연관관계 군집 분할 방법을 이용한 아이템 필터링 시스템

Item Filtering System Using Associative Relation Clustering Split Method

조동주*, 정경용*, 박양재**
상지대학교 컴퓨터정보공학부*, 가천의과학대학교 의료공학부 정보기술학과**

Dong-Ju Cho(queen8181@sangji.ac.kr)*, Kyung-Yong Jung(kyjung@sangji.ac.kr)*,
Yang-Jae Park(yjpark@gachon.ac.kr)**

요약

전자상거래에서 많은 아이템 중에 사용자에게 적합한 아이템을 추천하기 위해서는 많은 시간과 노력이 소요된다. 그러므로 추천 시스템이 사용자들을 대신하여 적합한 아이템을 추천해줄 수 있다면 만족을 얻을 수 있다. 본 논문에서는 정확성과 확장성을 향상시키기 위해서 협력적 필터링에서 연관관계 군집 분할 방법을 제안하였다. 평가한 데이터를 사용하여 연관 아이템간의 향상도를 산출하고 연관관계 군집의 효율성을 높이기 위해서 아이템으로 구성된 노드 군집을 분할하였다. 이는 군집들 중 하나의 아이템만이 연관성을 달리하고, 나머지 아이템들은 군집의 연관성이 충족되어진다면 결합하는 방법이다. 성능을 평가하기 위해서 MovieLens 데이터 집합에서 K-means와 EM에 의한 군집과 비교 평가하였다.

■ 중심어 : | 협력적 필터링 | 군집 | 연관규칙 | 추천 시스템 | 전자상거래 |

Abstract

In electronic commerce, it is important for users to recommend the proper item among large item sets with saving time and effort. Therefore, if the recommendation system can be recommended the suitable item, we will gain a good satisfaction to the user. In this paper, we proposed the associative relation clustering split method in the collaborative filtering in order to perform the accuracy and the scalability. We produce the lift between associative items using the ratings data. and then split the node group that consists of the item to improve an efficiency of the associative relation cluster. This method differs the association about the items of groups. If the association of groups is filled, the reminding items combine. To estimate the performance, the suggested method is compared with the K-means and EM in the MovieLens data set.

■ keyword : | Collaborative Filtering | Clustering | Association Rule | Recommendation System | Electronic Commerce |

1. 서 론

오늘날 웹의 성장과 확산으로 인해 전자상거래 기업이나 개인은 성향에 따라 선택하는 것보다는 자동적으로 추천받는 것을 선호하고 있다. 이는 근래의 인터넷 확산과 보급의 영향으로 정보의 양이 많아지면서 전자상거래를 이용하는 사용자들이 수많은 정보들 속에 있기 때문이다. 전자상거래에서 추천시스템은 인터넷을 통한 일대일 고객 관리와 개인화 서비스를 가능하게 하고 또한 추천시스템의 개인화 기법으로 가장 널리 이용

접수번호 : #061213-001
접수일자 : 2006년 12월 13일
심사완료일 : 2007년 03월 21일
교신저자 : 정경용, e-mail : kyjung@sangji.ac.kr

되고 있는 기술이 협력적 필터링이다[1]. 최근 전자상거래에서 추천시스템은 방대한 정보의 사용자와 아이템을 바탕으로 {사용자-아이템} 행렬에서 희박성 문제와 확장성 문제를 해결하려는 연구가 진행되고 있다. 대부분의 상업적 추천 시스템들이 대용량 아이템 집합들을 계산하지만 실제로 사용자들이 평가하거나 구매하는 아이템은 수백만 개의 아이템들 중 1% 미만에 불과하다. 따라서 같은 아이템을 구매하거나 평가한 사용자들의 데이터가 희박하며, 실제 아이템을 추천할 수 없는 경우가 많다. 또한 확장성 문제는 수백만 건의 사용자와 아이템의 데이터 처리를 한다는 점에서 실시간으로 아이템을 추천해야 하는 웹 기반 추천 시스템에게 상당한 부담감으로 작용한다. 본 논문에서는 {사용자-아이템} 행렬에서 아이템들 간의 연관관계를 유지하면서 행렬의 차원 수를 감소시키는 것을 이용하여 희박성 문제와 시간의 복잡도를 해결하기 위해 협력적 필터링에서 연관관계 군집 분할 방법을 제안하였다. 그리고 실험을 통해서 기존의 협력적 필터링과 제안한 방법과 예측의 정확도를 비교 평가하였다.

