선호도 재계산을 위한 연관 사용자 군집 분석과 Representative Attribute-Neighborhood를 이용한 협력적 필터링 시스템의 성능향상

정 경 용[†] · 김 진 수[†] · 김 태 용[††] · 이 정 현[†††]

요 약

추천 시스템에 있어서 협력적 필터링 기술은 많은 연구가 되고 있다. 그러나 협력적 필터링 기술을 이용한 추천 시스템은 초기 평가 문제와 희박성 문제가 발생한다. 이를 해결하기 위해서 본 논문에서는 선호도 재 계산을 위한 연관 사용자 군집과 베이지안 추정치를 이용한 사용자 선호도 예측 방법을 제안한다. 제안한 방법에서는 협력적 필터링 시스템에서 아이템의 속성을 고려하지 않는 단점을 보완하기 위해서 선호도에 가장 크게 영향을 미치는 대표 장르를 추출하여 유사한 이웃을 찾아 낼 때 예측에 이용하는 Representative Attribute-Neighborhood 방법을 사용한다. 협력적 필터링의 알고리즘에 군집 아이템 백터 내의 특정 아이템의 선호도를 재계산 하기 위한 연관 사용자 군집 분석을 적용하여 성능 향상을 하였다. 또 초기 평가 문제와 희박성 문제를 해결하기 위하여 Association Rule Hypergraph Partitioning 알고리즘을 사용하여 사용자를 장르별로 군집한다. 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 피어슨 상관 관계에 적용할 경우 결측치(Missing Value)로 인한 예측의 오류를 적게하여 예측의 정확도를 높일 수 있다. 제안된 방법은 기존의 방법보다 높은 성능을 나타냄을 보인다.

Performance Improvement of Collaborative Filtering System Using Associative User's Clustering Analysis for the Recalculation of Preference and Representative Attribute-Neighborhood

Kyung-Yong Jung[†] · Jin-Su Kim[†] · Tae-Yong Kim[††] · Jung-Hyun Lee[†††]

ABSTRACT

There has been much research focused on collaborative filtering technique in Recommender System. Howerer, these studies have shown the First-Rater problem and the Sparsity problem. The main purpose of this paper is to solve these problems. In this paper, we suggest the user's predicting preference method using Bayesian estimated value and the associative user clustering for the recalculation of preference. In addition to this method, to complement a shortcoming, which doesn't regard the attribution of item, we use Representative Attribute-Neighborhood method that is used for the prediction when we find the similar neighborhood through extracting the representative attribution, which most affect the preference. We improved the efficiency by using the associative user's clustering analysis in order to calculate the preference of specific item within the cluster item vector to the collaborative filtering algorithm. Besides, for the problem of the Sparsity and First-Rater, through using Association Rule Hypergraph Partitioning algorithm associative users are clustered according to the genre. New users are classified into one of these genres by Naive Bayes classifier. In addition, in order to get the similarity value between users belonged to the classified genre and new users, and this paper allows the different estimated value to item which user evaluated through Naive Bayes learning. As applying the preference granted the estimated value to Pearson correlation coefficient, it can make the higher accuracy because the errors that cause the missing value come less. We evaluate our method on a large collaborative filtering database of user rating and it significantly outperforms previous proposed method.

키워드 : 협력적 필터링 시스템(Collaborative Filtering system), 연관 규칙(Association Rule), 네이브 베이지안(Naïve Bayes)

1. 서 론

대부분의 추천 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. 또한 아이템의 속성에 대한 사용자의 선호도를 직접적으로 반영하지 못하는 문제점도 있다 [3, 8, 18].

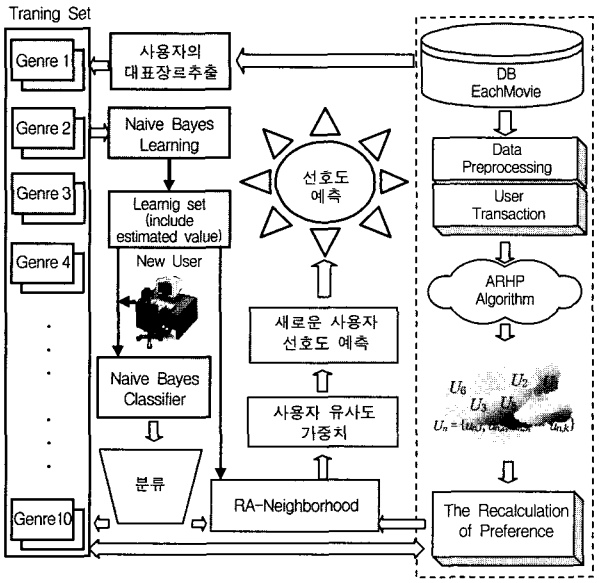
† 준 회 원 : 인하대학교 대학원 전자계산공학과
†† 정 회 원 : 문경대학교 웹마스터과 교수
††† 종신회원 : 인하대학교 컴퓨터공학부 교수
논문접수 : 2002년 2월 8일, 심사완료 : 2003년 4월 11일

협력적 필터링 기술에서 고려하기 힘든 부분에 대해서 내용 기반 필터링을 이용함으로써 문제점을 해결한다. 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 최근에 이루어지고 있다[3, 9, 14].

본 논문의 구성은 다음과 같다. 2장에서는 제안한 사용자 선호도 예측을 위한 시스템 구성도와 순서도를 소개하고, 3장에서는 협력적 필터링 알고리즘의 성능 향상을 위한 Representative Attribute-Neighborhood와 선호도 재 계산을 위한 연관 사용자 군집 분석 적용 방법을 제안한다. 4장에서는 베이지안 추정치를 이용한 예측 방법을 제안한다. 5장에서는 추천 시스템의 성능 평가를 위한 실험한다. 끝으로 6장에서는 결론을 맺는다.

