력되는 아이템은 적게 선호도가 입력되는 아이템에 비해서 사용자간의 차이를 보여주는데 덜 기여한다고 판단하여 가중치를 줄인다. 이를 위해 f_j를 정의하는데 f_j는 $\log\frac{n}{n_j}$로 정의되고 n_j는 아이템 j에 선호도를 입력한 사용자 수이며, n은 전체 사용자의 수이다. 따라서 모든 사용자에 의해서 선호도가 입력된 아이템 j에 대해서 $f_j = \log 1 = 0$이 되어 모든 사용자에 의해서 선호도 입력된 아이템 j는 사용자의 유사도 가중치를 구하는데 사용되지 않게 된다[17]. Inverse user frequency f_j를 피어슨 상관계수에 적용한 사용자 a와 사용자 i의 유사도 가중치는 식(5)와 같다.

$$\frac{w(a,i)=\sum_j f_j(\sum_j f_j v_{i,j}-(\sum_j f_j v_{a,j})(\sum_j f_j v_{i,j}))}{\sqrt{UV}} \quad (5)$$

$$\text{단 } U=\sum_j f_j(\sum_j f_j v_{a,j}^2)-(\sum_j f_j v_{a,j})^2$$

$$V=\sum_j f_j(\sum_j f_j v_{a,j}^2)-(\sum_j f_j v_{a,j})^2$$

Inverse User frequency 를 Vector Similarity에 적용하여 구한 사용자 a와 사용자 i의 유사도 가중치는 식(6)과 같다.

$$w(a,i)=\sum_j \frac{f_j v_{a,j}}{\sqrt{\sum_{k\in I_a}(f_j v_{a,k})^2}} \frac{f_j v_{i,j}}{\sqrt{\sum_{k\in I_i}(f_k v_{i,k})^2}} \quad (6)$$

3. 사용자 선호도 예측 시스템

서론 부분에서 설명한 기존의 협력적 필터링 기술을 이용한 사용자 선호도 예측 방법에서는 아이템에 대한 사용자의 선호도를 기반으로 이웃 선정 방법을 사용하고, 피어슨 상관계수에 의해 사용자의 유사도를 구하므로 아이템에 대한 내용을 반영하지 못할 뿐만 아니라 희박성 문제를

해결하지 못하였다[1,2,3,4]. 본 논문에서는 기존의 사용자 선호도 예측 방법의 문제점을 보완하기 위해서 베이지안 추정치를 부여한 유사도 가중치와 연관 사용자 군집을 이용한 선호도 예측 시스템을 제안한다.

제안한 방법에서는 협력적 필터링 기법에서의 희박성 문제를 해결하기 위해서 Association Rule Hypergraph Partitioning 알고리즘[11,12]을 이용하여 사용자 행렬의 차원수를 감소시킨다. 사용자 트랜잭션에 Apriori 알고리즘[13]을 적용하여 연관규칙과 신뢰도를 구한 후, 연관규칙에 포함되는 사용자를 Vertex로, 연관관계를 Hyperedge로 매핑한다. 신뢰도를 Hypergraph Partitioning을 위한 가중치로 하여 사용자를 군집시킨다. 또한 새로운 사용자는 Naive Bayes 분류자[21]에 의해 장르 중 하나로 분류하여 새로운 사용자 프로파일을 생성함으로써 초기 평가 문제를 해결한다[4]. 추정치가 부여된 학습 집단에서 분류된 장르에 따라 아이템에 대한 사용자의 선호도의 가중치를 달리 부여하여 결측치 값에 아이템의 정보를 반영한다. 이는 아이템의 속성에 대한 사용자의 선호도를 직접적으로 반영하지 못하는 문제점을 해결한다.

그림 1은 본 논문에서 설계한 베이지안 추정치가 부여된 유사도 가중치와 연관 사용자 군집을 이용한 선호도 예측 시스템에 대한 구성도로서 세부 단계의 작업 과정은 다음과 같다.