2. 관련연구

2.1 K-Means 군집

K-Means 군집은 데이터 분류에 있어 Maximum Likelihood(ML)방법의 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.

K-Means 군집을 이용하여 사용자를 군집하는 과정은 3단계로 구성한다. 첫 번째 단계에서는 군집의 개수 K와 중심들을 초기화한다. 두 번째 단계에서는 사용자 간의 유사도를 기반으로 사용자의 소속을 구한다. 세 번째 단계에서는 소속이 결정된 사용자들을 판별하기 위하여 유사도 평균의 변화치가 임계값보다 낮으면 종료한다[2].

2.2 아이템 기반 협력적 필터링

협력적 필터링 기술은 사용자 기반의 협력적 필터링과 아이템 기반의 협력적 필터링으로 나눈다. 사용자 기반의 협력적 필터링은 추천의 대상이 되는 고객에 대하여 그와 비슷한 취향을 갖는 유사 고객들을 찾고 이들 유사 고객들이 공통적으로 많이 구매하는 아이템들 중에서 추천 대상 고객이 구매하지 않은 아이템을 추천해 주는 방법이다. 아이템 기반의 협력적 필터링은 특정 고객에 대한 추천 시에 사용하므로 사용자기반의 협력적 필터링 보다 효율적이다. 또한 다양한 사용자의 관심분야를 반영하므로 추천 정확도를 높인다. 여기서 아이템간의 유사도를 계산하기 위한 기존 연구로는 코사인 기반의 유사도, 상관계수 기반의 유사도, 개선된 코사인 기반의 유사도가 있다[3].

2.3 Clique 기반의 하이퍼그래프 군집

[그림 1]은 하향식 순회방법을 사용한 Clique 기반의 하이퍼그래프 군집 방법이다.

```
Input : Equivalence Class group → EC[i]
Output : pCluster[]
 1. while EC[i] in Each EC
 2.      pNode[] = #{Src, Dsc[j]} in EC[i]
 3.      TreeRecursion(pNode)
 4. Assign(pCluster[]);
 5. Function TreeRecursion(pBtNode[])
 6.    L = #Count in pBtNode
 7.    for i=0 to N < L-1
 8.        k=0
 9.      for j=i+1 to N < L
10.       if possible to combination {pBtNode[i] and pBtNode[j]}
11.           pNode[k++]
                 = combination of pBtNode[i] and pBtNode[j]
12.       else
13.           pCluster[] = pBtNode[i] and pBtNode[j]
14.           if k>0
15.                  TreeRecursion(pNode)
16.                   pCluster[] = #(pBtNode[L-1]
17. return
```

그림 1. Clique 기반의 하이퍼그래프 군집 방법

[그림 1]의 알고리즘은 연관관계가 있는 아이템들을 결합하고 검증하여 군집하는 방법이다. 군집을 형성한 후 군집들 간의 연관 관계가 존재하면 포함된 작은 군집을 삭제함으로 차원의 수를 줄인다.

Clique 기반의 하이퍼그래프 군집은 트리 형식으로 구성된 아이템을 탐색하여 연관관계가 높은 아이템들을 기반으로 군집을 생성하는 방법이다. 순회방법으로는 하향식 순회방법, 상향식 순회방법, 두 가지 방법을 조합한 순회방법이 있다. 두 가지를 조합한 순회방법이 성능이 우수하다고 평가되나 효율성이 떨어지고 연결 노드간의 순서를 가지고 있어야 하는 단점이 있다[5]. 본 논문에서는 하향식 순회방법으로 Clique 기반의 하이퍼그래프 군집을 사용하였다.

