개인화 추천 시스템에서 속성 정보를 이용한 연관 사용자 군집 방법

한경수°, 조동주, 정경용*
상지대학교 컴퓨터정보공학부 지능시스템연구실
상지대학교 컴퓨터정보공학부
minosia@hotmail.com°, {queen8181, kyjung*}@sangji.ac.kr

Associative User Group Method using Attribute Information in Personalized Recommendation System

Kyung-Soo Han°, Dong-Ju Cho, Kyung-Yong Jung*
Intelligent System Lab, School of Computer Information Engineering, Sangji University
School of Computer Information Engineering, Sangji University*

요 약

유비쿼터스 상거래에서 사용자가 정보를 효율적으로 이용할 수 있도록 제어하고 필터링하는 일을 도와주는 개인화된 추천 시스템이 등장하였다. 더 나아가서는 사용자가 원하는 아이템을 예측하고 추천해주며, 이를 위해 협력적 필터링 기술을 적용하고 있다. 이는 사용자의 성향에 맞는 아이템을 예측하고 추천하기 위하여 비슷한 선호도를 가지는 사용자들 간의 유사도 가중치를 계산한다. 본 논문에서는 속성정보에 대한 사용자의 선호도를 고려하지 않은 문제점을 개선하기 위해서 속성정보를 이용한 연관 사용자의 선호도를 협력적 필터링 기술에 반영함으로써 추천의 정확도를 높이고자 한다. 그리고 협력적 필터링의 {연관 사용자-아이템} 행렬에서 사용자들 간의 연관 관계를 유지하면서 차원 수를 감소시키기 위해 ARHP 알고리즘을 이용하여 연관 사용자 군집을 한다. 제안된 방법의 성능 평가를 하기 위해 사용자가 아이템에 대해서 평가한 MovieLens 데이터 집합을 대상으로 평가되었으며, 기존의 Nearest Neighbor Model과 K-Means 군집보다 그 성능이 우수함을 보인다.

1. 서 론

개인화된 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 사용자들 간의 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[1].

사용자들을 군집하는 방법으로 MBR이나 K-NN과 같은 전체 사용자 탐색 방법이 있는데, 이는 정확도가 높으나 군집된 훈련 사용자나 군집할 실험 사용자간의 유사도를 모두 계산해야 하므로 많은 시간을 요구한다[2]. 또한 군집 기반 탐색 방법은 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 효과적인 군집을 할 수 있다. 본 논문에서는 협력적 필터링의 {연관 사용자-아이템} 행렬에서 사용자들 간의 연관 관계를 유지하면서 차원 수를 감소시키기 위해 ARHP 알고리즘[3]을 이용하여 연관 사용자 군집을 한다.

2. 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. 기존의 방법으로는 Nearest Neighbor Model[2]과 K-Means 군집[2,4]이 있다.

2.1. 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 두 가지 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[5]. 유사도 가중치가 구해진 모든 이웃들을 사용해서 선호도를 예측할 수 있으나 이러한 방법은 정확도나 성능 면에서 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.

2.2. K-Means 군집

K-Means 군집은 데이터 분류에 있어 Maximum Likelihood (ML) 방법[2]의 단순화된 형태이며, 절대적 수렴에 대한 보장

이 증명되지 않은 알고리즘이다. 또한 거리 기반 군집화 방법으로 사용자의 선호도를 다차원 공간상의 점으로 표시하고, 거리를 계산함으로써 전체 사용자의 집합을 여러 군집들로 나누는 방법이다. 이는 원활한 수행을 위하여 초기에 군집해야 할 개수를 미리 정해야 하고 또 군집 중심의 초기 값에 따라 군집된 결과의 수렴성이 달라지는 단점이 있다. 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.