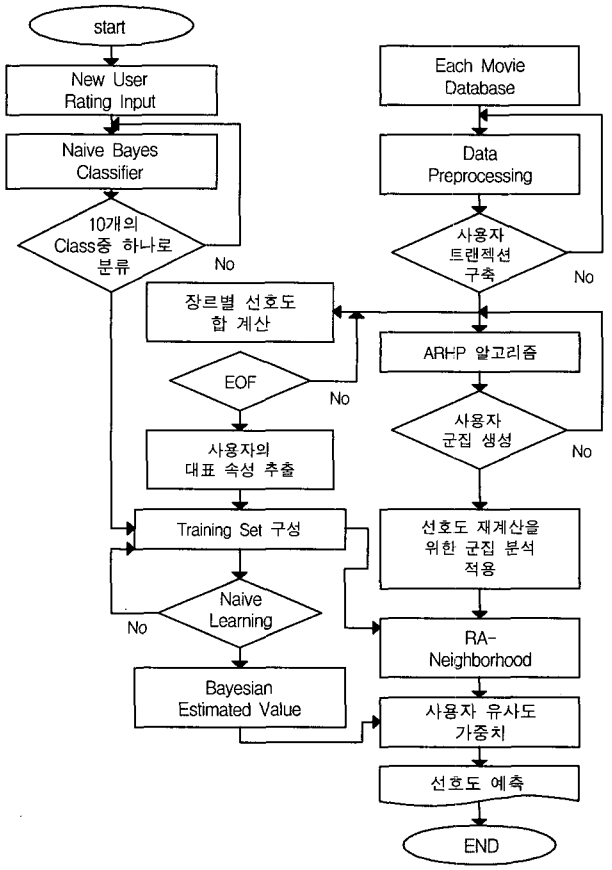
2. 시스템 구성도

(그림 1)은 본 논문에서 설계한 Representative Attribute-Neighborhood 방법과 베이지안 추정치를 이용하여 선호도 예측 시스템에 대한 구성도로서 세부 단계는 다음과 같다.



(그림 1) 사용자 선호도 예측을 위한 시스템 구성도

예측 시스템의 성능 향상을 위해서 선호도 재 계산을 위한 군집 분석을 적용하였다. 본 논문에서는 협력적 필터링 시스템에서 아이템의 속성을 고려하지 않고 사용자의 선호도만을 기반으로 하는 이웃 선정 방법의 문제점[3, 16, 17]을 보완하기 위해서 선호도에 가장 크게 영향을 미치는 사용자 대표 속성을 추출하여 훈련 집합을 만들 때 사용하고, 유사한 이웃을 찾아 낼 때 이를 예측에 이용하는 Representative Attribute-Neighborhood 방법을 사용한다. (그림 2)는 선호도 재계산을 위한 연관 사용자 군집 분석과 Representative Attribute(RA)-Neighborhood를 이용한 사용자 선호도 예측 시스템을 위한 순서도이다. 본 논문에서 사용하는 전체 알고리즘의 구성을 하나의 순서도로 나타내었다.



(그림 2) 사용자 선호도 예측 시스템의 순서도

군집 아이템 벡터내의 특정 아이템의 선호도를 재 계산하기 위해서 군집 내에서 특정 아이템에 선호도를 보인 사용자들의 선호도와 사용자가 군집에 속할 확률을 곱하여 합계한 후 특정 아이템에 선호도를 평가한 사용자들이 군집에 속할 확률의 합으로 값을 나누어 새로운 아이템의 선호도를 계산한다. 이렇게 구해진 군집의 대표 아이템 벡터를 새로운 복합 사용자로 간주하여 특정 사용자와 복합 사용자간의 이웃을 구하는 방식으로 협력적 필터링에 적용한다. 또 협력적 필터링 시스템에서의 초기 평가 문제(First-rater problem)와 희박성 문제(Sparsity problem)를 해결하기 위해서 Association Rule Hypergraph Partitioning 알고리즘[10, 15]을 사용하여 연관 사용자를 장르별로 군집하며, 새로운 사용자는 Naive Bayes 분류자에 의해 이들 장르 중 하나로 분류된다. 분류된 장르 내에 속한 사용자들과 새로운 사용자의 유사도를 구하기 위해 Naive Bayes 학습[20, 22]을 통해 사용자가 평가한 아아템에 추정치를 달리 부여한다. 추정치가 부여된 선호도를 피어슨 상관 관계에 적용할 경우 결측치로 인한 예측의 오류를 적게하여 예측의 정

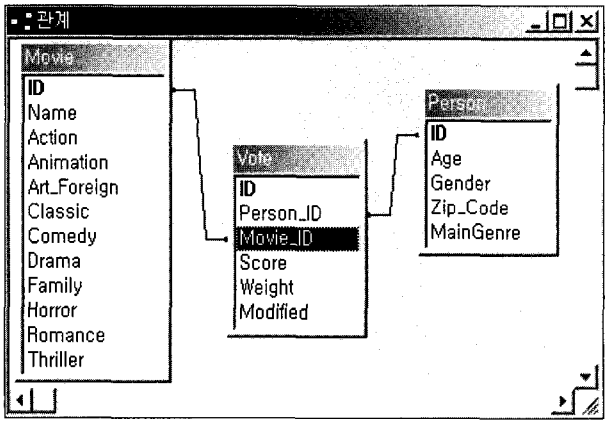
확도를 높일 수 있다[23, 24]. 제안된 방법의 성능을 평가하기 위해서 기존의 방법과 비교 평가 하였다[11-14].

3. Representative Attribute-Neighborhood와 선호도 재 계산을 위한 연관 사용자 군집 분석 적용

3.1 데이터 정제 작업

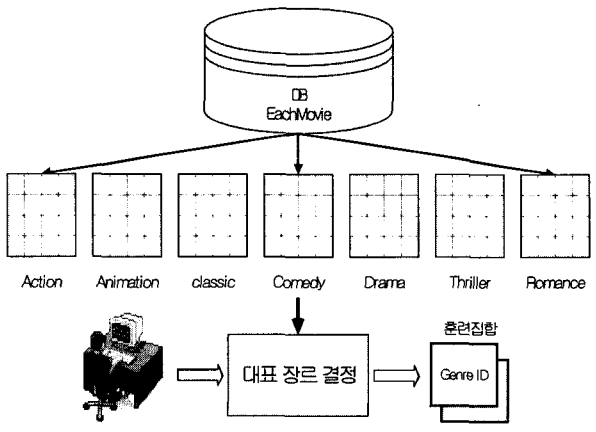
사용자의 대표장르 추출 및 연관 사용자 군집하기 위해서 EachMovie 데이터[19]를 사용한다. 이 EachMovie 데이터는 사용자가 선호도를 평가한 아이템들이 영화별 장르 정보에 속하지 않는 것이 있고, 사용자의 정보 또한 누락이 된 것이 있다. 이를 해결하기 위해서 본 연구에서는 데이터 정제 작업을 하여 데이터의 무결성 검사를 하였다[7].