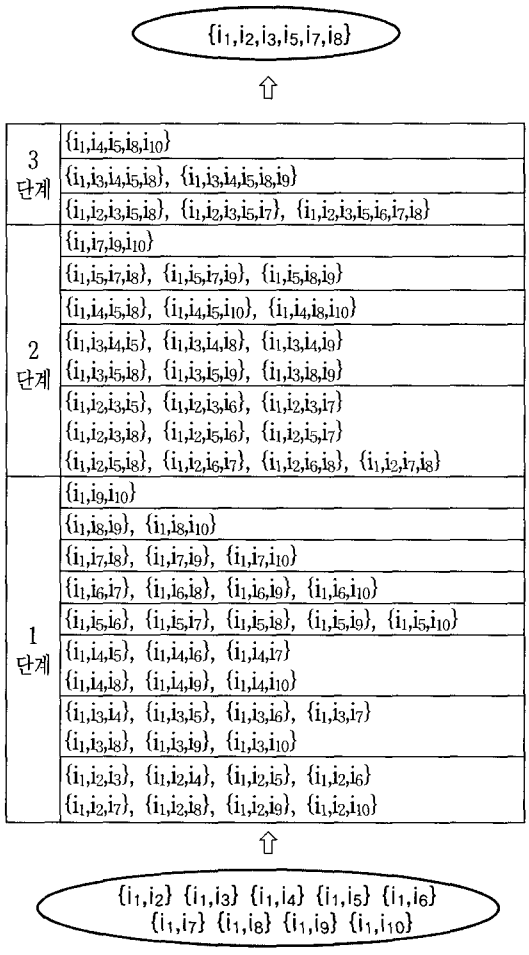

그림 2. 상향식 순회방법 적용의 예

[그림 2]는 상향식 순회방법을 이용하여 아이템 $\{i_1,i_2,i_3,i_4,i_5,i_6,i_7,i_8,i_9,i_{10}\}$을 군집하는 과정을 나타낸다. 여기서 연관성이 높은 아이템 집합을 $\{\{i_1,i_2\}, \{i_1,i_3\}, \{i_1,i_4\}, \{i_1,i_5\}, \{i_1,i_6\}, \{i_1,i_7\}, \{i_1,i_8\}, \{i_1,i_9\}, \{i_1,i_{10}\}, \{i_2,i_3\}, \{i_2,i_5\}, \{i_2,i_6\}, \{i_2,i_7\}, \{i_2,i_8\}, \{i_3,i_4\}, \{i_3,i_5\}, \{i_3,i_8\}, \{i_3,i_9\}, \{i_4,i_5\}, \{i_4,i_8\}, \{i_4,i_{10}\}, \{i_5,i_7\}, \{i_5,i_8\}, \{i_5,i_9\}, \{i_6,i_8\}, \{i_6,i_9\}, \{i_7,i_9\}, \{i_7,i_{10}\}, \{i_8,i_{10}\}, \{i_9,i_{10}\}\}$이라고 가정한다. [그림 2]는 연관관계가 높은 아이템들을 기본 노드로 정한 후 단계가 올라갈수록 아이템들이 군집되는 것을 보인다. 여기서 연관성이 없는 아이템들은 삭제하고, 연관성이 높은 아이템들의 단계를 진행시킨다. 생성된 군집들은 아이템들 간의 노드 위치가 동일하며 군집들 중 하나의 아이템만 연관성을 달리하고, 나머지 아이템들의 군집 연관성이 충족되면 결합하는 방식이다. 마지막 단계에서는 연관성이 높은 아이템들을 결합하여 군집을 형성한다. [그림 2]의 마지막 단계에서 생성된 군집은 $\{\{i_1,i_2,i_3,i_5,i_8\}, \{i_1,i_2,i_3,i_5,i_7\}, \{i_1,i_2,i_3,i_5,i_6,i_7,i_8\}, \{i_1,i_3,i_4,i_5,i_8\}\ \{i_1,i_3,i_4,i_5,i_8,i_9\}, \{i_1,i_4,i_5,i_8,i_{10}\}\}$이다. 여기서 가장 높은 연관관계에 있는 그룹을 최종적으로 결합하면 $\{i_1,i_2,i_3,i_5,i_7,i_8\}$라는 군집 결과를 얻을 수 있다.