3. 개인화 추천 시스템에서 연관 사용자 구성

3.1. ARHP 알고리즘에 의한 연관 사용자 군집

ARHP(Association Rule Hypergraph Partitioning) 알고리즘은 연관 규칙과 하이퍼 그래프 분할을 이용하여 트랜젝션 기반의 데이터베이스에서 연관된 아이템들을 군집하는 방법이다[3]. 하이퍼 그래프 H={V, E}는 아이템들로 구성된 정점들의 집합 V와 빈번한 아이템 집합들을 나타내는 하이퍼 간선들의 집합 E로 구성된다. 하이퍼 그래프 분할 알고리즘은 항목들 간의 거리가 아닌 가중치를 이용하기 때문에 아이템들 간의 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 사용자 군집을 구한다. 하이퍼 그래프 분할에서의 클러스터는 유사한 사용자들을 분류하거나 예측할 때 사용되고 관심 없는 연관 규칙을 제거함으로써 연관 규칙의 차원 수를 감소시키는데 사용된다.

3.2. 연관 사용자 군집 절차

본 절에서는 Apriori 알고리즘을 이용하여 사용자 트랜젝션으로부터 연관 사용자의 군집하는 절차를 보인다. MovieLens 평가 데이터에서 사용자에 의해 선호도를 평가한 아이템들을 표 1의 사용자 트랜젝션으로 재구성한다. 표 1에서 트랜젝션 번호는 사용자가 평가한 아이템을 의미하며 추출된 사용자는 후보 사용자 집합과 고빈도 사용자 집합을 구성하기 위한 것이다. 표 1의 사용자 트랜젝션으로부터 Apriori 알고리즘으로 연관 규칙을 표 2에서 제시한 방법으로 마이닝한다[6,7].

표 1. 연관 사용자 군집을 위한 사용자 트랜젝션

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

표 2는 표 1에 나타난 추출된 사용자를 Apriori 알고리즘에 적용한 결과를 보인다. 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 두 번째 단계에서는 C_2, L_2를 구성하며, Apriori 알고리즘의 세 번째 단계에서는 C_3, L_3를 구성한다. 표 2에서 제시한 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.

표 2. Apriori 알고리즘에 의한 연관 사용자를 추출

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

연관 사용자 집합에서 모든 연관 규칙과 신뢰도를 구한 후, ARHP 알고리즘을 이용하여 연관 사용자 군집을 한다. 여기서 표 2의 L_3의 고빈도 사용자 집합에서 하이퍼 그래프의 분할을 위한 가중치는 연관 규칙의 평균 신뢰도를 사용한다. 예를 들어 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.

표 3. 신뢰도를 이용한 가중치 부여

$\{u_1\} \rightarrow \{u_2,u_3\}$	80%
$\{u_1,u_2\} \rightarrow \{u_3\}$	40%
$\{u_1,u_3\} \rightarrow \{u_2\}$	60%
$\{u_2,u_3\} \rightarrow \{u_1\}$	80%
$\{u_3\} \rightarrow \{u_1,u_2\}$	60%
$\{u_1,u_2,u_3\}$의 평균 신뢰도	60%

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 3과 같이 부여할 수 있다. 여기서 추출된 연관 사용자 집합에서 모든 연관 규칙들의 평균 신뢰도를 구한다. 표 3에서 $\{u_1,u_2,u_3\}$의 평균 신뢰도 60%가 하이퍼 그래프 분할을 위한 가중치이다. 연관 사용자 집합에서 ARHP 알고리즘을 이용해서 하이퍼 그래프 분할을 한 군집 결과는 표 4와 같다. 결과적으로, ARHP 알고리즘은 표 1에 나타난 25명의 사용자를 표 4와 같이 3개의 연관 사용자 군집을 한다.

표 4. 하이퍼 그래프 분할 결과

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	$\{u_1,u_6,u_7,u_8,u_{12},u_{21},u_{24},u_{25}\}$

4. 속성정보를 이용한 연관 사용자 군집

협력적 필터링은 자동화된 프로세스로 쉽게 분석될 수 없는

정보의 질을 기존 사용자들의 선호도를 통해 어느 정도 반영한다는 장점이 있으나, 속성정보에 대한 사용자의 선호도는 고려하지 않는다는 문제점을 가지고 있다. 따라서 본 논문에서는 속성정보에 대한 연관 사용자의 선호도를 협력적 필터링에 반영함으로써 추천의 정확도를 높이고자 한다.