(그림 3)과 같이 Vote 테이블을 기준으로 Movie와 Person 테이블의 무결성 검사를 하였고, Person과 Movie 테이블을 기준으로 Vote 테이블의 무결성 검사를 하였다. 그리고 실험을 위하여 Movie, Vote 테이블의 ID를 Vote 테이블의 Movie_ID에 관계 설정하였다.



(그림 3) Person, Vote, Movie 테이블의 관계

3.2 사용자의 대표속성 추출



(그림 4) 사용자의 대표장르 결정 개념도

사용자가 선호도를 평가한 아이템을 이용하여 사용자의 대표속성을 추출한다. 대표속성을 추출하기 위해서는 사용자의 장르별 아이템의 선호도 합을 계산한 후 선호도의 합이 가장 큰 장르를 대표속성으로 정한다[3, 23]. (알고리즘 1)은 사용자의 대표속성을 추출하는 알고리즘이다. 사용자의 대표속성 추출은 훈련집합(train set)을 구성할 때 사용하고, Representative Attribute-Neighborhood[12, 14]를 만들 때 사용한다. (그림 4)은 사용자의 대표속성을 추출하기 위한 개념도이다.

아이템의 속성을 고려하지 않는 협력적 필터링의 단점을 보완하여 좀 더 효율적인 필터링을 수행하기 위해 사용자의 대표속성을 추출하는 것이다. 본 논문에서 사용자의 대표장르는 선호도에 가장 크게 영향을 미치는 대표속성이라 가정한다. 기존의 협력적 필터링 기술은 사용자의 각 정보에 대한 선호도의 정도를 반영하여 예측을 수행하기 위해 전체 정보에 대하여 유사도를 계산하여 예측에 반영하게 된다. 그러나 전체 정보를 모두 사용하여 유사도를 구하는 것은 대표 속성에 대해서 사용자가 차별적인 선호도를 가지는 경우 이를 제대로 반영하지 못하는 단점이 있다. 그러므로 본 논문에서는 이를 보완하기 위해서 각 대표 속성에 한정하여 훈련집합을 구성할 때 사용하고, 유사한 이웃을 찾아 낼 때 이를 예측에 이용하는 방법에 사용한다.

```
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 ;
   } // 아이템에 대해서 장르별 선호도의 합을 구한다.
 If (GenreMaxSum > MainGenreMaxSum) {
        MainGenreID ← GenreID of j th ;
        MainGenreMaxSum ← GenreMaxSum ;
  } // 선호도의 합이 가장 큰 장르의 ID을 Return
}
Assign (MainGenreID) ;  // 대표 장르에 의한 대표 속성 결정
Representative Attribute-Neighbor[MainGenreID] ← Add UserID ;
// Representative Attribute-Neighborhood
```

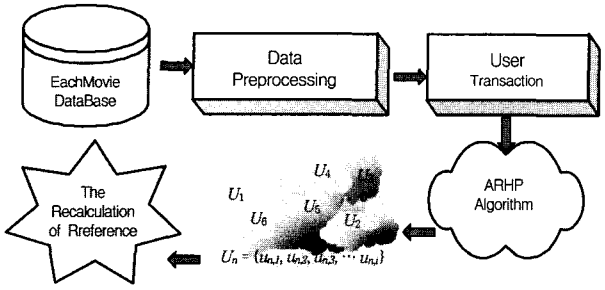
(알고리즘 1) 사용자의 대표속성을 추출하는 알고리즘

본 논문에서의 사용자 대표 속성 추출은 EachMovie 데이터[19]의 영화를 평가한 고객정보에서 아이템에 대한 선호도의 합만을 고려하고, 실제로 영화를 보았는지의 여부를 알 수 있는 가중치 정보는 고려하지 않는다. 사용자의 대표속성을 추출하는 알고리즘에서 선호도의 합을 계산한 후 선호도의 합이 가장 큰 장르를 대표 장르라고 한것은 선호도에 가장 크게 영향을 미치는 대표 속성을 의미하기 때문이다[12-14].

3.3 Association Rule Hypergraph Partitioning 알고리즘에 의한 연관 사용자 군집

Association Rule Hypergraph Partitioning(ARHP) 알고리즘은 연관 규칙과 Hypergraph Partitioning을 이용하여 트랜잭션 기반의 데이터베이스에서 연관된 항목들을 클러스터링 하는 방법이다[10, 15]. 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 어려운 다차원 데이터 집합에 대한 클러스터링에 유용하다. 연관 규칙의 신뢰도를 Hypergraph Partitioning의 가중치로 사용한다.

ARHP 알고리즘에 의한 연관 사용자 군집을 위한 단계적 흐름은 (그림 5)와 같이 진행된다. 사용자에 의해 선호도가 표시된 아이템들을 사용자 트랜잭션으로 재구성한다[7, 25]. 이를 연관 규칙 탐사 방법[1, 2]을 이용하여 사용자 트랜잭션 안에 빈번하게 동시에 출현하는 사용자들의 집합을 찾는다. 사용자들에 대한 Large 항목집합을 가지고 Apriori 알고리즘[1, 2]을 이용하여 연관 규칙과 신뢰도를 계산한 후, 연관규칙(association rule)에 포함되는 항목을 vertex로, 연관 관계를 Hyperedge로 매핑처리 한다. 그리고 신뢰도를 Hypergraph Partitioning을 위한 가중치로 사용자들간의 군집을 형성한다[10].



