논문 2006-43CI-4-6

대표 속성을 이용한 최적 연관 이웃 마이닝

(Optimal Associative Neighborhood Mining using Representative Attribute)

정 경 용*

(Kyung-Yong Jung)

요 약

최근 정보 기술의 발전에 따라 다양하고 폭넓은 정보들이 디지털 형태로 빠르게 생산 및 배포되고 있다. 사용자가 이러한 정보과잉 속에서 자신이 원하는 정보를 단시간 내에 검색하는 것은 그리 쉬운 일이 아니다. 따라서 유비쿼터스 상거래에서 사용자가 정보를 효율적으로 이용할 수 있도록 제어하고 필터링하는 일을 도와주는 개인화된 추천 시스템이 등장하였으며, 더 나아가 사용자가 원하는 아이템을 예측하고 추천해주고 있으며 이를 위해 협력적 필터링을 적용하고 있다. 이는 사용자의 성향에 맞는 아이템을 예측하고 추천하기 위하여 비슷한 선호도를 가지는 사용자들간의 유사도 가중치를 계산한다. 본 연구는 정보의 속성에 대한 사용자의 선호도를 고려하지 않은 문제를 개선하기 위하여 연관 이웃 마이닝을 사용하여 대표속성에 대한 연관 사용자의 선호도를 협력적 필터링에 반영하였다. 연관 이웃 마이닝은 선호도에 가장 크게 영향을 미치는 속성을 추출하여 유사한 성향을 가진 연관 사용자를 군집한다. 제안된 방법은 사용자가 아이템에 대해서 평가한 MovieLens 데이터 집합을 대상으로 평가되었으며, 기존의 nearest neighbor model과 K-means 군집보다 그 성능이 우수함을 보인다.

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), 데이터마이닝(Data Mining), 군집(Clustering)

Ⅰ. 서 론

유비쿼터스 상거래에서 추천 시스템은 사용자의 선호도를 추출하고 분석하여 사용자에게 적합한 아이템을 정확하게 예측하여 추천해줄 수 있어야 한다. 이를 위해 일반적으로 협력적 필터링이라고 하는 정보 필터링 기술을 사용한다. 협력적 필터링은 아이템에 대한 선호도 상관관계에 따른 사용자들간의 선호도의 유사도를 구하고 이를 예측하여 아이템에 대한 추천 여부를 결정한다. 유사한 선호도를 갖는 이웃들의 평가에 근거하기 때문에 사용자에게 가장 적합한 이웃들을 적절히 선정해 내는 것이 추천의 정확도를 위해 필요하다[8].

사용자들을 군집하는 방법 중에서 MBR이나 K-NN과 같은 전체 사용자 탐색 방법은 정확도는 높으나 군

* 정회원, 상지대학교 컴퓨터정보공학부
(School of Computer Information Engineering, Sangji University, Korea)
접수일자: 2006년4월27일, 수정완료일: 2006년7월3일

집된 훈련 사용자나 군집할 실험 사용자간의 유사도를 모두 계산해야 하므로 많은 시간을 요구한다[6]. 또한 군집 기반 탐색 방법[13]은 사용자를 분류하는데 소요되는 시간은 단축되나 어떠한 알고리즘으로 군집을 구성하느냐에 따라 군집의 효율성에서 차이를 보인다. 반면, 카테고리 기반 탐색 방법은 같은 카테고리를 갖는 사용자들을 대상으로 군집을 생성하므로 군집의 효율성이 높다. 이러한 방법들은 대부분 데이터의 차원 수가 상대적으로 적을 때 효과적으로 군집할 수 있다. 본 논문에서는 협력적 필터링에서의 {연관 사용자-아이템} 행렬에서 사용자들간의 연관 관계를 유지하면서 차원 수를 감소시키기 위해서 ARHP 알고리즘[7]을 이용하여 연관 사용자 군집을 하며, 연관 사용자 군집을 대상으로 대표속성에 의한 연관 이웃 마이닝을 함으로써 차원수를 감소시킨다.

Ⅱ. 관련연구

협력적 필터링을 이용하는 추천 시스템에서 가장 중요한 단계는 사용자간의 유사도를 계산하는 것이다. 이를 수행하기 위해 먼저 특정 사용자와 유사한 선호도를 가진 이웃 집단을 형성해야 한다.

1. Nearest-Neighbor Model

이웃 선택 과정을 통해 모델을 만들거나 학습하는 과정이다. 이웃 선택의 목적은 각각의 사용자에 대해 순위화된 사용자 리스트를 찾는 것이다. 따라서 이웃 집단을 형성하기 위해서는 두가지 단계를 거쳐야 한다. 먼저 특정 사용자와 모든 다른 사용자 사이의 유사도를 구한다. 그 다음으로 이웃 집단의 규모를 결정한다. 즉, 모든 사용자에 대해 계산된 유사도를 가지고 추천 아이템을 예측하기 위해 몇 명의 이웃을 사용할지 결정한다[3]. 유사도 가중치가 구해진 모든 이웃들을 사용해서 선호도를 예측할 수 있으나 이러한 방법은 정확도나 성능면에서 권장할 방법은 아니다. 반면, 너무 높은 유사도의 이웃들만을 예측에 사용할 경우에는 다른 사용자들과 유사도가 높지 않은 사용자의 아이템에 대해서는 예측할 수 없는 문제점이 발생한다. 그러므로 추천 시스템이 예측할 수 있는 적절한 이웃의 수를 결정하는 것이 무엇보다도 중요하다.