3. 연관관계 군집 분할 방법

본 논문에서는 연관관계 군집 분할 방법을 이용한 아이템 필터링 시스템을 제안하였다. 기존 연구인 Clique 기반의 하이퍼그래프 분할방법은 방대한 양의 데이터에서는 성능이 낮은 단점이 있으며, 상향식 순회방법을 사용하므로 시간의 복잡도가 O(n!)이다[4]. 이러한 문제점을 해결하기 위해 본장에서는 제안한 연관관계 군집 분할 방법을 기술한다.

3.1 연관관계 군집 분할 방법

Clique 기반의 하이퍼그래프 분할방법은 아이템의 수를 증가시키면서 군집하므로 방대한 양의 데이터에서는 성능이 매우 낮은 단점이 있다. 이를 개선하기 위해 본 연구에서는 전체 군집 중 하나의 군집에서부터 하향 단계의 작은 군집을 분리해 진행하는 연관관계 군집 분할 방법을 사용한다. 상위 단계에 있는 노드에서 분리된 소규모 군집들 사이에서 연관성을 파악한 후 연관관

계가 성립하는 군집들을 차례대로 삭제해 나간다. 삭제한 전체 노드를 이용하여 최종적으로 분할된 군집의 연관관계를 검증하여 적용한다. 본 연구에서 제안한 방법의 시간 복잡도는 O(n3)이다[4]. [그림 3]은 본 논문에서 제안한 연관관계 군집 분할 방법이다.

```
Input : Delete node class group → DC[i]
        Equivalence Class group → EC[i]
Output : Split cluster group → SC[i]
 1. for i=0 to N ⟨ #EC group
 2.   pNode[0] = Add Unique Item in EC[i]
 3.   for j=0 to N ⟨#DC group
 4.     for k=0 to N ⟨ #pNode
 5.       if pNode[k] ⊃ DC[j]
 6.         Split pNode[k]
 7.         for k=0 to N ⟨ #Split pNode
 8.           for l=0 to N ⟨ #Unsplited pNode
 9.             if pNode[l] ⊃ pNode[k]
10.               Delete pNode[k]
11.       SC = All pNode
12.       Clear pNode
13. return
```

그림 3. 연관관계 군집 분할 방법

연관관계 군집 분할 방법에서는 Clique 기반의 하이퍼그래프 군집의 상향식 순회방법에서 사용했던 10개의 노드를 이용하여 제안한 군집 분할방법을 사용하였다. 이 과정에서는 단계가 하향으로 내려가면서 각각의 단계 사이에 제시한 향상도가 낮은 노드를 적용하고 그 노드가 포함된 군집은 분리시켜 나간다. 여기서 분리된 군집은 분리되지 않은 군집들과 연관성을 비교한다. 분할되지 않은 군집에 속하면 삭제하는 방식으로 향상도가 낮은 노드를 제안한 방법에 적용하면 {{i1,i2,i3,i5,i8}, {i1,i2,i3,i5,i7},{i1,i2,i3,i5,i6,i7,i8},{i1,i3,i4,i5,i8} {i1,i3,i4,i5,i8,i9},{i1,i4,i5,i8,i10}} 이라는 군집들이 생긴다. 군집의 아이템들은 [그림 2]에서 제시한 Clique 기반의 하이퍼그래프 군집의 상향식 순회방법을 적용한 결과와 동일하나 제안한 연관관계 군집 분할 방법의 시간 복잡도면에서는 성능이 향상됨을 알 수 있다.

3.2 연관군집에서 아이템간의 유사도

연관성이 높은 아이템들 간의 군집이 생성되면 아이템 기반의 협력적 필터링에서 아이템을 예측을 하기 위하여 연관 군집내에서 아이템간의 유사도를 계산한다. 아이템간의 유사도는 피어슨 상관계수를 사용한다[3].

$$sim(x,y) = \frac{\sum_u (P_{x,u} - \overline{P_x})(P_{y,u} - \overline{P_y})}{\sqrt{\sum_u (P_{x,u} - \overline{P_x})^2 \sum_u (P_{y,u} - \overline{P_y})^2}} \quad \text{(식 1)}$$