(그림 5) 선호도 재 계산을 위한 연관 사용자 군집 분석 적용

3.4 선호도 재 계산을 위한 연관 사용자 군집 분석 적용

협력적 필터링에 군집 분석을 적용하는 방법은 기존의 데이터 베이스에 있는 사용자들을 평가한 아이템의 상위 분류에 의해서 군집화하고, 군집의 대표 아이템 벡터를 구해서 군집의 대표 아이템 벡터를 새로운 복합 사용자(composite user)로 간주하는 것이다[22]. 즉, m명의 전체 사용자를 n개의 군집으로 군집화하여 n명의 복합 사용자로 축소하는 것이다. n명의 복합 사용자의 아이템 벡터는 군집에 속한 사용자들 중 한 사람이라도 선호도를 평가한 아이템을 대상으로 구성한다. 선호도 재계산을 위한 군집 분석을 적용하기 위한 군집 아이템 벡터는 식 (1)과 같이 정의한다.

$$U_n = \{u_{n,1}, u_{n,2}, u_{n,3}, \cdots, u_{n,k}\} \quad (1)$$

U_n은 n번째 군집이고, $\{u_{n,1}, u_{n,2}, u_{n,3}, \cdots, u_{n,k}\}$는 n번째 군집의 k개의 아이템 벡터이다. k는 n번째 군집의 사용자들이 한번이라도 선호도를 평가한 아이템의 수이다. 군집 아이템 벡터 내에서 특정 아이템에 대한 선호도는 식 (2)와 같이 정의한다.

$$u_{n,k} = \frac{\sum_{n=1}^{n}(v_{i,k} \times p_{i,U_n})}{\sum_{i=1}^{n} p_{i,U_n}} \quad (2)$$

$u_{n,k}$는 n번째 군집에서 k번째 아이템에 대해서 평가한 값이고, $V_{i,k}$는 사용자 i가 아이템 k에 대해서 평가한 선호도이다. P_{i,U_n}는 사용자 i가 n번째 군집에 속할 확률이다. 식 (2)의 분모는 군집 내에서 $V_{i,k}$가 존재하는 사용자 i에 대해서만 계산한다. 군집 내에서 한번이라도 아이템에 선호도를 평가한 사용자들을 대상으로 합을 구한다는 의미이다. 군집 아이템 벡터내의 특정 아이템의 선호도를 재 계산하기 위해서 군집 내에서 특정 아이템에 평가한 사용자들의 선호도와 사용자가 군집에 속할 확률을 곱하여 합계한 후 특정 아이템에 선호도를 보인 사용자들이 군집에 속할 확률의 합으로 값을 나누어 새로운 아이템의 선호도를 구한다. 이렇게 구해진 군집의 대표 아이템 벡터를 새로운 복합 사용자로 간주하여 특정 사용자와 복합 사용자간의 이웃을 구하는 방식으로 협력적 필터링 기술에 적용한다. 또한 군집 분석을 통해서 사용자의 수가 복합 사용자들로 대치되면서 사용자의 아이템에 대한 상세 벡터 정보가 손실되어 발생하는 정확도 저하 부분을 Association Rule Hypergraph Partitioning 알고리즘에의한 연관 사용자 군집을 통해서 보완한다.

4. 베이지안 추정치를 이용한 예측 방법

4.1 Naive Bayes 알고리즘을 적용한 사용자 유사도 가중치

사용자들이 추정치가 부여된 학습집단을 구축하기 위해서는 우선 훈련 집합을 만들어야 한다. 훈련집합은 (알고리즘 1)에 의해서 사용자의 대표 장르를 계산한 후 장르별 사용자 군집을 기반으로 아이템들을 장르별로 수집한 데이터이다. 훈련집합의 아이템에 추정치를 부여하기 위하여 Naive Bayes 학습 알고리즘[20]을 사용한다. 본 논문에서는 아이템의 발생여부만 사용하는 방법이 아닌 아이템의 출현빈도를 고려하는 다항 베이지안 학습법을 사용한다[6, 20].

사용자 유사도 가중치는 사용자가 평가한 아이템의 선호도를 대상으로 장르별로 추정치를 다르게 적용한다. 이를 위해서 사용자가 선호도를 평가한 아이템에 추정치가 부여된 학습집단을 적용하면 장르별 아이템인 $P(nut_i|GenreID)$

에 $V_{a,k}$를 곱한다. $V_{a,k}$는 사용자 a와 아이템 k에 대해서 평가한 선호도이다.

$$\beta_{a,k} = P(nut_i | GenreID) \times V_{a,k} \quad (3)$$

식 (3)을 기반으로 피어슨 상관 계수[4, 5, 11]에 적용하면, 사용자 a와 사용자 i의 유사도 가중치는 식 (4)와 같이 재정의된다.

$$B(a,i) = \frac{Cov(a,i)}{\delta_a \cdot \delta_i} = \frac{\sum_k (\beta_{a,k} - \overline{\beta_a})(\beta_{i,k} - \overline{\beta_i})}{\sqrt{\sum_k (\beta_{a,k} - \overline{\beta_a})^2 (\beta_{i,k} - \overline{\beta_i})}} \quad (4)$$

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

4.2 Naive Bayes 분류자에 의한 사용자 분류

새로운 사용자를 장르별로 분류하기 위해 Naive Bayes 분류자[20]를 사용한다. Naive Bayes 분류자는 추정치가 부여된 학습집단을 사용하여 식 (5)에 의해 사용자를 장르별로 분류할 수 있다.

$$GenreID = \underset{Genre1 \in GenreTot}{\arg\max}\ p(GenreID) \prod_{i \in I} v_{a,k}\, p(nut_i | GenreID) \quad (5)$$

각 아이템에 대한 선호도 값을 가지는 새로운 사용자는 $u_{new} = \{x \in p \mid nut_1(x), nut_2(x), \cdots, nut_n(x)\}$로 표현하며, $nut_n(x)$는 새로운 사용자가 선호도(p)를 표시한 아이템들이다. nut_{new}가 분류될 장르는 $GenreID$로, 전체 장르는 $GenreTot$로 표현한다.

$P(nut_i | GenreID)$는 사용자가 선호도를 평가한 아이템들이 장르에 포함될 확률의 곱을 나타낸다. 새로운 사용자의 장르 결정은 확률 값이 가장 높은 장르(Genre class)에 할당한다.

사용자가 평가한 아이템의 선호도는 장르별로 추정치가 다르게 적용된다. 이를 위해 사용자가 선호도를 평가한 아이템에 추정치가 부여된 학습 집단을 적용하고, 장르별 아이템에 사용자의 선호도 값을 곱한다. 이렇게 함으로써 결측치(Missing Value)로 인한 예측의 오류를 줄일 수 있다. 또한 아이템을 분류하여 분류된 장르에 따라 아이템에 대한 사용자의 선호도에 가중치를 달리 부여한다. 이는 사용자의 선호도만을 이용하는 것이 아닌 통계적인 값에 의해 가중치를 부여한다.