2. K-Means Clustering

K-Means 군집은 데이터 분류에 있어 Maximum Likelihood (ML) 방법[6]의 단순화된 형태이며, 절대적 수렴에 대한 보장이 증명되지 않은 알고리즘이다. 또한 거리 기반 군집화 방법으로 사용자의 선호도를 다차원 공간상의 점으로 표시하고, 거리를 계산함으로써 전체 사용자의 집합을 여러 군집들로 나누는 방법이다. 이는 원할한 수행을 위하여 초기에 군집해야 할 개수를 미리 정해야 하고 또 군집 중심의 초기 값에 따라 군집된 결과의 수렴성이 달라지는 단점이 있다[4]. 그러나 간결성으로 인하여 사용자 군집에 효율적으로 응용되어 왔다. K-Means 군집[2]을 이용하여 사용자를 군집하는 과정은 3단계로 구성한다. 첫 번째 단계에서는 군집의 개수 K와 중심들을 초기화한다. 두 번째 단계에서는 사용자간의 유사도를 기반으로 사용자의 소속을 구한다. 세 번째 단계에서는 소속이 결정된 사용자들을 판별하기 위하여 유사도 평균의 변화치가 임계값보다 낮으면 종료한다.

Ⅲ. 연관 사용자 군집

연관 사용자 군집은 사용자들간의 연관 관계를 유지하면서 {연관 사용자-아이템} 행렬의 차원수를 줄임으로써 예측의 정확도를 높이기 위해서 사용된다. 그림 1은 연관 사용자 군집을 위한 단계적 흐름을 보여주고 있다.

사용자에 의해 선호도가 평가된 아이템들을 사용자 트랜잭션으로 재구성한다. 이를 연관 규칙 탐사 방법을 이용하여 사용자 트랜잭션 안에 빈번하게 동시에 출현하는 사용자들의 집합을 찾는다. 이는 Apriori 알고리즘[1]을 이용하여 고빈도 사용자 집합에서 사용자들간의 연관 규칙을 생성시키고 연관 규칙의 신뢰도를 가중치로 하여 하이퍼 그래프 분할에 의해 군집시키는 ARHP

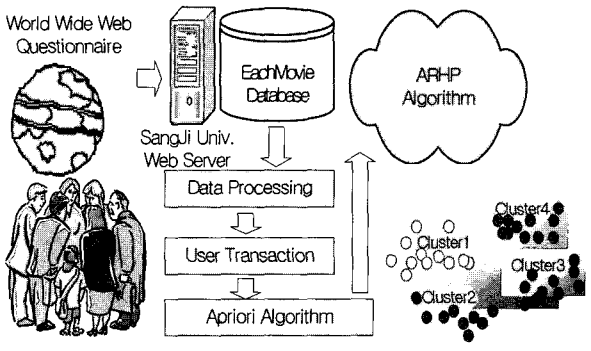

그림 1. 연관 사용자 군집을 위한 흐름도
Fig. 1. Flow for Associative User Group.

알고리즘[7]을 적용함으로써 연관 사용자들을 군집한다.

1. 연관 사용자 마이닝

Apriori 알고리즘을 이용하여 연관 사용자를 마이닝하는 과정을 기술한다. 사용자에 의해 선호도가 평가된 아이템들을 사용자 트랜잭션으로 재구성한다. 고빈도 사용자 집합과 후보 사용자 집합은 사용자 트랜잭션에 나타나는 사용자들이다. 이러한 경우, n번의 데이터베이스 검색이 있을 경우 n단계를 통하여 n개의 사용자로 구성된 연관 사용자를 마이닝한다. 마이닝한 결과 각 사용자는 연관 사용자들의 집합으로 나타내어진다. 이에 따라 연관 사용자 집합{AU_i}은 식 (1)과 같이 표현된다.

$$\{AUi\}=\{(u_{11}\&u_{12}\cdots\&u_{1(r-1)}\ u_{1r}),(u_{21}\&u_{22}\cdots\&u_{2(r-1)}\ u_{2r}),\cdots,(u_{k1}\&u_{k2}\cdots\&u_{k(r-1)}\ u_{kr}),\ldots,(u_{p1}\&u_{p2}\cdots\&u_{p(p-1)}\ u_{pr})\}i\ (i=1,2,\cdots,m) \quad (1)$$

식 (1)의 ($u_{11}\&u_{12}\cdots\&u_{1(r-1)}\&u_{1r}$)는 연관 사용자 집합에서 연관 사용자, ($u_{11}\&u_{12}\cdots\&u_{1(r-1)}\&u_{1r}$)는 연관 사용자를 구성하는 사용자들의 구성, p는 하나의 연관 사용자 집합에서 연관 사용자의 수, r은 연관 사용자를 구성하는 사용자의 수, m은 연관 사용자 집합을 대표하는 연관 사용자의 수이다. 또한 &는 각각의 사용자들이 연관되었음을 의미하는 기호이다.

연관 사용자 형태의 특징을 추출할 경우, 사용되는 지지도, 신뢰도, 향상도에 따라 추출되는 특징은 연관 사용자 군집의 반응 시간과 정확도에서 많은 차이를 보인다. 신뢰도를 결정하기 위한 식 (2)는 다음과 같이 구해진다.

$$Confidence(u_1 \rightarrow u_2) = Pr(u_1|u_2) \quad (2)$$

식 (2)는 사용자 u_1과 u_2의 모든 사용사를 포함하고 있는 사용자 트랜잭션의 수를 사용자 u_i을 포함하고 있는 사용자 트랜잭션의 수로 나눈 결과값이다.