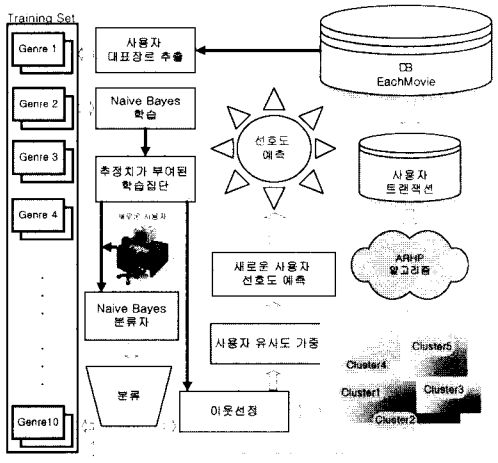

그림 1 사용자 선호도 예측 시스템에 대한 구성도

3.1 연관 사용자 군집 및 사용자의 대표장르 추출

3.1.1 데이타 정제 작업

사용자의 대표장르 추출 및 연관 사용자 군집을 하기 위해서 EachMovie 데이타[22]를 사용한다. 이 데이타는 사용자가 선호도를 표시한 아이템들이 영화별 장르 정보에 속하지 않는 것이 있고, 사용자의 정보 또한 누락이 된 것이 있다. 이를 해결하기 위해서 데이타 정제 작업을 하여 데이타의 무결성을 검사하였다[23].

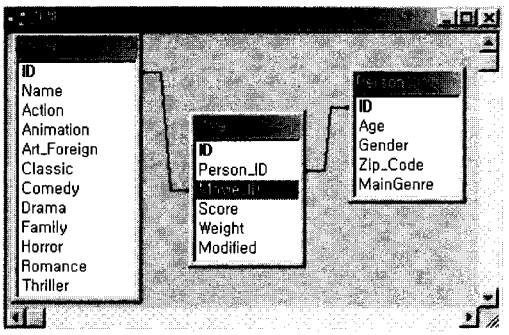

그림 2 Person, Vote, Movie 테이블의 관계

그림 2와 같이 Vote 테이블을 기준으로 Movie와 Person 테이블의 무결성 검사를 하였고, Person과 Movie 테이블을 기준으로 Vote 테이블의 무결성 검사를 하였다.

3.1.2 ARHP 알고리즘에 의한 연관 사용자 군집

Association Rule Hypergraph Partitioning 알고리즘은 연관 규칙과 Hypergraph Partitioning을 이용하여 트랜잭션 기반의 데이타 베이스에서 연관된 항목들을 클러스터링 하는 방법이다[11,12]. Hypergrah H={V, E}는 사용자들로 구성된 정점(vertex)들의 집합 V와 빈번한 항목 집합들을 나타내는 Hyperedge들의 집합 E로 구성된다. Hypergraph Partitioning 알고리즘은 항목들간의 거리가 아닌 가중치를 이용하기 때문에 항목들간의 거리 계산이

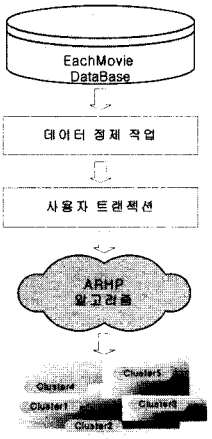

그림 3 연관사용자 군집의 개념도

어려운 다차원 데이타 집합에 대한 클러스터링에 유용하다. 이때 Hypergraph Partitioning을 위한 가중치로 연관 규칙의 신뢰도를 사용한다. ARHP 알고리즘에 의한 연관 사용자 군집을 위한 단계적 흐름은 그림 3와 같이 진행된다. 사용자에 의해 선호도가 표시된 아이템들을 사용자 트랜잭션으로 재구성한다[11,12]. 이를 연관 규칙 탐사 방법[13]을 이용하여 사용자 트랜잭션 안에 빈번하게 동시에 출현하는 사용자들의 집합을 찾는다. 사용자들에 대한 Large 항목집합을 가지고 Apriori 알고리즘[13]을 이용하여 연관 규칙과 신뢰도를 구한 후, 연관규칙에 포함되는 항목을 vertex로, 연관 관계를 Hyperedge로 매핑한다. 그리고 신뢰도를 Hypergraph Partitioning을 위한 가중치로 사용자들 간의 군집을 만든다[12,23].

3.1.3 사용자의 대표 장르 추출

사용자가 선호도를 보인 아이템으로 사용자의 대표장르를 추출한다. 대표장르를 추출하기 위해서는 사용자의 장르별 아이템의 선호도 합을 구한 후 선호도의 합이 가장 큰 장르를 대표장르로 정한다. 알고리즘 1은 사용자의 대표장르를 추출하는 알고리즘이다. 사용자의 대표장르 추출은 훈련집합을 구성할 때 사용한다[4]. 그림 4는 사용자의 대표 장르 결정을 위한 개념도이다.