4.3 새로운 사용자의 선호도 예측

새로운 사용자의 선호도 예측은 Naive Bayes 추정치를 적용한 사용자 a와 사용자 i의 유사도 가중치를 기존의 협력적 필터링 기술의 피어슨 상관계수에 적용한다. 이는 사용자의 선호도만을 이용하는 것이 아닌 통계적인 값에 의해 가중치를 부여하기 때문에 예측의 정확도가 향상된다. 특정 아이템에 대한 Representative Attribute-Neighborhood에 의해 선정된 이웃들의 선호도와 각 이웃들의 선호도 평균과의 거리를 이웃들과의 유사도로 가중 평균함으로써 특정 사용자의 아이템에 대한 선호도는 예측된다. 이를 수식으로 표현하면 식 (6)과 같이 정의한다.

$$p_{a,k} = \overline{v_a} + \frac{\sum_{i=0}^{n} \beta(a,i)(v_{i,k} - \overline{v_i})}{\sum_{i=0}^{n} \beta(a,i)} \quad (6)$$

P_{ak}는 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 필터링 시스템에서 사용자의 선호도만을 사용하여 유사도 가중치를 계산하나, $\beta(a,i)$는 Naive Bayes 추정치를 이용한 사용자 유사도 가중치에 의해서 계산된다.

5. 성능 평가

5.1 실험 방법 및 결과

본 논문에서 제안한 사용자 선호도 예측 방법은 Visual Studio C++ 6.0과 MS SQL Server 2000으로 구현되었으며, 실제 실험 환경은 PentiumIII 450Mhz, 256MB Ram 환경에서 수행되었다. 실험 데이터로는 컴팩 연구소에서 18개월 동안 협력적 필터링 알고리즘을 연구하기 위해서 영화에 대한 사용자의 선호도를 조사한 EachMovie 데이터[19]를 사용한다. EachMovie 데이터를 무결성 검사를 하여 30861명의 사용자와 1612종류의 영화에 대해서 실험을 진행하였다. 이는 Naive Bayes 학습을 위한 훈련 집합이다.

무결성 처리하여 전처리한 EachMovie 데이터 중 500명을 Systematic 샘플링하였다. 샘플링된 사용자 중 125명을 Random 샘플링하고 125명 사용자 각각의 선호도 자료에서 50%씩을 샘플링하여 Test 집합으로 구성하였다. 나머지 선호도 자료 50%와 375명의 전체 선호도 자료를 합하여 Train 집합으로 구성하였다. Train 집합의 영화에 대한 총 선호도 개수는 25396개이고 예측해야 하는 Test 집합의 영화에 대한 총 선호도 개수는 2937개이다. 선호도 재 계산을 위한 군집 분석을 하기 위해서 군집의 개수는 500명의 전체 사용자에 대해서 10개의 군집(10, 20, 30, 40, 50, 60, 70, 80, 90, 100)을 만들어 실험을 진행 하였다.

사용자 트랜잭션에서는 2601개의 연관 규칙과 신뢰도를 생성하였고, 연관 규칙의 평균 길이는 3이다. 500명의 사용

자들에 대해 Association Rule Hypergraph Partitioning 알고리즘 적용한 결과, 최소 지지도 30%를 만족하는 사용자 클러스터를 생성하였다[25]. 사용자 트랜젝션에서의 연관 규칙은 아래 <표 1>과 같이 생성하였다. <표 1>에서 "AND"는 두 사용자가 연관이 되어 있다는 의미이다.

〈표 1〉 사용자 트랜잭션에서 연관 규칙

Group	Support	Conf	Pvalue	Lift	UserID ==> UserID
1	58.333	77.78	15.28	1.24	[34] => [305]
1	58.333	77.78	15.28	1.24	[162] => [305]
1	45.833	61.11	15.28	1.33	[34] => [489]
1	62.500	83.33	16.67	1.25	[162] => [157]
1	58.333	93.33	18.33	1.24	[305] => [162]
1	62.500	93.75	18.75	1.25	[157] => [162]
1	54.166	86.67	24.17	1.39	[29] AND [129] => [305]
1	58.333	92.31	33.33	1.50	[34] AND [305] => [157]
1	58.515	82.54	33.54	1.45	[35] AND [205] => [117]
1	58.511	82.98	33.74	1.57	[33] AND [245] => [127]
…	…	…	…	…	…

대표 장르가 결정된 사용자들은 훈련집합을 만드는데 사용되고, 유사한 이웃을 찾아 내어 예측을 하는 Representative Attribute-Neighborhood 방법에 사용된다. <표 2>에서 대표 장르가 결정된 사용자들의 많은 수가 Action 장르와 Drama 장르로 결정 되었다. 이것은 대부분의 사용자들이 이 두 장르를 선호하기 때문이다. 대표 속성 추출을 하는 실험은 EachMovie 데이터 안에 있는 사용자가 평가한 아이템을 대상으로 실험을 진행하였기 때문에 실제적인 한국인의 선호도 및 정서와는 차이가 난다[14].

〈표 2〉 대표 속성이 추출된 사용자들

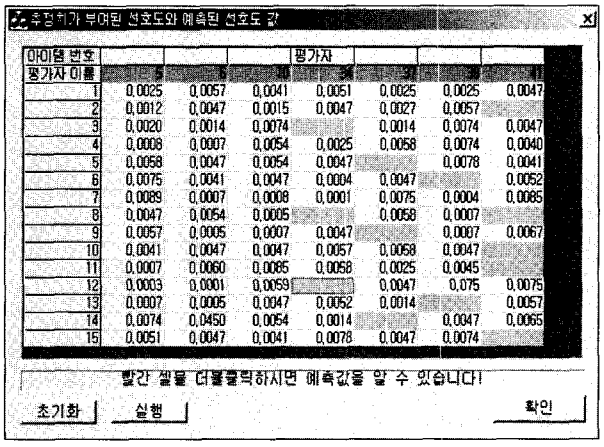
MainGenre	Representative Attribute UserID	총사용자
Action	User 9, User 10, User 11, User 25, User 26, User 48, User 49, User 86, User 87, User 88, …	13590
Animation	User 7, User 8, User 12, User 24, User 27, User 50, User 51, User 52, User 84, User 85, …	125
Art/Foreign	User 19 User 6, User 21, User 22, User 23, User 53, User 54, User 55, User 56, User 83, …	385
Classic	User 1, User 4, User 20, User 35, User 36, User 57, User 58, User 59, User 80, User 82, …	249
Comedy	User 2, User 3, User 33, User 34, User 48, User 76, User 77, User 78, User 79, User 81, …	4107
Drama	User 10, User 16, User 23, User 39, User 47, User 60, User 61, User 62, User 74, User 75, …	11559
Family	User 13, User 14, User 17, User 40, User 46, User 63, User 64, User 73, User 96, User 97, …	158
Horror	User 5, User 28, User 31, User 41, User 42, User 65, User 67, User 72, User 94, User 95, …	74
Romance	User 19, User 29, User 32, User 43, User 44, User 66, User 68, User 71, User 89, User 90, …	166
Thriller	User 15, User 30, User 37, User 38, User 45, User 69, User 70, User 91, User 92, User 93, …	448