지지도를 결정하기 위한 식 (3)은 전체 사용자들의 쌍 중에 각 연관 사용자의 출현 빈도이다.

$$Support(u_1 \rightarrow u_2) = Pr(u_1 \cap u_2) \quad (3)$$

식 (3)은 사용자 u_1과 u_2의 모든 사용자를 포함하고 있는 사용자 트랜잭션의 수를 데이터베이스 내의 전체 사용자 트랜잭션의 수로 나눈 결과값을 계산한다.

향상도를 결정하기 위한 식 (4)는 다음과 같이 구해진다. 사용자 트랜젝션의 수에서 사용자 u_2를 포함하고 있는 트랜젝션의 비율보다는 사용자 u_1를 포함하는 트랜젝션에서 사용자 u_2를 포함하고 있는 트랜젝션의 비율이 더 클 것이다. 따라서 사용자 u_1과 u_2를 포함하고 있는 트랜젝션이 서로 상호 관련이 없다면, $Pr(u_2|u_1)$은 $Pr(u_2)$와 같게 된다.

$$Lift(u_1 \rightarrow u_2) = \frac{Pr(u_2|u_1)}{Pr(u_2)} = \frac{Pr(u_1 \cap u_2)}{Pr(u_1)Pr(u_2)} \quad (4)$$

식 (4)는 $Pr(u_2|u_1)$의 값은 $Pr(u_2)$의 값보다 향상도가 배수만큼 크다. 그러므로 향상도는 1에 가까우면 독립에 가까운 사건, 1보다 크면 연관 사용자들이 양의 연관 관계, 1보다 작으면 연관 사용자들이 음의 연관 관계를 의미한다. 본 논문에서는 의미있는 연관 사용자의 규칙이 되려면 향상도의 값이 1이상이 되어야 한다.

Apriori 알고리즘을 사용하여 추출된 연관 사용자는 이를 구성하는 수에 따라 반응 시간과 정확도의 차이가 보인다. 따라서 연관 이웃 마이닝의 성능을 향상시키기 위해서 연관 사용자의 구성에 사용되는 사용자의 수를 어떻게 지정해야만 가장 효율적인가를 보인다.

2. 연관 사용자 구성

Apriori 알고리즘은 데이터베이스에 있는 사용자들의 집합에서 연관 사용자들을 마이닝한다. 이때 연관 사용자 벡터를 표현하기 위한 공간은 한정되지 않으며 연관 사용자를 구성하기 위한 사용자의 수도 두 명에서부터 다섯 명으로 다양하게 구성된다. 표 1은 1,000명의 사용자들을 대상으로 연관 사용자를 구성하기 위한 사용자의 수에 따라 차이가 있음을 보인다.

표 1에서 2-AU는 2명의 단일 사용자로 구성된 연관 사용자, 3-AU는 3명의 단일 사용자로 구성된 연관 사용자, 4-AU는 4명의 단일 사용자로 구성된 연관 사용자, 마지막으로 5-AU는 5명의 단일 사용자로 구성된 연관 사용자를 나타낸다. 표 1은 연관 사용자의 구성이 2-AU 형태일 경우 추출된 특징의 개수는 149,894개로, 특징의 개수가 각각 12,936개, 3,822개, 191개인 3-AU, 4-AU, 5-AU보다 많음을 보인다.

표 2는 2-AU, 3-AU, 4-AU 형태의 연관 사용자를 ARHP 알고리즘을 이용하여 군집하였을 경우의 반응

표 1. 추출된 특징의 수
Table 1. Number of Extracted Feature.

	2-AU	3-AU	4-AU	5-AU
연관 사용자의 특징 수	149,894	12,936	3,822	191

표 2. 연관 사용자 군집의 반응 시간 및 정확도
Table 2. Response Time & Accuracy of Associative User Group.

	2-AU	3-AU	4-AU
반은시간(sec)	40	20	10
정확도	89	92	71

시간과 정확도를 보인다. 5-AU 형태의 연관 사용자는 그림 3에서와 같이 특징으로 추출될 확률이 거의 희박하므로 반응 시간과 정확도 실험의 대상에서 제외한다. 2-AU, 3-AU, 4-AU 형태로 구성된 연관 사용자의 군집의 반응 시간을 비교하기 위하여 초(second)를 사용한다. 반응 시간은 ARHP 알고리즘이 연관 사용자를 군집하는데 걸리는 시간이다. 연관 사용자 구성에 따른 연관 사용자의 군집의 정확도는 F-Measure 측정식을 이용하여 비교 평가하였다.

표 2에서와 같이 2-AU 형태로 연관 사용자를 군집하였을 경우 반응 시간이 크게 증가함을 알 수 있다. 이는 표 1에서와 같이 2-AU 형태의 특징의 수가 많으므로 연관 사용자의 군집하는데 걸리는 반응 시간이 증가한 것이다. 3-AU, 4-AU 형태의 반응 시간은 비교적 좋은 성능을 보인다. 표 2에서와 같이 3-AU 형태로 연관 사용자의 군집을 하였을 경우 정확도는 89.3으로, 2-AU보다는 4.07%, 4-AU보다는 13.87% 높다. 2-AU의 형태로 군집하였을 경우, 군집된 연관 사용자 수가 많으므로 다차원 벡터로 표현은 가능하나, 이로 인한 잡음의 영향으로 군집의 정확도가 저하된다. 4-AU의 형태로 연관 사용자 군집을 하였을 경우, 군집된 연관 사용자의 수가 너무 적으므로 연관 사용자의 군집을 구성할 수 없다. 따라서 4-AU 형태의 연관 사용자 군집의 정확도는 2-AU, 3-AU보다 낮다. 그러므로 연관 사용자 군집을 위한 연관 사용자 형태는 3명의 단일 사용자로 구성된 연관 사용자 형태(3-AU)로 특징 추출하는 것이 반응 시간과 정확도를 높이는데 가장 효과적이다.