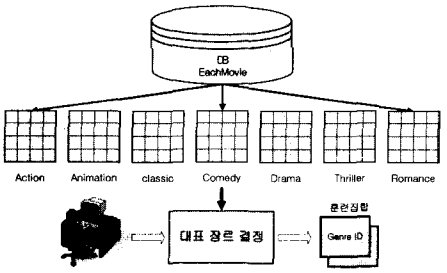

그림 4 사용자의 대표장르 결정 개념도

본 논문에서의 사용자의 대표장르 추출은 EachMovie 데이타의 영화를 평가한 고객정보에서 아이템에 대한 선호도의 합만을 고려한다.

```
Num_class ← # of item in GenreID;
MainGenreID ← Null;
MainGenreMaxSum ← 0;
For(j=1; j ≤ Num_class; j++){
    GenreMaxSum ( 0;
  For(each item){
      GenreMaxSum ← GenreMaxSum + Score;
} // item에 대해서 장르별 선호도의 합을 구한다.
    If (GenreMaxSum > MainGenreMaxSum){
          MainGenreID ← GenreID of j th;
          MainGenreMaxSum ← GenreMaxSum;
      } // 선호도의 합이 가장 큰 장르의 ID을 Return
}
Assign(MainGenreID);   // 대표장르 결정
```

알고리즘 1 사용자의 대표장르를 추출하는 알고리즘

3.2 Naive Bayes 알고리즘을 적용한 사용자 유사도 가중치

사용자들이 추정치가 부여된 학습집단을 구축하기 위해서는 우선 훈련 집합을 만들어야 한다. 훈련집합은 알고리즘 1에 의해서 사용자의 대표 장르를 구한 후 장르별 사용자 군집을 기반으로 아이템들을 장르별로 수집한 데이타이다. 훈련집합의 아이템에 추정치를 부여하기 위하여 Naive Bayes 학습 알고리즘[21]을 사용한다. 본 논문에서는 아이템의 발생여부만 사용하는 방법이 아닌 아이템의 출현빈도를 고려하는 다항 베이지안 학습법을 사용한다[21,23].

사용자 유사도 가중치는 사용자가 평가한 아이템의 선호도를 대상으로 장르별로 추정치를 다르게 적용한다. 이를 위해서 사용자가 선호도를 표시한 아이템에 추정치가 부여된 학습집단을 적용하면 장르별 아이템인 P(nuti|GenreID)에 Va,k를 곱한다. Va,k는 사용자 a와 아이템 k에 대해서 보여준 선호도이다.

$$\beta a,k = P(nuti|GenreID) \times Va,k \quad (7)$$

식(7)을 기반으로 피어슨 상관 계수[14,15]에 적용하면, 사용자 a와 사용자 i의 유사도 가중치는 식(8)과 같이 재 정의된다.

$$\beta(a,i) = \frac{Cov(a,i)}{\delta_a \cdot \delta_i} = \frac{\sum_k (\beta_{a,k} - \overline{\beta_a})(\beta_{i,k} - \overline{\beta_i})}{\sqrt{\sum_k (\beta_{a,k} - \overline{\beta_a})^2 \sum_k (\beta_{i,k} - \overline{\beta_i})^2}} \quad (8)$$

$\beta_{a,k}$는 사용자 a와 아이템에 대해서 가중치가 부여된 선호도이고, 는 사용자 a가 선호도를 입력한 아이템들에 대한 가중치가 부여된 선호도 평균값이다. $\overline{\beta_a}$는 사용자 a와 사용자 i가 공통으로 선호도를 입력한 아이템들이다.

```
/*Naive Bayes 학습을 이용한 훈련집합의 아이템에 추정치 부여*/
 TIS ← 전체 아이템의 수;
 Data ← 훈련 집합;
 For Each Class Variable{
          ut_k ← Item which class variable is;
          P(GenreID) ← |ut_k| / |Data|;
 //각 클래스에서 아이템이 출현할 확률
        n ← Total Number in GenreID;
     For each utk in TIS{
        n_k ← frequency of utk in n;
P(ut_k|GenreID) ← (n_k + 1) / (n + |TIS|);
        //아이템인 utk 에 추정치를 부여하기 위한 식
     }
 }
/*Naive Bayes 분류자를 이용한 새로운 사용자의 장르 분류*/
  Num_Class ← # of Item in GenreID;
  MainGenreID ← Null;
MainMaxG ← 0;
For(j=1; j≤ Num_Class; j++){
          // Naive Bayes 분류자에 의한 사용자 분류
For Each Item of Class {
```

```
Prob ← Prob +Cal_W(utk) * Calc_ P(nuti|GenreID));
}
G ← Calc_GenreID*Prob;
      If(G > MainMaxG){
           MainGenreID ← GenreID of i th;
          MainMaxG ← G;
    }      // 가장 큰 G값을 Return한다.
}
Assign(MainGenreID); // 새로운 사용자 대표 장르 결정
/*사용자 a와 사용자 i의 유사도 가중치*/
```