훈련집합은 ARHP 알고리즘을 의해서 연관된 사용자 군집과 사용자의 대표장르를 기반으로 선호도를 표시한 아이템들을 장르별로 만든 후 Naïve Bayes 학습 알고리즘에 의한 추정치가 부여된 학습 집단을 구성 할 때 사용한다. 10개의 장르별로 아이템을 분류한 것은 아래 <표 3>과 같다.

〈표 3〉 훈련집합(Training Set)

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, …	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, …	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, …	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

훈련집합의 아이템들은 추정치를 부여하기 위해서 Naive Bayes 알고리즘에 의해서 학습한다. 추정치가 부여된 학습 집단에서 분류된 장르에 따라 아이템에 대한 사용자의 선호도의 가중치를 달리 부여하여 결측치 값(Missing Value)에 아이템의 정보를 반영한 것은 (그림 6)과 같다. 새로운 사용자는 Naïve Bayes 분류자에 의해서 장르가 분류되면, 분류된 장르 내에 속한 사용자들과 새로운 사용자의 유사도를 계산하기 위하여 추정치가 부여된 학습집단을 통해 사용자가 평가한 아이템에 추정치를 달리 부여한다.



(그림 6) 추 정치가 부여된 선호도와 예측된 선호도 값

(그림 6)의 결측값(Missing value)은 시스템에 의해서 추

정치가 부여된 선호도를 예측한 값이다. 그 결과는 <표 4>에서 시스템에 의해서 예측된 선호도 값을 나타내었다. 추정치가 부여된 선호도와 예측된 선호도 값을 계산하기 위한 프로그램은 http://dragondie.hihome.com/Resume/paperlink.htm에서 다운 받을 수 있다.

〈표 4〉 (그림 6)에서 예측된 선호도 값

UserID(item 번호)	
User34(3) = 0.0066	User34(8) = 0.0024
User34(12) = 0.0165	User37(5) = 0.0040
User37(9) = 0.0057	User37(14) = 0.0038
User39(6) = 0.0081	User39(13) = 0.0116
User41(2) = 0.0043	User41(8) = 0.0053
User41(10) = 0.0045	User41(11) = 0.0021
User41(15) = 0.0043	

5.3 분석 및 평가

5.3.1 성능 평가 기준

예측 알고리즘을 평가하는 여러 가지 방법 중에서 추천의 성능을 평가 하기 위해 본 논문에서는 Breese[5]에 의해 제안된 예측 값과 실제 값의 차이를 표시하여 정확성 측면에서 성능을 평가하기 위해 MAE(Mean absolute error)방식과 예측할 수 있는 아이템의 전체 대비 비율인 Coverage 방식을 사용하여 성능평가 하였다[5, 11, 13, 14, 23, 24].

MAE는 예측의 정확도를 측정하기 위해서 실제로 사용자가 평가한 값과 예측된 값의 차이에 대한 절대값의 평균을 나타낸다. MAE는 절대적으로 알고리즘이 얼마나 정확하게 예측을 했는지를 알 수 있으며 식 (7)에 의해 정의된다.

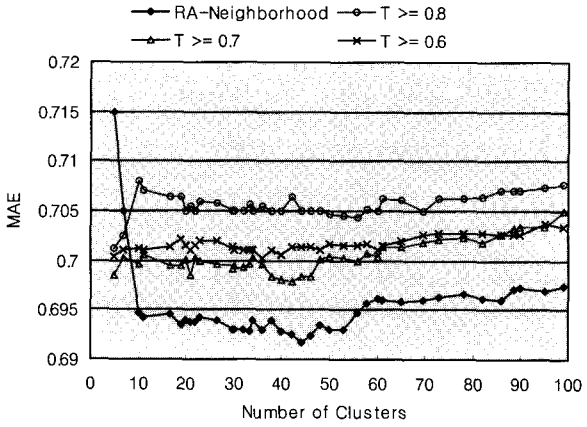
$$s_a = \frac{\sum_{j \in p_a} |p_{a,j} - v_{a,j}|}{m_a} \qquad (7)$$

식 (7)에서 $p_{a,j}$는 예측된 선호도이며 $v_{a,j}$는 실제로 사용자가 평가한 선호도이다. 또한 m_a는 새로운 사용자에 의해 평가된 아이템의 수를 의미한다.

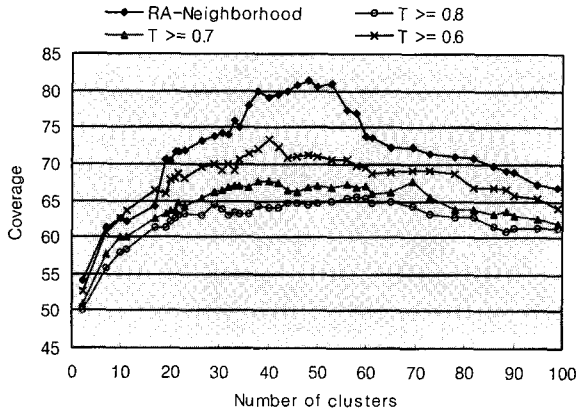
Coverage는 추천 알고리즘이 예측을 할 수 있는 아이템의 전체 대비 비율이다. Coverage는 예측의 대상이 되는 벡터 집합의 백터 개수를 전체로 보았을 때 특정한 알고리즘이 예측을 해 내는 개수의 전체 대비 비율이다. 예측에 사용하는 이웃의 개수를 적게 설정하게 되면 Coverage는 줄어들게 된다. 또한 신규 아이템과 같이 전체 사용자 중 어떤 사람에 의해서도 선호도가 평가되지 않은 아이템이 예측의 대상이 되면 이 아이템의 경우에도 사용자간의 유사도 가중치가 0이 될 가능성이 존재하므로 역시 아이템의 선호도를 예측할 수 없을 수 있다. Coverage는 전적으로 협력적 알고리즘의 정확도와 실행 성능과 반비례하는 경우가 많고 적절한 접점을 찾는 것이 필요하다.