3. ARHP 알고리즘에 의한 연관 사용자 군집

ARHP(Association Rule Hypergraph Partitioning) 알고리즘은 연관 규칙과 하이퍼 그래프 분할을 이용하여 트랜잭션 기반의 데이터베이스에서 연관된 아이템들을 군집하는 방법이다[7]. 하이퍼 그래프 H={V, E}는 아이템들로 구성된 정점들의 집합 V와 빈번한 아이템 집합들을 나타내는 하이퍼 간선들의 집합 E로 구성된다. 하이퍼 그래프 분할 알고리즘은 항목들간의 거리가 아닌 가중치를 이용하기 때문에 아이템들간의 거리 계산이 어려운 다차원 데이터 집합에 대한 군집에 유용하다. 본 논문에서는 ARHP 알고리즘은 군집하기 위한 연관 사용자 집합들의 모든 연관 규칙과 신뢰도를 구한 후, 연관 규칙에 포함되는 사용자를 정점으로, 연관 관계를 하이퍼 간선으로 매핑한다. 그리고 신뢰도를 하이퍼 그래프 분할을 위한 가중치로 하여, 연관 사용자들의 군집을 구한다. 하이퍼 그래프 분할에서의 클러스터는 유사한 사용자들을 분류하거나 예측할 때 사용되고 관심없는 연관 규칙을 제거함으로써 연관 규칙의 차원 수를 감소시키는데 사용된다.

4. 연관 사용자 군집 절차

본 절에서는 Apriori 알고리즘을 이용하여 사용자 트랜잭션으로부터 연관 사용자의 군집하는 절차를 보인다. MovieLens 평가 데이터에서 사용자에 의해 선호도를 평가한 아이템들을 표 3의 사용자 트랜잭션으로 재구성한다. 표 3에서 트랜잭션 번호는 사용자가 평가한 아이템을 의미하며 추출된 사용자는 후보 사용자 집합과 고빈도 사용자 집합을 구성하기 위한 것이다. 표 3의 사용자 트랜잭션으로부터 Apriori 알고리즘으로 연관 규칙을 표 4에서 제시한 방법으로 마이닝한다[9,10].

표 4는 표 3에 나타난 추출된 사용자를 Apriori 알고리즘에 적용한 결과를 보인다. 여기서 사용자는 2절의 연관 사용자 구성에서 제시한 바와 같이 3개의 단일 사용자로 구성된 연관 사용자 형태를 사용한다.

Apriori 알고리즘은 첫 단계에서 후보 사용자 집합(C_1)을 구성하며 이들의 지지도를 확인하기 위해 데이터베이스를 검색하고, 고빈도 사용자 집합(L_1)을 구성할 수 있다. 이와 같은 방법으로 Apriori 알고리즘의 두번째 단계에서는 C_2, L_2를 구성하며, Apriori 알고리즘의 세번째 단계에서는 C_3, L_3를 구성한다. 표 2에서 제시한

표 3. 연관 사용자 마이닝을 위한 사용자 트랜잭션
Table 3. User Transaction for Associative User Mining.

트랜잭션 번호	추출된 사용자
1	u_1, u_2, u_3, u_4
2	u_2, u_3, u_1, u_5, u_6, u_7, u_{12}, u_8
3	u_9, u_3, u_2, u_{10}, u_5, u_{11}
4	u_{13}, u_3, u_{14}, u_{15}, u_{16}, u_{17}
5	u_{18}, u_{13}, u_3
6	u_{13}, u_3, u_{19}, u_{20}, u_{15}
7	u_{21}, u_{22}
8	u_{23}, u_{24}, u_{25}

표 4. Apriori 알고리즘에 의한 연관 사용자를 추출
Table 4. Extract of Associative User using Apriori Alg.