$\overline{\beta_a}$ 사용자 a의 가중치가 부여된 선호도의 평균값;

$\overline{\beta_i}$ 사용자 i의 가중치가 부여된 선호도의 평균값;

$P_1 \leftarrow P_2 \leftarrow P_3 \leftarrow 0$;

For Each Item k in($I_a \cap I_i$) {

$V_{a,k}$←사용자 a가 아이템 k에 대해서 보여준 선호도;

$V_{i,k}$←사용자 i가 아이템 k에 대해서 보여준 선호도;

$\beta_{a,k}=P(ut_k|GenreID)\times V_{a,k}$; $\beta_{i,k}=P(ut_k|GenreID)\times V_{i,k}$;

$P_1 \leftarrow P_1+(\beta_{a,k}-\overline{\beta_a})(\beta_{i,k}-\overline{\beta_i})$;

$P_2 \leftarrow P_2+(\beta_{a,k}-\overline{\beta_a})^2$;

$P_3 \leftarrow P_3+(\beta_{i,k}-\overline{\beta_i})^2$;

}

$\beta(a,i) \leftarrow P_1/\sqrt{(P_2 \times P_3)}$;

Assign($\beta(a,i)$); //사용자 a와 사용자 i의 유사도 가중치

알고리즘 2 Naive Bayes 알고리즘을 적용한 사용자 유사도 가중치

본 논문에서 제안한 Naive Bayes 알고리즘을 적용한 사용자 유사도 가중치는 기존의 협력적 여과 필터링에서 피어슨 상관계수를 이용한 유사도 가중치를 구하는 방식[3]에 Naive Bayes 학습 알고리즘[21]을 적용하는 방식이다.

3.3 새로운 사용자의 선호도 예측

새로운 사용자의 선호도 예측은 Naive Bayes 추정치를 적용한 사용자 a와 사용자 i의 유사도 가중치를 기존의 협력적 여과 필터링 기술의 피어슨 상관계수에 적용한다. 이는 사용자의 선호도만을 이용하는 것이 아닌 통계적인 값에 의해 가중치를 부여하기 때문에 예측의 정확도가 향상된다. 특정 아이템에 대한 이웃들의 선호도와 각 이웃들의 선호도 평균과의 거리를 이웃들과의 유사도로 가중 평균함으로써 특정 사용자의 아이템에 대한 선호도는 예측된다. 이를 수식으로 표현하면 식(9)와 같이 정의한다.

$$p_{a,k}=\overline{V_a}+\frac{\sum_{i=1}^{n}\beta(a,i)(v_{i,k}-\overline{v_i})}{\sum_{i=1}^{n}\beta(a,i)} \quad (9)$$

Pa.k는 사용자 a의 아이템 k에 대한 추정치가 부여된 선호도를 예측한 값이고, $\overline{v_a}$는 사용자 a의 가중치가 부여된 선호도 평균값이다. n은 사용자 a와 다른 사용자들간의 유사도가 0이 아닌 사용자 수이다. 기존의 협력적 여과 시스템에서 사용자 유사도 가중치를 구하는 방법은 사용자의 선호도만을 사용하여 계산하나, $\beta(a, i)$는 알고리즘 2에서 Naive Bayes 추정치를 이용한 사용자 유사도 가중치에 의해서 계산된다.

4. 실험 및 결과

4.1 실험 데이타

실험 데이타로는 컴팩 연구소에서 18개월 동안 협력적 필터링 알고리즘을 연구하기 위해서 영화에 대한 사용자의 선호도를 조사한 EachMovie 데이타[22]를 사용한다. 이 데이타는 총 72916명의 사용자와 1628종류의 영화에 대해서 0.0에서부터 1.0까지 0.2간격으로 명시적으로 평가한 선호도로 구성되어 있다. 또한 사용자가 실제로 영화를 보았는지의 여부를 알 수 있는 가중치 정보가 존재한다. 영화의 장르는 액션, 애니메이션, 외국예술, 고전, 코미디, 드라마, 가족, 공포, 로맨스, 스릴러의 10 가지로 구분되어 있다. 영화별 장르 정보는 1612개의 영화에 대한 장르 정보를 담고 있는 Data set이며, 영화를 평가한 고객정보는 영화를 평가한 30861 명의 사용자 정보이다. 사용자의 영화에 대한 평가는 사용자가 영화에 대해서 0.0에서부터 1.0까지 0.2간격으로 6단계로 평가한 데이타이다.