5.3.2 제안한 방법의 성능평가

사용자 대표 장르 추출하여 유사한 이웃을 찾아내어 선호도를 예측하는 Representative Attribute-Neighborhood 방법과 기존의 협력적 필터링의 Thresholding(T)를 이용한 이웃 선정 방법[21]간에 실험을 통하여 예측의 정확도를 비교 수행하였다. 또 선호도 재 계산을 통한 연관 사용자 군집 분석을 적용하여 군집의 개수는 500명의 사용자에 대해서 10, 20, 30, 40, 50, 60, 70, 80, 90, 100개의 군집을 만들어 실험하였다.



(그림 7) MAE에 의한 RA-Neighborhood 성능 평가

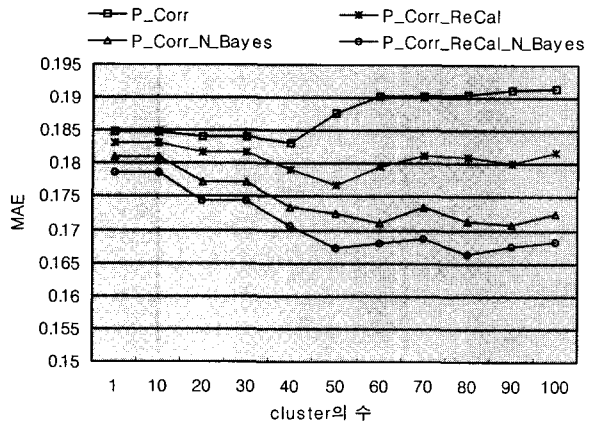


(그림 8) Coverage에 의한 RA-Neighborhood 성능 평가

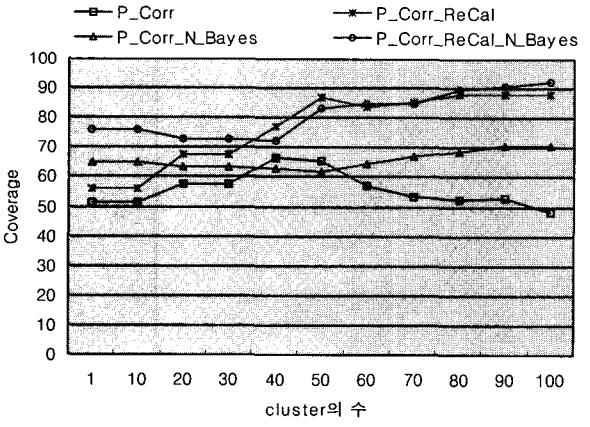
(그림 7)의 MAE를 보면, 아이템의 속성을 고려하지 않는 기존의 협력적 필터링의 단점을 보완하기 위해 사용자의 대표 장르 추출하여 유사한 이웃을 찾아 내어 선호도 예측에 이용하는 방법이 Thresholding(T)를 이용한 이웃 선정 방법[16, 18]보다 예측의 오차가 적음을 알 수 있다. (그림 8)의 Coverage를 보면, 선호도 재 계산을 위한 군집 분석을 적용한 결과 군집의 개수가 증가함에 따라 일관성 있게 결과가 좋아지지는 않으며 대략 50개의 군집, 즉 전체 대비 1/4 수준의 군집 수를 적용했을 때가 가장 좋은 결과를 보인다. 군집 분석을 적용하는 경우 500명의 사용자가

검색해야 하는 작업을 125명의 사용자가 검색하는 작업으로 줄일 수 있어 시스템의 속도를 높일 수 있다[24].

본 논문에서 제안한 베이지안 추정치를 적용한 사용자 유사도 가중치를 3가지 방법으로 실험을 진행하였다. 첫 번째 방법(P_Corr_ReCal)은 기존의 협력적 필터링 기술에 선호도 재 계산을 위한 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_ReCal_N_Bayes)은 선호도 재 계산을 위한 연관 사용자 군집 안에서 두 번째 방법을 적용한다.



(그림 9) 클러스터 수의 변화에 따른 MAE



(그림 10) 클러스터 수의 변화에 따른 Coverage

(그림 9)와 (그림 10)을 보면 기존의 협력적 필터링 방식과 MAE와 Coverage로 비교해 볼때 정확도가 향상되었다. 첫 번째 방법(P_Corr_ReCal)과 마지막 방법(P_Corr_ReCal_N_Bayes)은 처음에는 기존 방식과 정확도가 비슷하지만, 사용자의 군집의 수가 많아질수록 정확도가 높아지는 것을 볼 수 있다. 클러스터의 수가 적을 때에는 선호도 재계산을 위한 연관 사용자 군집의 의미가 크지 않으므로 이 결과는 예측과 부합한다고 할 수 있다. 두 번째 방법(P_Corr_N_Bayes)은 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 선호도 재 계산을 위한 연관 사용자 군집만을 적용했을 경우 다른 방법보다 심각하게 감소하는 경향이 보인다. 그러나 군집을 이용하여 연산할 수 있는 데이터량이 줄어들기 때문에 연산 시간이 줄어드는 것을 알 수 있다. 전체적인 EachMovie 데이터의 예측 결과를 보면 선호도 재 계산을 위한 연관 사용자 군집과 베이지안 추정치를 이용한 방법이 기존의 피어슨 상관계수 만을 적용했을 경우보다 우수한 결과를 보인다. 그리고 유사한 이웃을 찾아 낼 때 Representative Attribute-Neighborhood 방법을 선호도 예측에 이용하는 방법이 기존의 방법들보다 예측의 오차가 적음을 알 수 있다. 그러나 선호도 재 계산을 위한 연관 사용자 군집만을 적용한 경우에는 Coverage가 줄어든 문제가 있으므로 이를 보완해야 할 것으로 보인다.