후보 사용자 집합(C_1)	$u_{1(2)}$,$u_{2(3)}$,$u_{3(6)}$,$u_{4(1)}$,$u_{5(2)}$,$u_{6(1)}$,$u_{7(1)}$,$u_{8(1)}$, $u_{9(1)}$,$u_{10(1)}$,$u_{11(1)}$,$u_{12(1)}$,$u_{13(1)}$,$u_{14(1)}$,$u_{15(2)}$, $u_{16(1)}$,$u_{17(1)}$,$u_{18(1)}$,$u_{19(1)}$,$u_{20(1)}$,$u_{21(1)}$,$u_{22(1)}$, $u_{23(1)}$,$u_{24(1)}$,$u_{25(1)}$
고빈도 사용자 집합(L_1)	$u_{1(2)}$,$u_{2(3)}$,$u_{3(6)}$,$u_{5(2)}$,$u_{13(3)}$,$u_{15(2)}$
후보 사용자 집합(C_2)	$(u_1,u_2)_{(2)}$,$(u_1,u_3)_{(2)}$,$(u_1,u_5)_{(1)}$,$(u_1,u_{13})_{(0)}$, $(u_1,u_{15})_{(0)}$,$(u_2,u_3)_{(3)}$,$(u_2,u_5)_{(2)}$,$(u_2,u_{13})_{(0)}$, $(u_2,u_{15})_{(0)}$,$(u_3,u_5)_{(2)}$,$(u_3,u_{13})_{(3)}$,$(u_3,u_{15})_{(2)}$, $(u_5,u_{13})_{(0)}$,$(u_5,u_{15})_{(0)}$,$(u_{13},u_{15})_{(2)}$
고빈도 사용자 집합(L_2)	$(u_1,u_2)_{(2)}$,$(u_1,u_3)_{(2)}$,$(u_2,u_3)_{(3)}$,$(u_2,u_5)_{(2)}$, $(u_3,u_5)_{(2)}$,$(u_3,u_{13})_{(3)}$,$(u_3,u_{15})_{(2)}$,$(u_{13},u_{15})_{(2)}$
후보 사용자 집합(C_3)	$(u_1,u_2,u_3)_{(2)}$,$(u_1,u_2,u_5)_{(0)}$,$(u_1,u_2,u_{13})_{(0)}$, $(u_1,u_2,u_{15})_{(0)}$,$(u_1,u_3,u_5)_{(1)}$,$(u_1,u_3,u_{13})_{(0)}$, $(u_1,u_3,u_{15})_{(0)}$,$(u_2,u_3,u_5)_{(2)}$,$(u_2,u_3,u_{13})_{(0)}$, $(u_2,u_3,u_{15})_{(0)}$,$(u_2,u_5,u_{15})_{(2)}$,$(u_2,u_5,u_{13})_{(0)}$, $(u_3,u_5,u_{13})_{(0)}$,$(u_3,u_5,u_{15})_{(0)}$,$(u_3,u_{13},u_{15})_{(2)}$, $(u_{13},u_{15},u_1)_{(0)}$,$(u_{13},u_{15},u_2)_{(1)}$,$(u_{13},u_{15},u_3)_{(0)}$, $(u_{13},u_{15},u_5)_{(0)}$
고빈도 사용자 집합(L_3)	$(u_1,u_2,u_3)_{(2)}$,$(u_2,u_3,u_5)_{(2)}$,$(u_2,u_5,u_{15})_{(2)}$, $(u_3,u_{13},u_{15})_{(2)}$

표 5. 신뢰도를 이용한 가중치 부여
Table 5. Weighting using Confidence.

연관규칙	신뢰도	{u_1, u_2, u_3}의 평균 신뢰도
{u_1} {u_2, u_3}	80%	60%
{u_1, u_2} {u_3}	40%	
{u_1, u_3} {u_2}	60%	
{u_2, u_3} {u_1}	80%	
{u_3} {u_1, u_2}	60%	

표 6. 하이퍼 그래프 분할한 결과
Table 6. Result of Hyper Graph Partitioning.

연관 사용자 군집	추출된 사용자
1	{u_{13}, u_{14}, u_{15}, u_{16}, u_{17}, u_{18}, u_{19}, u_{20}, u_{22}, u_{23}}
2	{u_2, u_3, u_4, u_5, u_9, u_{10}, u_{11}}
3	{u_1, u_6, u_7, u_8, u_{12}, u_{21}, u_{24}, u_{25}}

바와 같이 L_3의 연관 사용자 집합, {u_1, u_2, u_3}, {u_2, u_3, u_5}, {u_2, u_5, u_{15}}, {u_3, u_{13}, u_{15}}으로 추출된다.

연관 사용자 집합에서 모든 연관 규칙과 신뢰도를 구한 후, ARHP 알고리즘을 이용하여 연관 사용자 군집을 한다. 여기서 표 4의 L_3의 고빈도 사용자 집합에서 하이퍼 그래프의 분할을 위한 가중치는 연관 규칙의 평균 신뢰도를 사용한다. 예를들면 L_3의 고빈도 사용자 집합이 {u_1, u_2, u_3}라면, ARHP 알고리즘을 위한 하이퍼 그래프는 사용자들로 구성된 정점들의 집합 {u_1, u_2, u_3}과 연관 규칙으로 연결된 하이퍼 간선들의 집합으로 구성된다.

하이퍼 그래프 분할을 위한 가중치는 하이퍼 간선의 모든 사용자들을 포함하는 연관 규칙의 신뢰도를 사용한다. 연관 사용자 집합 {u_1, u_2, u_3}에서 연관 규칙의 신뢰도를 이용하여 가중치는 표 5와 같이 부여할 수 있다. 여기서 추출된 연관 사용자 집합에서 모든 연관 규칙들의 평균 신뢰도를 구한다. 표 5에서 {u_1, u_2, u_3}의 평균 신뢰도 60%가 하이퍼 그래프 분할을 위한 가중치이다. 연관 사용자 집합에서 ARHP 알고리즘을 이용해서 하이퍼 그래프 분할을 한 군집 결과는 표 6과 같다. 결과적으로, ARHP 알고리즘은 표 3에 나타난 25명의 사용자를 표 6과 같이 3개의 연관 사용자 군집을 한다.

Ⅳ. 대표속성을 이용한 최적 연관 이웃 마이닝

협력적 필터링은 자동화된 프로세스로는 쉽게 분석될 수 없는 정보의 질을 기존 사용자들의 선호도를 통해 어느 정도 반영한다는 장점이 있으나, 정보의 속성에 대한 사용자의 선호도는 고려하지 않는다는 문제점을 가지고 있다. 따라서 본 논문에서는 정보의 대표적인 속성에 대한 연관 사용자의 선호도를 협력적 필터링에 반영함으로써 추천의 정확도를 높이고자 한다.

ARHP 알고리즘에 의한 연관 사용자 군집에서 사용자의 대표속성을 추출한 후, 성별과 나이에 의한 연관 이웃 마이닝을 한다. 협력적 필터링에서의 유사한 이웃 선택에서 연관 사용자들 간의 유사도 가중치를 구하기 위해서 성별과 나이에 의한 대표속성을 사용한다. 최적 연관 이웃 마이닝은 이를 구성하는 연관 이웃의 수에 따라 정확도의 차이가 보인다. 따라서 정확도의 성능을 향상시키기 위해서 연관 이웃 마이닝의 수를 어떻게 지정해야만 효율적인가를 제시한다.