ARHP 알고리즘에 의한 연관 사용자 군집에서 사용자의 속성을 추출한 후, 성별과 나이에 의한 연관 사용자를 군집한다. 협력적 필터링에서 유사한 사용자 선택에 연관 사용자들 간의 유사도 가중치를 구하기 위해서는 성별과 나이에 의한 속성을 사용한다.

4.1. 연관 사용자의 속성정보 추출

효율적인 협력적 필터링을 수행하기 위해서 속성정보를 중심으로 특정 사용자와 유사한 선호도를 가지는 연관 사용자를 찾아내는 것이다. 기존의 사용자 선정에 사용된 방법들은 정보에 대한 선호도의 정도만을 반영하여 사용자의 수를 결정하는데 사용하였다. 이는 전체 선호도 정보들을 모두 사용하여 유사도 가중치를 구하는 것이므로 속성정보들의 값에 대해 사용자가 차별적인 선호도를 가지는 경우 이를 제대로 반영하지 못하는 단점이 있다. 이를 보완하기 위해서 속성정보 추출에 의해 얻어진 속성정보의 값에 한정하여 연관 사용자를 군집한다.

알고리즘 1. 연관 사용자의 속성정보 추출

```
Algorithm 연관 사용자의 속성을 결정
 Input: 연관 사용자가 선호도를 평가한 아이템들
     → Score of Item[k]
 Output: 연관 사용자의 속성 → MainAttributeID
        AttributeSum[Num_Attribute] ← 0,
        AttributeCount[Num_Attribute] ← 0
  for k is items that user rated do
    for c is the Attribute of Item[k] do
      AttributeSum[c] ← AttributeSum[c] + Score of Item[k]
      AttributeCount[c]++
    endfor
  endfor
  for j=1 to Num_Attribute do
    MainAttributeID
      ← Max(MainAttributeID, AttributeSum[j]/AttributeCount[j])
  endfor // 사용자의 속성을 결정
  Representative Attribute[MainAttributeID]← Add UserID
return
```

알고리즘 1은 연관 사용자의 속성정보를 추출하는 방법이다. 연관 사용자가 선호도를 평가한 아이템을 이용하여 사용자의 속성정보를 구한다. 속성정보는 장르별 아이템의 선호도 합을 구한 후 선호도 합의 평균이 가장 큰 장르이다. 연관 사용자의 속성정보를 추출하는 이유는 실험 데이터로 쓰이는 MovieLens 평가 데이터[8]에서 아이템에 대한 속성정보가 하나 이상인 것이 많아서 사용자의 관점에 따라 속성정보가 달라지기 때문이다.

4.2. 속성정보를 이용한 연관 사용자 군집

본 논문에서는 ARHP 알고리즘에 의한 연관 사용자 군집에서 사용자의 속성정보를 추출한다. 여기에 협력적 필터링에서 연관 사용자를 선택하기 위하여 성별과 나이에 의한 속성정보를 사용한다. 연관 사용자 군집은 같은 성별 또는 같은 나이를 가진 사람들이 각각의 아이템에 대해서 유사한 선호도를 가진다고 가정한다. 성별과 나이를 연관 사용자 군집에 적용한 이유는 남성과 여성간의 성별 차이와 세대차를 통해서 추천의 정확도를 높이기 위함이다.