6. 결 론

본 논문에서는 Representative Attribute(RA)-Neighborhood 방법과 베이지안 추정치을 이용하여 사용자의 선호도 예측 방법을 제안하였고, 선호도 재 계산을 위한 군집 분석을 적용하여 예측 알고리즘의 성능 향상을 하였다. Representative Attribute-Neighborhood 방법은 사용자의 대표속성을 추출하여 유사한 이웃을 찾아내어 예측을 하는데 이용하고, 베이지안 추정치를 이용하여 사용자의 선호도의 가중치를 달리 부여하여 결측치 값에 아이템의 정보를 반영한다. 선호도 재 계산을 위한 연관 사용자 군집 분석을 적용하여 추천 시스템의 속도를 높일 수 있다. 제안된 방법의 성능을 평가하기 위해 기존의 협력적 필터링 시스템과 비교한 결과 기존 방법보다 높은 성능을 보였다.

향후 제안된 방법에 사용자가 실제로 영화를 보았는지의 여부를 알 수 있는 가중치 정보를 이용하여 유사도 가중치를 구하는 데 이용한다면 좋은 결과를 기대할 수 있을 것이다.

참 고 문 헌

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994.

[2] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases," In Proceedings of the 1993 ACM SIGMOD Conference, Washington DC, USA, 1993.

[3] C. Basu, 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, Madison, WI, pp.714-720, 1998.

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

[5] J. S. Breese, D. Heckerman and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. of the 14th Conference on Uncertainty in Artificial Intelligence, 1998.

[6] Y. H. Chien and E. I. George, "A Bayesian Model for Collaborative Filtering," In Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics, San Francisco, 1999.

[7] R. Cooley, et al., "Data Preparation for Mining World Wide Web Browsing Patterns," Knowledge and Information Systems, Vol.1, No.1, 1999.

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

[9] N. Good, B. Schafer, J. Konstan, A. Borchers, B. Sarwar, J. Riedl, "Combining Collaborative filtering with Personal Agents for Better Recommendation," AAAI/IAAI, 1999.

[10] 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.

[11] J. Herlocker, J. Konstan, A. Borchers and J. Riedl, "An Algorithm Framework for Performing Collaborative Filtering," In Proceedings of ACM SIGIR '99, 1999.

[12] K. Y. Jung, J. K. Ryu and J. H. Lee, "A New Collaborative Filtering Method using Representative Attributes-Neighborhood and Bayesian Estimated Value," Proceedings of International Conference on Artificial Intelligence : Las Vegas, USA, pp.24-27, June, 2002.

[13] K. Y. Jung, Y. J. Park and J. H. Lee, "Integrating User Behavior Model and Collaborative Filtering Methods in Recommender Systems," International Conference on Computer and Information Science, Seoul, Korea, August, 2002.

[14] K. Y. Jung, J. H. Lee, "Prediction of User Preference in Recommendation System using Association User Clustering and Bayesian Estimated Value," Lecture Notes in Computer Science 2557, 15th Australian Joint Conference on Artificial Intelligence, December, 2002.

[15] G. Karypis, V. Kumar, "Multilevel k-way Hypergraph Partitioning," DAC, pp.343-348, 1999.

[16] G. Karypis, "Evaluation of Item-Based Top-N Recommendation Algorithms," Technical Report CS-TR-00-46, Computer Science Dept., University of Minnesota, 2000.

[17] S. J. Ko and J. H. Lee, "Feature Selection using Association Word Mining for Classification," In Proceedings of the Conference on DEXA2001, LNCS2113, pp.211-220, 2001.

[18] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon and J. Riedl, "GroupLens : Applying Collaborative Filtering to Usenet News," Communications of the ACM, Vol.40, No.3, pp.77-87, 1997.

[19] P. McJones, EachMovie collaborative filtering dataset, URL : http://www.research.digital.com/SRC/eachmovie, 1997.

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

[21] 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.

[22] 정영미, *정보검색론*, 구미무역 출판부, 1993.

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

[24] 김진현, 정경용, 김태용, 이정현, "연관 관계 군집에 의한 협력적 여과 방법", 제29회 한국정보과학회 추계학술발표 논문집(II), pp.331-333, 2001.

[25] 양신모, 정경용, 김진수, 최성용, 이정현, "아이템의 범주적 속성과 수량적 속성에 기반한 연관규칙 발견," 제12회 한국정보과학회, HCI° CG° DESIGN 학술대회, pp.456-461, 2003.

정 경 용

e-mail : kyjung@gcgc.ac.kr

2000년 인하대학교 전자계산공학과(공학사)

2002년 인하대학교 전자계산공학과(공학석사)

2002년~현재 인하대학교 전자계산공학과 박사과정

2001년~현재 에이플러스전자(주) 선임연구원

2003년~현재 가천길대학 뉴미디어과 겸임교수

관심분야 : 웹 마이닝, 기계학습, 정보검색, CRM, 협력적 필터링, 자연어처리, 전자상거래

김 진 수

e-mail : kjspace@nlsun.inha.ac.kr

1998년 인천대학교 전자계산공학과(공학사)

2001년 인하대학교 전자계산공학과(공학석사)

2001년~현재 인하대학교 전자계산공학과 박사과정

2002년~현재 김포대학 컴퓨터계열 겸임교수

관심분야 : 웹 마이닝, 데이터마이닝, 기계학습, 정보검색, 자연어처리, 웹 마이닝, 정보검색

김 태 용

e-mail : tykim@mkc.ac.kr

1992년 인천대학교 전자계산공학과(공학사)

1995년 인하대학교 전자계산공학과(공학석사)

2000년 인하대학교 전자계산공학과 박사과정 수료

1995년~1998년 (주)현대정보기술 정보기술연구소 선임연구원 재직

1998년~현재 문경대학교 웹마스터과 교수

관심분야 : 웹 마이닝, 텍스트마이닝, 정보검색, 자연어처리

이 정 현

e-mail : jhlee@inha.ac.kr

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

관심분야 : 자연어처리, HCI, 정보검색, 음성인식, 음성합성, 컴퓨터구조