1. 연관 사용자의 대표속성 추출

효율적인 협력적 필터링을 수행하기 위해서 대표속성을 중심으로 특정 사용자와 유사한 선호도를 가지는 연관 이웃을 찾아내는 것이다. 기존의 이웃 선정에 사용된 방법들은 정보에 대한 선호도의 정도만을 반영하여 이웃의 수를 결정하는데 사용하였다. 이는 전체 선호도 정보들을 모두 사용하여 유사도 가중치를 구하는 것이므로 정보의 대표속성값들에 대해 사용자가 차별적인 선호도를 가지는 경우 이를 제대로 이웃 선정에 반영하지 못하는 단점이 있다. 이를 보완하기 위해서

알고리즘 1. 연관 사용자의 대표속성
Alg. 1. Representative Attribute for Associative User.

```
Algorithm 연관 사용자의 대표속성을 결정
    Input: 연관 사용자가 선호도를 평가한 아이템들
           Score of Item[k]
    Output: 연관 사용자의 대표속성  MainGenreID
GenreSum[Num_Genre]  0, GenreCount[Num_Genre]  0
for k is items that user rated do
    for c is the genre of Item[k] do
        GenreSum[c]  GenreSum[c] + Score of Item[k]
        GenreCount[c]++
    endfor
endfor
for j=1 to Num_Genre do
    MainGenreID
        Max(MainGenreID, GenreSum[j]/GenreCount[j])
endfor // 사용자의 대표속성을 결정
Representative Attribute[MainGenreID]  Add UserID
return
```

대표속성 추출에 의해 얻어진 대표속성값에 한정하여 연관 사용자 군집 속에서 연관 이웃들을 찾아내어 예측에 이용하는 대표속성을 이용한 최적 연관 이웃 마이닝을 사용한다.

알고리즘 1은 연관 사용자의 대표속성을 결정하는 방법이다. 연관 사용자가 선호도를 평가한 아이템을 이용하여 사용자의 대표속성을 구한다. 대표속성은 장르별 아이템의 선호도 합을 구한 후 선호도 합의 평균이 가장 큰 장르이다. 연관 사용자의 대표속성을 추출하는 이유는 실험 데이터로 쓰이는 MovieLens 평가 데이터[11]에서 아이템에 대한 속성이 하나 이상인 것이 많아 사용자가 보는 관점에 따라 속성이 달라지기 때문이다.

2. 대표속성을 이용한 최적 연관 이웃 마이닝

본 논문에서는 연관 사용자를 마이닝하여 ARHP 알고리즘에 의한 연관 사용자 군집에서 사용자의 대표속성을 추출한다. 여기에 협력적 필터링에서 최적 연관 이웃을 선택하기 위하여 성별과 나이에 의한 대표속성을 사용한다. 이는 협력적 필터링에서 이웃 선정할 때 대표속성에 의한 연관 이웃 마이닝을 사용한다. 연관 이웃 마이닝은 같은 성별 또는 같은 나이를 가진 사람들이 각각의 아이템에 대해서 유사한 선호도를 가진다고 가정한다. 성별과 나이를 연관 이웃 마이닝에 적용한 이유는 남성과 여성간의 성별 차이와 세대차를 통해서 추천의 정확도를 높이기 위함이다.

알고리즘 2는 대표속성에 의한 최적 연관 이웃을 마이닝하는 방법이다. 이는 협력적 필터링에서 아이템의

알고리즘 2. 대표속성을 이용한 최적 연관 이웃 마이닝
Alg. 2. Optimal Associative Neighborhood Mining using Representative Attribute.

```
Algorithm Represnetative Attribute-Neighborhood
    Input: Num_class   # of associative user in GenreID
           Num_gender   # of associative user in Gender
           Num_age   # of associative user in Age
    Output: AssociativeUserGroup(i,j,k)
for i=1 to Num_class do
    for j=1 to Num_gender do
        for k=1 to Num_age do
            AssociativeUserGroup(i,j,k)
                조건 만족하는 연관 사용자 군집
        endfor
    endfor
endfor
return
```

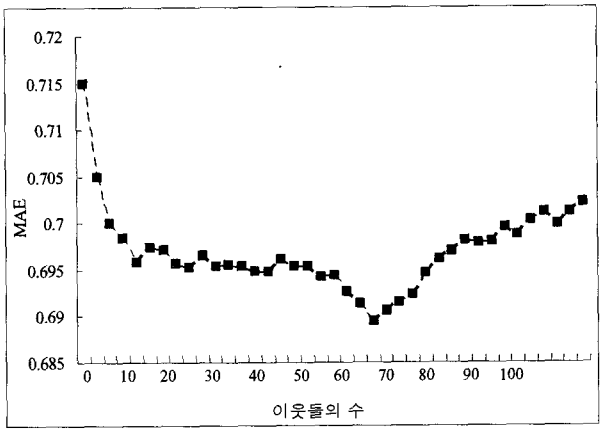

그림 2. 이웃들의 수에 따른 예측의 정확도
Fig. 2. Prediction Accuracy of varying Neighborhood.

예측에 사용될 연관 이웃의 수를 결정하기 위해서 사용된다. 개인화 아이템 추천 시스템에서 실시간 예측을 하기 위해서 대표속성을 이용한 최적 연관 이웃 마이닝은 적절한 연관 이웃의 수를 결정해야 한다. 실험을 통해서 적절한 연관 이웃의 수를 결정하기 위해서 이웃의 수를 증가시킴에 따라 정확도를 비교 평가하였다. 그림 2는 연관 이웃의 수에 따른 예측의 정확도이다.