(식 1)는 피어슨 상관계수를 이용해서 아이템간의 유사도를 계산하는 식이다. 여기서 $P_{x,u}$는 아이템 x가 사용자 u에 대해서 평가한 선호도이고, $\overline{P_x}$는 아이템 x에 대해서 평가한 선호도의 평균값이다. u는 아이템 x와 y의 선호도를 공통으로 평가한 사용자이다. (식 1)에 의해서 군집내의 아이템간의 유사도를 계산되면 군집간의 가중치를 계산할 수 있다. 본 연구에서는 군집간의 가중치를 계산할 때 고려할 요인으로 분산 가중치를 적용하였다. 분산 가중치는 영향력이 큰 아이템에 대해서 가중치를 다르게 부여하기 위한 강도이다. 분산 가중치를 아이템간의 유사도에 적용하면 (식 2)와 같다.

$$W(u,i) = \frac{\sum_j Var_j \times z_{a,j} \times z_{i,j}}{\sum_j Var_j} \quad \text{(식 2)}$$

$$Var_j = \frac{\left(\frac{\sum_{k=1}^{n}(v_{k,j} - \overline{v_j})^2}{n-1}\right)_j - \left(\frac{\sum_{k=1}^{n}(v_{k,j} - \overline{v_j})^2}{n-1}\right)_{\min}}{\left(\frac{\sum_{k=1}^{n}(v_{k,j} - \overline{v_j})^2}{n-1}\right)_{\max}}$$

(식 2)에서 $z_{a,i}$는 Var_j를 평균 0, 표준편차 1로 변환한 값이다. n은 아이템 i에 선호도를 평가한 사용자의 수이다. max는 모든 아이템에 대한 분산 중 최대이고, min은 모든 아이템에 대한 최소 분산이다[6].

3.3 아이템 기반의 협력적 필터링에서 예측

연관군집에서 아이템들을 분리 군집한 후 군집속의 아이템에 대하여 유사도와 분산 가중치를 계산하여 아이템의 선호도를 예측한다. 이는 아이템간의 유사도를 계산할 때 평균만 이용하는 것이 아니라 분산 가중치를 적용했기 때문에 더욱더 정확한 예측을 할 수 있다. 비중을 다르게 하는 아이템에 대하여 평균을 계산할 때 단순히 평균만으로는 성능이 낮으므로 비중에 따른 아

이템에 맞는 중요도를 결정하고 적용하여 평균을 계산한다. 기본적으로, 특정 아이템과 가까운 아이템이 선택되면 예측을 위해 아이템에 대한 선호도를 같은 분포를 따르도록 변환하고 조합한다. 모든 아이템의 선호도를 사용하는 것이 기본이지만 가중치를 적용하여 선호도를 유사도 기준의 가중평균으로 계산하는 방법이 많이 쓰인다. 그리하여 모든 선호도가 근사적으로 같은 분포를 따른다는 가정을 기본으로 하게 된다. 이 같은 방법으로 특정 아이템에 대한 유사도 평균과 그 유사도를 가중 평균하여 아이템의 선호도를 예측하는 것이다. 또한 모든 아이템의 유사도 값은 동일한 분포와 동일한 분산을 가진다고 가정한다. 예측할 아이템이 속한 군집에 대해서 아이템들 간의 가중치를 적용하여 선호도의 가중치 평균을 (식 3)과 같이 계산한다.

$$P_{u,i} = \overline{V_n} + \frac{\sum_{d}^{D} w(u,i) \sum_{k=1}^{D} sim(x,y)(v_{k,i} - \overline{v_k})}{\sum_{d}^{D} w(u,i) \sum_{k=1}^{D} sim(x,y)} \quad \text{(식 3)}$$

$P_{u,i}$는 아이템 i에 대한 사용자 u의 선호도 예측값이고 $\overline{V_u}$는 사용자 u의 가중치가 부여된 선호도의 평균이다. D는 연관 군집내의 아이템의 수를 나타낸 것이고 $w(u,i)$는 분산가중치를 아이템간의 유사도 가중 평균을 나타낸다. $sim(x,y)$는 아이템 x와 아이템 y간에 피어슨 상관계수를 적용한 유사도이다.