4.2 실험 방법 및 결과

본 논문에서 제안한 사용자 선호도 예측 방법은 Microsoft Visual C++ 6.0으로 구현되었으며, 실제 실험 환경은 PentiumⅢ 450Mhz, 256MB Ram 환경에서 수행되었다. 실험 방법은 3장에서 제안된 방식을 3가지로 구분하였다. 첫 번째 방법(P_Corr_A)은 기존의 협력적 필터링 기술(P_Corr)에 연관 사용자 군집만을 적용한 방법이고, 두 번째 방법(P_Corr_N_Bayes)은 아이템을 분류하여 분류된 카테고리에 따라 아이템에 대한 사용자의 선호도에 가중치를 달리 부여하여 아이템의 정보를 반영한다. 마지막 방법(P_Corr_A_N_Bayes)은 연관 사용자 군집 안에서 두 번째 방법을 적용한다.

본 논문에서는 EachMovie 데이타를 전처리하여 30861명의 사용자와 1612종류의 영화에 대해서 실험을 진행하였다. 이는 Naive Bayes 학습을 위한 훈련집합이다. 사용자 트랜잭션에서는 1601개의 연관 규칙과 신뢰도를 생성하였고, 연관 규칙의 평균 길이는 3이다. 500명의 사용자들에 대해 ARHP 알고리즘 적용결과, 최소 지지도 30%를 만족하는 5개의 사용자 클러스터를 아래 표 1과 같이 생성되었다.

표 1 클러스터 포함된 사용자 리스트

클러스터번호	클러스터에 포함된 사용자	사용자수
1	[5][6][10][13][18][19][25][26][27][37] [38][46][47][51][52][60][90][101][107]	114
2	[1][2][3][4][11][12][17][20][28][35][36] [45][53][63][68][69][76][102][103][104]	89
3	[9][10][16][24][34][42][43][54][55][64] [65][66][77][82][85][92][100][105][106]	104
4	[7][8][15][21][23][29][32][29][50][56] [57][70][73][78][79][82][94][98][109]	105
5	[14][22][30][31][40][41][47][49][58][59] [71][72][80][81][89][95][96][97][108]	88

훈련집합은 ARHP 알고리즘을 의해서 연관된 사용자 군집과 사용자의 대표장르를 기반으로 선호도를 표시한 아이템들을 장르별로 만든다. 10개의 장르별로 아이템을 분류한 것은 아래 표 2와 같다.

훈련집합의 아이템들은 추정치를 부여하기 위해서 Naive Bayes 알고리즘에 의해서 학습한다. 추정치가 부여된 학습집단에서 분류된 장르에 따라 아이템에 대한 사용자의 선호도의 가중치를 달리 부여하여 결측치 값(Missing Value)에 아이템의 정보를 반영한 것은 표 3와 같다. 새로운 사용자는 Naive Bayes 분류자에 의해서 장르가 분류되면, 분류된 장르 내에 속한 사용자들과 새로운 사용자의 유사도를 구하기 위해 추정치가 부여된 학습집단을 통해 사용자가 평가한 아이템에 추정치를 달리 부여한다. 표 2에서 대표장르가 결정된 사용자들의 많은 수가 Action 장르와 Drama 장르로 결정되었다. 이는 대부분의 사용자들이 이 두 장르를 선호하기 때문이다.

4.3 분석 및 평가

예측 알고리즘을 평가하는 여러 가지 방법 중에서 예측값과 실제 값의 차이를 표시하여 정확성 측면에서 성능을 평가하기 위해 MAE(Mean absolute error)방식과 예측

표 3 추정치가 부여된 선호도와 예측된 선호도 값

Item	User ID in Cluster 1							
	5	6	10	127	13	18	Newuser	...
1	0.0025	0.0038	0.0098	0.0035	0.0029	0.0097	0.0084	...
19	0.0012	0.0015	0.0024	0.0018	0.0027	**0.0075**	**0.0027**	...
21	0.0020	0.0037	0.0034	**0.0023**	0.0020	0.0037	**0.0037**	...
35	0.0008	0.0013	0.0006	0.0019	**0.0075**	0.0026	0.0008	...
45	0.0012	0.0020	0.0017	0.0026	0.0007	**0.0047**	0.0007	...