그림 2에서 MAE[12]의 정확도를 보면 연관 이웃의 수가 증가함에 따라 예측의 정확도가 일관성있게 좋아지지 않는다. 대략 연관 이웃의 수가 69 정도에 해당되는 곳에서부터 정확도가 감소되는 것을 볼 수 있다. 연관 이웃의 수를 결정하는 실험은 협력적 필터링에서 사용하는 피어슨 상관계수에 의한 사용자 유사도 가중치를 구하는 부분에 적용한 것이다.

V. 성능 평가

대표속성을 이용한 최적 연관 이웃 마이닝의 성능 평가를 위해 실험 데이터로는 Grouplens Research Center의 MovieLens 평가 데이터를 사용하였다. MovieLens 평가 데이터 집합[11]은 6,040의 사용자들이 3,960의 영화에 대해서 총 1,000,000의 평가를 하였다. 본 논문에서는 469명의 사용자들 데이터 집합으로부터 무작위로 선택하였으며, 그 사용자들은 0에서 1까지 0.2의 간격으로 아이템에 대하여 평가를 하였다.

실험을 위해 대표속성에 의한 최적 연관 이웃 마이닝은 협력적 필터링에서 기존의 이웃 선정에 사용되었던 Nearest Neighbor Model[2]과 K-Means 군집[5,6]의 결과와 비교하였다. 이를 위해 내용 정보 데이터베이스의 사용자 469명을 대상으로 연관 이웃 마이닝을 위한 실험을 진행하였다. 그 결과 최종적으로 20개의 군집으로 전체 360명의 사용자가 각 그룹으로 군집되었다. 성능을 평가하기 위해서 각 그룹으로 군집된 연관 사용자들을 대상으로 정보 검색에서 쓰이는 재현율과 정확도를 사용한다. 여기서 재현율과 정확도를 합한 단위인 F-measure 측정식은 식 (5)와 같이 정의한다.

$$F-Measure=\frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R} \quad (5)$$

식 (5)에서 P는 정확도, R은 재현율을 의미하며, F-measure의 값이 클수록 분류가 우수함을 의미한다. 여기서 β는 정확도에 대한 재현율의 상대적인 가중치를 나타내는 수치로 1.0일 경우 정확도와 재현율의 가중치가 같다[12]. 본 실험에서는 β의 값을 1.0으로 설정하여 그룹별로 F-measure의 분류 결과를 분석해 보았다. 대표속성에 의한 최적 연관 이웃 마이닝은 ANS, Nearest Neighbor Model은 NNM, K-Means 군집 방법은 K-MC으로 표기하였다.

그림 3과 그림 4는 식 (5)를 이용해서 정확도와 F-measure의 성능 곡선을 나타낸다. 그림 3에서 대표속성에 의한 최적 연관 이웃 마이닝은 K-MC 방법보다는 22.33%, NNM 방법보다는 5.09%의 높은 정확도를 나타낸다. 그림 4에서 β=1.0일 경우, 대표속성에 의한 최적 연관 이웃 마이닝(ANS)이 K-MC 방법보다는 15.44%, NNM 방법보다는 4.1% 향상된 F-measure의 결과를 나타낸다.

그림 5는 사용자의 수에 따른 F-measure의 성능 변화를 나타낸다. 세가지 방법에 대해서 사용자의 수가

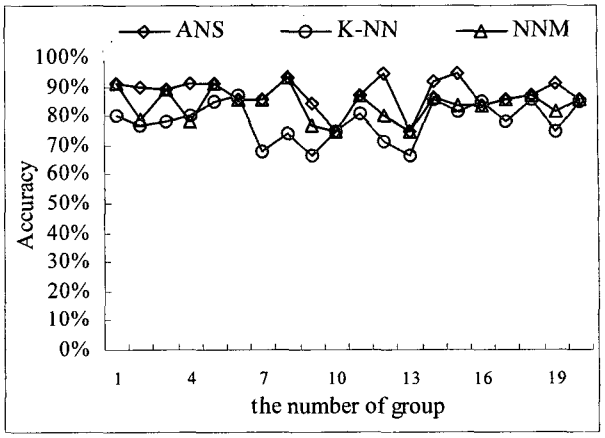

그림 3. 사용자 군집의 정확도에 의한 성능 평가
Fig. 3. Accuracy of varying Groups.

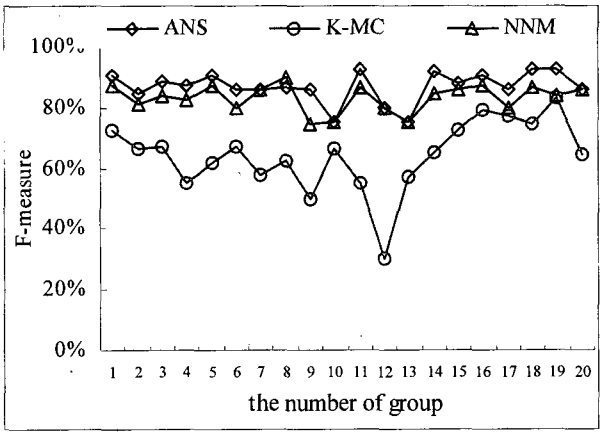

그림 4. F-measure에 의한 그룹별 성능 평가
Fig. 4. F-measure of varying Groups.