4. 성능평가

성능 평가를 하기 위한 실험 데이터로는 Grouplens 연구 기관에서 제공받은 MovieLens 데이터[7] 집합을 사용하였다. 사용자들은 0.0에서부터 1.0까지 0.2간격으로 아이템에 대하여 평가하였으며 또한 사용자가 실제로 영화를 보았는지의 여부를 알 수 있는 가중치 정보가 존재한다. 1.0의 선호도는 아이템에 대해서 긍정적인 면을 의미하고 0.0의 선호도는 부정적인 면을 의미한다. 아이템의 장르는 액션, 애니메이션, 외국 예술, 고전, 코미디, 드라마, 가족, 공포, 로맨스, 스릴러의 10 가지로 구분되어 있다. 여기서 장르는 중복을 허용한다. [그림 4]는 1,612개의 아이템에 대한 장르 정보를 포함하는 MovieLens 데이터의 일부분이다.

ID	Name	Action	Animation	Art_Foreign	Classic	Comedy	Drama	Family	Horror	Romance	Thriller
1	Toy Story	☐	☑	☐	☐	☐	☐	☑	☐	☑	☐
2	Jumanji	☐	☐	☐	☐	☐	☐	☑	☐	☐	☐
3	Grumpier Old Men	☐	☐	☐	☐	☑	☐	☐	☐	☐	☐
4	Waiting to Exhale	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐
5	Father of the Bride Part II	☐	☐	☐	☐	☑	☐	☐	☐	☐	☐
6	Heat	☑	☐	☐	☐	☐	☐	☐	☐	☐	☑
7	Sabrina	☐	☐	☐	☐	☑	☐	☐	☐	☑	☐
8	Tom and Huck (Tom Sawyer)	☐	☐	☐	☐	☐	☐	☑	☐	☐	☐
9	Sudden Death	☑	☐	☐	☐	☐	☐	☐	☐	☐	☐
10	Goldeneye	☑	☐	☐	☐	☐	☐	☐	☐	☐	☐
11	The American President	☐	☐	☐	☐	☑	☐	☐	☐	☑	☐
12	Dracula: Dead and Loving It	☐	☐	☐	☐	☑	☐	☐	☐	☐	☐
13	Balto	☐	☑	☐	☐	☐	☑	☑	☐	☐	☐
14	Nixon	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐
15	Cutthroat Island	☑	☐	☐	☐	☐	☐	☐	☐	☐	☐
16	Casino	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐
17	Sense and Sensibility	☐	☐	☑	☐	☐	☐	☐	☐	☑	☐
18	Four Rooms	☐	☐	☐	☐	☑	☐	☐	☐	☐	☐
19	Ace Ventura: When Nature Calls	☐	☐	☐	☐	☑	☐	☐	☐	☐	☐
20	Money Train	☑	☐	☐	☐	☐	☐	☐	☐	☐	☐
21	Get Shorty	☑	☐	☐	☐	☑	☐	☐	☐	☐	☐
22	Copycat	☐	☐	☐	☐	☐	☐	☐	☑	☐	☐
23	Assassins	☐	☐	☐	☐	☐	☐	☐	☐	☐	☑

그림 4. 아이템별 장르 정보

4.1 실험 방법 및 결과

본 논문에서 제안한 아이템 기반의 협력적 필터링에서의 선호도 예측방법은 Visual C++ 6.0으로 구현되었으며 실험 환경은 Pentium Ⅳ, 1.9GHz, 256MB RAM 환경에서 수행되었다. 실험 방법은 연관관계 군집에서 기존 연구인 Clique 기반의 하이퍼그래프 군집과 제안한 연관관계 군집 분할 방법과 실험을 통해서 비교 평가하였고, 기존의 피어슨 상관계수 기반의 협력적 필터링 방법(IbCF)과 제안한 방법을 협력적 필터링에 적용하여 실험(ARCS-CF)하였다. 본 연구에서 성능 평가하기 위한 실험 데이터는 MovieLens 데이터 집합을 전처리하여 30,861명의 사용자와 1,612종류의 영화에 대해서 실험을 진행하였다. 또한, 사용자의 평가 데이터가 1,428,362개이다. 모든 데이터를 적용하기에 성능 평가가 어려워 데이터를 분할하여 실험을 하였다. 본 연구에서는 아이템을 100개씩 10개의 그룹으로 나누어 실험을 하였다. 10개의 트랜잭션, 100개의 아이템, 1395k의 데이터 로그라는 의미에서 T10I100D1395k를 사용하였다[4]. [표 1]은 기존 연구인 Clique 기반의 하이퍼그래프 군집과 연관관계 군집 분할 방법의 실행 시간(초)을 통하여 비교 평가하였으며. 아이템의 개수가 증가 할수록 연관관계 군집 분할 방법의 성능이 실행 시간 면에서 우수하다는 결과를 얻을 수 있었다.