■ : 결측치 값에 추정치를 부여된 선호도를 예측한 값

할 수 있는 아이템의 전체 대비 비율인 Coverage 방식을 사용하여 성능평가 하였다[3,14,15]. 표 4와 그림 8는 본 논문에서 제안하는 방식을 아이템에 대한 사용자의 선호도를 기반으로 이웃 선정 방법을 사용하고, 피어슨 상관계수에 의해 사용자의 유사도를 구하는 기존의 협력적 필터링 방식(P_Corr)[3]과 비교하여 실험한 예측 값들의 MAE / Coverage를 평균한 값들이다. 본 실험에서는 평가를 백분율로 표시한다.

표 4와 그림 8을 보면 기존의 방식과 비교해 볼 때 정확도가 향상되었다. 방법 1과 방법 3은 처음에는 기존 방식과 정확도가 비슷하지만, 사용자가 많아질수록 정확도가 높아지는 것을 볼 수 있다. 사용자의 수가 적을 때에는 연관 사용자 군집의 의미가 크지 않으므로 이 결과는 예측과 부합한다고 할 수 있다. 방법 2는 기존 방식과 비교하면 사용자의 수에 관계없이 정확도가 높은 것으로 나타난다. 이는 아이템에 대한 정보를 반영하여 통계적인 값에 의해 가중치를 부여하기 때문에 예측의 정확도는 향상된다.

Coverage는 연관 사용자 군집만을 적용했을 경우 다른 방법보다 심각하게 감소하는 경향이 보인다. 그러나 군집을 이용하여 연산할 수 있는 데이타 량이 줄어들기 때문에 연산 시간이 줄어드는 것을 알 수 있다. 전체적인 EachMovie 데이타의 예측 결과를 보면 연관 사용자 군집과 베이지안 추정치를 이용한 방법이 기존의 피어슨 상관

표 2 훈련집합

GenreID		선호도를 표시한 아이템	사용자수	아이템수
1	Action	Golden eye, Clueless, 12Monkeys, Star gate, Star Wars, Drop Zone, Mission, ..	13590	1502
2	Animation	Toy Story, Exit to Eden, Heavy Metal, Pocahontas, Space Jam, Robin Hood, ..	125	163
3	Art/Foreign	Four Rooms, Birdcage, Antonia's Line, Birdcage, Stalker, Diva, Shine, The Delta, ..	385	961
4	Classic	Jumanji, Balto, Happy Gilmore, Foreign Student, Alien, Amadeus Annie Hall, ..	249	1541
5	Comedy	Ace Ventura, Bronx Tale, Fatal Instinct, Four Rooms, Palookaville, Heather, ..	4107	1545
6	Drama	Sabrina, Nixon, Ace Ventura, Clerks, Get Shorty, High Noon, Cape Fear, Power, ..	11559	1604
7	Family	Casper, Apollo13, Bad Boys, Batman Forever, Gordy, Fly Away Home, Shiloh, ..	158	1248
8	Horror	Copycat, Screamers, Mary Reilly, Babe, Clueless, M, Braindead, Scream, Alien3, ..	74	453
9	Romance	American President, Swiss Family Robinson, Beautiful Thing, Benny & Joon, ..	166	603
10	Thriller	Die Had, Taxi Driver, Crimson Tide, The Net, Breakdown, Head Above Water, ..	448	916

계수 만을 적용했을 경우보다 우수한 결과를 보이나 연관 사용자 군집만을 적용한 경우에는 Coverage가 줄어든 문제가 있으므로 이를 보완해야 할 것으로 보인다.

표 4 MAE / Coverage

(단위%)

성능평가	10명	100명	250명	500명
P_Corr	18.76 /81.22	18.43 /80.87	18.12 /80.25	17.91 /79.22
P_Corr_A	18.81 /81.33	17.01 /81.33	16.79 /80.01	16.45 /81.34
P_Corr_N _Bayes	17.99 /81.54	16.89 /81.59	16.32 /81.6	16.12 /82.13
P_Corr_A _N_Bayes	18.21 /82.23	17.59 /82.34	16.11 /82.82	15.41 /84.22

기존의 협력적 필터링 방식: P_Corr,
방법1: P_Corr_A, 방법2: P_Corr_N_Bayes,
방법3: P_Corr_A_N_Bayes