알고리즘 2. 속성정보를 이용한 연관 사용자 군집

```
Algorithm Associative User Group using Attribute Information
 Input: Num_class ← # of associative user in AttributeID
     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
```

알고리즘 2는 속성정보를 이용한 연관 사용자 군집 방법이다. 이는 협력적 필터링에서 아이템의 예측에 사용될 연관 사용자의 수를 결정하기 위해서 사용된다. 개인화 아이템 추천 시스템에서 실시간 예측을 하기 위해서 속성정보를 이용한 연관 사용자 군집은 적절한 연관 사용자의 수를 결정해야 한다. 실험을 통해서 적절한 연관 사용자의 수를 결정하기 위해서 사용자의 수를 증가시킴에 따라 정확도를 비교 평가하였다. 그림 1은 연관 사용자의 수에 따른 예측의 정확도이다.

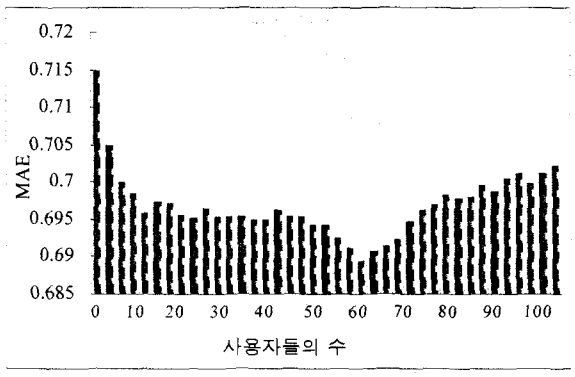

그림 1. 사용자들의 수에 따른 예측의 정확도

그림 1에서 MAE의 정확도를 보면 연관 사용자의 수가 증가함에 따라 예측의 정확도가 일관성 있게 좋아지지 않는다. 대략 연관 사용자의 수가 69 정도에 해당되는 곳에서부터 정확도가 감소되는 것을 볼 수 있다. 연관 사용자의 수를 결정하는 실험은 협력적 필터링에서 사용하는 피어슨 상관계수에 의한 사용자 유사도 가중치를 구하는 부분에 적용한 것이다.

5. 성능 평가

속성정보를 이용한 연관 사용자 군집의 성능 평가를 위해 실험 데이터로는 Grouplens Research Center의 MovieLens 평가 데이터를 사용하였다. MovieLens 평가 데이터 집합[8]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 간격

으로 아이템에 대하여 평가를 하였다.

실험을 위해 속성정보를 이용한 연관 사용자 군집은 협력적 필터링에서 기존의 이웃 선정에 사용되었던 Nearest Neighbor Model[2]과 K-Means 군집[2,4]의 결과와 비교하였다. 이를 위해 내용 정보 데이터베이스의 사용자 469명을 대상으로 연관 사용자 군집을 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 (1)과 같이 정의한다.

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

식 (1)에서 P는 정확도, R은 재현율을 의미하며, F-measure의 값이 클수록 분류가 우수함을 의미한다. 본 실험에서는 사용자의 수에 따라 그룹별로 F-measure의 분류 결과를 분석해 보았다. 속성정보를 이용한 연관 사용자 군집은 AUG-AI, Nearest Neighbor Model은 NNM, K-Means 군집 방법은 K-MC으로 표기하였다.

식 (1)을 이용한 그림 2는 정확도의 성능 곡선을 나타내고, 그림 3은 F-measure의 성능 곡선을 나타낸다. 그림 2에서 속성정보를 이용한 연관 사용자 군집은 K-MC 방법보다는 22.33%, NNM 방법보다는 5.09%의 높은 정확도를 나타낸다. 그림 3에서 속성정보를 이용한 연관 사용자 군집(AUG-AI)이 K-MC 방법보다는 15.44%, NNM 방법보다는 4.1% 향상된 F-measure의 결과를 나타낸다.