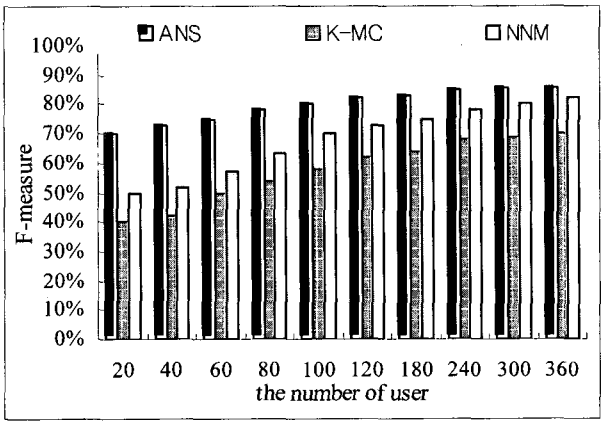

그림 5. 사용자의 수에 따른 F-measure의 성능 평가
Fig. 5. F-measure of varying Users.

증가함에 따라 점차 F-measure의 성능이 점차 향상됨을 보인다. 특히, 대표속성에 의한 연관 이웃 마이닝(ANS)은 사용자의 수가 적은 경우에도 높은 성능을 나타낸다. 그러나 K-MC 방법과 NNM 방법은 사용자 수

가 적은 경우 낮은 성능을 나타낸다. 전체적으로 대표속성에 의한 최적 연관 이웃 마이닝이 K-MC 방법과 NNM 방법보다 성능이 우수함을 알 수 있다.

VI. 결 론

협력적 필터링에서 이웃 선정에 사용되었던 기존의 방법들은 정보에 대한 선호도의 정도만을 반영하여 이웃의 수를 결정하는데 사용하였다. 이는 전체 선호도 정보들을 모두 사용하여 유사도 가중치를 구하는 것이므로 정보의 대표속성값들에 대해 사용자가 차별적인 선호도를 가지는 경우 이를 제대로 이웃 선정에 반영하지 못하는 단점이 있다. 이를 보완하기 위해서 본 논문에서 대표속성을 이용한 최적 연관 이웃 마이닝을 제안하였다. 정보의 속성에 대한 사용자의 선호도를 고려하지 않은 문제를 개선하기 위하여 연관 이웃 마이닝을 사용하여 대표속성에 대한 연관 사용자의 선호도를 협력적 필터링에 반영하였다. 여기서 연관 이웃 마이닝은 선호도에 가장 크게 영향을 미치는 속성을 추출하여 유사한 성향을 가진 연관 사용자를 군집한다. 제안한 방법의 성능을 평가하기 위하여 기존의 Nearest Neighbor Model과 K-Means 군집과 비교하여 분석하였다. 그 결과, 제안한 방법이 K-Means 군집보다는 18.88%, Nearest Neighbor Model보다는 4.58%의 높은 성능 차이를 보였다. 또한 K-Means 군집이나 Nearest Neighbor Model은 사용자가 적은 환경에서 낮은 성능을 나타냈으나 제안한 방법은 비교적 높은 성능을 나타냄을 보였다.

참 고 문 헌

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

[2] K. Alsabti, S. Ranka, and V. Singh, "An Efficient K-means Clustering Algorithm," Proceedings of the 1stWorkshop on High-Performance Data Mining, 1998.

[3] S. Brin, "Near Neighbor Search in Large Metric Spaces," In Proc. of the 21th International Conference on Very Large Data Bases, pp. 574-584, 1995.

[4] G. Casella and E. I. Gerge, "Explaining the Gibbs Sampler," Journal of the American Statistician, Vol. 46, pp. 167-174, 1992.

[5] M. O. Connor and J. Herlocker, "Clustering Items for Collaborative Filtering," In Proc. of the ACM SIGIR Workshop on Recommender Systems, Berkeley, CA, 1999.1.

[6] C. Ding and X. He, "K-Means Clustering via Principal Component Analysis," In Proc. of the 21th Int. Conf. on Machine Learning, pp. 225-232, 2004.

[7] E. H. Han, G. Karypis, and V. Kumar, "Clustering based on Association Rule Hypergraphs," In Proc. of the SIGMOD'97 Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 9-13, 1997.

[8] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating Collaborative Filtering Recommender Systems," ACM Transactions on Information Systems (TOIS) archive, Vol. 22, No. 1, pp. 5-53, 2004.

[9] K. Y. Jung and J. H. Lee, "User Preference Mining through Hybrid Collaborative Filtering and Content-based Filtering in Recommendation System," IEICE Transaction on Information and Systems, Vol. E87-D, No. 12, pp. 2781-2790, 2004.

[10] S. J. Ko and J. H. Lee, "Feature Selection using Association Word Mining for Classification," LNCS 2113, In Proc. of the International Conference on Database and Expert Systems Applications, pp. 211-220, 2001.

[11] MovieLens Collaborative Filtering Data Set, http://www.cs.umn.edu/research/GroupLens/, Grouplens Research Project, 2000.

[12] R. Raymond, J. Mooney, and L. Roy, "Content-Based Book Recommending Using Learning for Text Categorization," In Proc. of the 5th ACM Conference on Digital Libraries, pp. 195-204, 2000.

저 자 소 개

정 경 용(정회원)

2000년 인하대학교 전자계산 공학과 (공학사)

2002년 인하대학교 컴퓨터정보 공학과 (공학석사)

2005년 인하대학교 컴퓨터정보 공학과 (공학박사)

2001년~2005년 에이플러스전자 책임연구원

2002년~2005년 가천길대학 겸임교수

2006년~현재 상지대학교 컴퓨터정보공학부 교수

<주관심분야 : 데이터마이닝, HCI, 정보검색, 감성공학, 임베디드시스템, 컴퓨터구조>