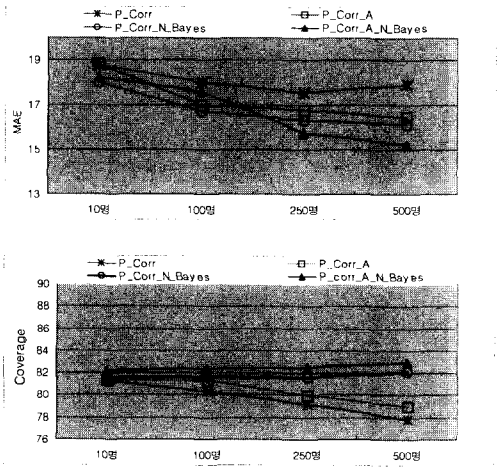

그림 8 MAE / Coverage 성능평가

6. 결 론

기존의 협력적 필터링 기술을 이용한 사용자 선호도 예측 방법에서는 아이템에 대한 사용자의 선호도를 기반으로 이웃 선정 방법을 사용하고, 피어슨 상관 계수에 의해 사용자의 유사도를 구한다. 그러므로 사용자가 평가하지 않은 아이템들은 사용자에게 추천되지 않는 초기 평가문제와 사용자-아이템의 데이타 데이타 집합의 희박한 특징을 보인다. 또한 아이템의 속성에 대한 사용자의 선호도를 직접적으로 반영하지 못하는 문제점이 있다.

본 논문에서는 기존의 사용자 선호도 예측 방법의 문제점을 보완하기 위하여 베이지안 추정치가 부여된 유사도 가중치와 연관 사용자 군집을 이용한 선호도 예측 방법을 사용하였고, 희박성 문제를 해결하기 위하여 Association Rule Hypergraph Partitioning 알고리즘을 사용하여 사용자를 장르별로 군집하였다. 새로운 사용자는 Naive Bayes 분류자에 의해 이들 장르 중 하나로 분류된다. 또한, 분류된 장르 내에 속한 사용자들과 새로운 사용자의 유사도를 구하기 위해 Naive Bayes 학습을 통해 사용자가 평가한 아이템에 추정치를 달리 부여한다. 추정치가 부여된 선호도를 기존의 피어슨 상관관계에 적용할 경우 결측치(Missing Value)로 인한 예측의 오류를 적게하여 예측의 정확도를 높일 수 있다. 제안된 방법의 성능을 평가하기 위해서 기존의 협력적 필터링 기술과 비교 평가하였다.

본 논문에서는 연관 사용자 군집을 적용하여 사용자들을 군집하고, 이 군집에 속한 사용자들의 선호도를 기반으로 사용자의 선호도에 가중치를 달리 부여하여 아이템에 대한 정보를 반영하여 통계적인 값에 의해 가중치를 부여하여 사용자의 선호도를 예측하였다. 제안한 방법에 대한 성능을 기존의 협력적 필터링 기술과 비교 실험한 결과 예측의 정확도가 향상되었기 때문에 본 논문에서 제안한 방식이 효과적임을 알 수 있었다.

향후 연구 과제로는 사용자 유사도 간의 상관 계수에 대한 연구와 영화의 장르뿐만 아니라 아이템의 속성들과 사용자가 실제로 영화를 보았는지의 여부를 알 수 있는 가중치 정보를 이용하여 사용자 유사도 가중치를 구하는데 이용한다면 좋은 결과를 기대할 수 있을 것이다.

참 고 문 헌

[1] D. Billsus and M. J. Pazzani, "Learning collaborative information filters," In proceedings of the International Conference on Machine Learning, 1998.

[2] M. O'Connor and J. Herlocker, "Clustering Item for Collaborative Filtering," In Proceedings of the ACM SIGIR Workshop on Recommender Systems, Berkeley, CA, 1999.

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

[4] 정경용, 김진현, 이정현, "연관 사용자 군집과 베이지안 분류를 이용한 사용자 선호도 예측 방법," 제28회 한국정보과학회 추계학술발표 논문집(II)-우수논문, pp. 109-111, 2001.