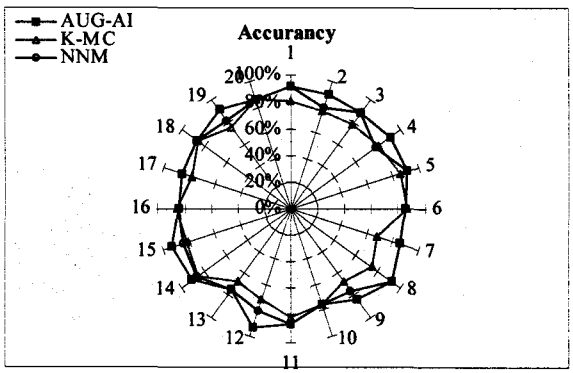

그림 2. 사용자 군집의 정확도에 의한 성능 평가

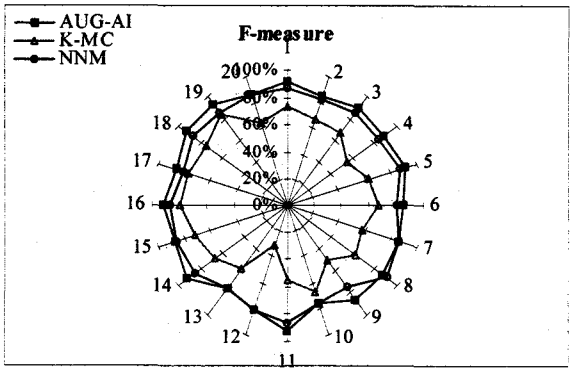

그림 3. F-measure에 의한 그룹별 성능 평가

그림 4는 사용자의 수에 따른 F-measure의 성능 변화를 나타낸다. 세 가지 방법에 대해서 사용자의 수가 증가함에 따라 점차 F-measure의 성능이 점차 향상됨을 보인다. 특히, 속성정보를 이용한 연관 사용자 군집(AUG-AI)은 사용자의 수가 적은 경우에도 높은 성능을 나타낸다. 그러나 K-MC 방법과 NNM 방법은 사용자 수가 적은 경우 낮은 성능을 나타낸다. 전체적으로 속성정보를 이용한 연관 사용자 군집이 K-MC 방법과 NNM 방법보다 성능이 우수함을 알 수 있다.

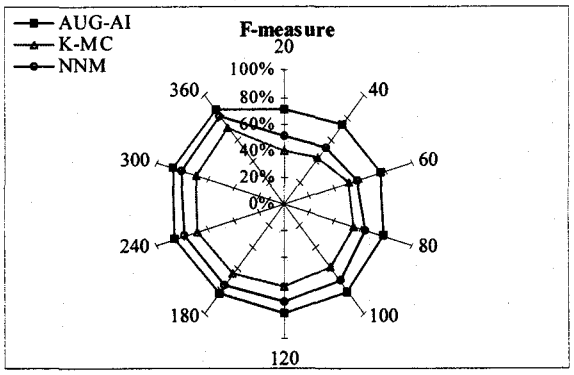

그림 4. 사용자의 수에 따른 F-measure의 성능 평가

6. 결 론

협력적 필터링에서 사용자 선정에 사용되었던 기존의 방법들은 정보에 대한 선호도의 정도만을 반영하여 사용자의 수를 결정하는데 사용하였다. 이는 전체 선호도 정보들을 모두 사용하여 유사도 가중치를 구하는 것이므로 속성정보의 값에 대해 사용자가 차별적인 선호도를 가지는 경우 이를 제대로 반영하지 못하는 단점이 있다. 이를 보완하기 위해서 본 논문에서 속성정보를 이용한 연관 사용자 군집을 제안하였다. 속성정보에 대한 사용자의 선호도를 고려하지 않은 이러한 문제점을 개선하기 위하여 연관 사용자 군집을 사용하고 그 속성에 대한 연관 사용자의 선호도를 협력적 필터링에 반영하였다. 여기서 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] 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.

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

[3] T. H. Kim and S. B. Yang, "An Effective Recommendation Algorithm for Clustering-Based Recommender Systems," Proc. of the Conference on Artificial Intelligence, pp. 115-1153, 2005

[4] M. O. Connor and J. Herlocker, "Clustering Items for

Collaborative Filtering," Proc. of the ACM SIGIR Workshop on Recommender Systems, Berkeley, CA, 1999.1.

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

[6] 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.

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

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