[5] C. Basu and H. Hirsh and W. W. Cohen, "Recommendation as classification: Using social and content-based information in recommendation." In

proceedings of the Fifteenth National Conference on Artificial Intelligence, pp. 714-720, Madison, WI, 1998.
[6] M. Balabanovic and Y. Shoham, "Fab : Content-based, collaborative recommendation," Communication of the Association of Computing Machinery, Vol. 40, No. 3, pp. 66-72, 1997.
[7] N. Good, J. B. Schafer and J. A. Konstan, A. Borchers, B. Sarwar, J. Herlocker, and J. Riedl, "Combining collaborative filtering with personal agents for better recommendations," In Proceedings of National Conference on Artificial Intelligence (AAAI-99), pp. 439-446, 1999.
[8] W. S. Lee, "Collaborative learning for recommender systems," In Proceedings of the Conference on Machine Learning, 1997.
[9] I. Soboroff and C. Nicholas, "Combining content and collaboration in text filtering," In Proceedings of the IJCAI'99 Workshop on Machine Learning in Information filtering, pp. 86-91, 1999.
[10] M. J. Pazzani, "A framework for collaborative, content-based and demographic filtering," Artificial Intelligence Review, pp. 393-408, 1999.
[11] E. H. Han, et al., "Clustering Based On Association Rule Hypergraphs," Proc. of SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery(DMKD), May, 1997.
[12] G. Karypis, V. Kumar, "Multilevel k-way Hypergraph Partitioning," DAC, pp. 343-348, 1999.
[13] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules," Proc. of the 20th VLDB Conference, pp. 487-499, 1994.
[14] J. S. Breese and D. Heckerman and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. of the 14th Conference on Uncertainty in Artificial Intelligence, 1998.
[15] J. Herlocker, J. Konstan, A. Borchers and J. Riedl, "An Algorithm Framework for Performing Collaborative Filtering," In Proceedings of ACM SIGIR'99, 1999.
[16] R. J. Kwok, "Automated text categorization using support vector machine," In Proceedings of the International Conference on Neural Information Processing, PP. 347-451, October, 1998.
[17] 정영미, 정보검색론, 구미무역 출판부, 1993.
[18] G. Salton and C. Buckley. "Term Weighting Approaches in Automatic Text Retrieval," Information Processing & Management, Vol. 24, No. 5, pp. 513-523, 1988.
[19] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Analysis of Recommendation Algorithms for E-Commerce." The ACM E-Commerce 2000 Conference, 2000.
[20] G. Salton and M. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.
[21] T. Michael, Maching Learning, McGraq-Hill, pp. 154-200, 1997.
[22] P. McJones, EachMovie collaborative filtering data-set, URL:http://www.research.digital.com/SRC/eachmovie, 1997.
[23] K. Y. Jung, J. H. Lee, "Prediction of User Preference in Recommendation System using Association User Clustering and Bayesian Estimated Value," Lecture Notes in Artificial Intelligence 2557, 15th Australian Joint Conference on Artificial Intelligence, December 2-6, 2002.

정 경 용
2000년 인하대학교 전자계산공학과(공학사). 2002년 인하대학교 전자계산공학과(공학석사). 2002년~현재 인하대학교 전자계산공학과 박사과정. 2001년~현재 에이플러스전자(주). 선임연구원. 2003년~현재 가천길대학 뉴미디어과 겸임교수. 관심분야는 웹 마이닝, 기계학습, 정보검색, CRM, 협력적 필터링, 자연어처리, 전자상거래

최 성 용
1993년 인하대학교 통계학과 졸업(이학사). 2001년 인하대학교 대학원 통계학과(이학석사). 2001년~현재 인하대학교 전자계산공학과 박사과정. 2001년~현재 김포대학 소프트웨어개발 겸임교수. 관심분야는 베이지안 학습, 신경망, 지능형 에이전트

임 기 욱
1944년 인하대학교 공과대학 전자공학과 졸업. 1987년 한양대학교 전자계산학 석사. 1994년 인하대학교 전자계산학 박사 1977년~1983년 한국전자기술연구소 선임연구원. 1983년~1988년 한국전자통신연구소 시스템소프트웨어 연구실장. 1988년~1989년 미 캘리포니아주립대학(Irvine)방문 연구원 1989년~1997년 한국전자통신연구원 시스템연구부장 주전산기(타이컴) Ⅲ,Ⅳ개발 사업책임자. 1997년~2000년 정보통신연구진흥원 정보기술 전문의원. 2000년~현재 선문대학교 교수. 관심분야는 실시간 데이타 베이스시스템, 운영체제, 컴퓨터구조

이 정 현

1977년 인하대학교 전자공학과 졸업 1980년 인하대학교 대학원 전자공학과 (공학석사). 1988년 인하대학교 대학원 전자공학과 (공학박사). 1979년~1981년 한국전자기술연구소 시스템 연구원. 1984년~1989년 경기대학교 전자계산학과 교수. 1989년~현재 인하대학교 컴퓨터공학부 교수. 관심분야는 자연어처리, HCI, 정보검색, 음성인식, 음성합성, 컴퓨터구조