표 1. 제안한 방법과 기존 연구와 실행시간 비교

	아이템 수									
	10	20	30	40	50	60	70	80	90	100
Clique기반의 하이퍼그래프군집방법	1	2	2	6	11	21	52	63	77	94
제안한 연관관계 군집분할방법	1	1	2	3	3	5	11	18	24	30

4.2 분석 및 성능평가

예측 알고리즘을 평가하는 여러 가지 방법 중에서 예측 값과 실제 값의 차이를 표시하는 MAE 방식을 사용하여 성능 평가하였다[8][9]. (식 4)에서 제시한 MAE는 예측의 정확성을 판단하는데 가장 많이 쓰이는 방법이며, 실제 선호도 값과 예측된 선호도 값과의 오차로 정의되고 MAE는 오차들의 절대값 평균을 의미한다. MAE는 (식 4)과 같이 구할 수 있다.

$$|E| = \frac{\sum_{i=0}^{N} |\epsilon_i|}{N} \qquad \text{(식 4)}$$

여기서 N은 총 예측 횟수이고, ϵ_i는 예측 값과 실제 값의 오차를 나타내며 i는 각 예측 단계를 나타낸다. (식 4)에 실험 데이터 집합의 사용자 수를 증가시킴으로서 성능평가를 한다.

표 2. 사용자 수에 따른 MAE 성능 평가

사용자 수	MAE	
	ARCS-CF	IbCF
1	1.40	1.42
180	1.27	1.29
400	1.11	1.20
860	0.80	1.00
1,490	0.61	0.81
2,460	0.51	0.74
2,734	0.52	0.80
3,400	0.57	0.83
3,680	0.60	0.90
4,010	0.61	0.92
4,380	0.64	0.96
5,048	0.65	0.99
5,720	0.67	1.04

[표 2]는 MAE를 이용하여 협력적 필터링에 의한 피어슨 상관계수의 방식과 연관관계 군집 분할 방법을 실험하여 사용자 수에 따른 성능평가이다.

아이템에 대해서 평가한 사용자가 적을 경우 전반적으로 연관관계가 높게 나타나는 경향이 있다. 이는 평가한 사용자의 수가 적을수록 모든 아이템에 대해서 평가하기 때문에 영향력에 따른 가중치의 변별력이 없기 때문이다. 이러한 경우 아이템이 속한 군집의 크기는 매우 큰 반면 아이템이 속한 군집의 수는 매우 적어지게 된다. 따라서 평가한 사용자의 수가 많을수록 예측의 정확도는 높아지게 된다. 또한 많은 사용자들이 평가한 아이템에 대해서는 많은 군집에 속하게 되기 때문에 상대적으로 정확도가 낮아지는 경향이 있다.

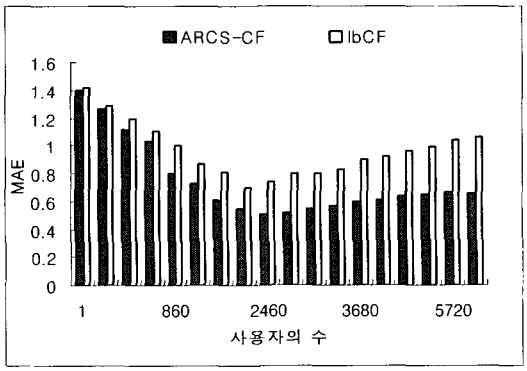

그림 5. ARCS-CF와 CF에서 사용자 수에 따른 MAE

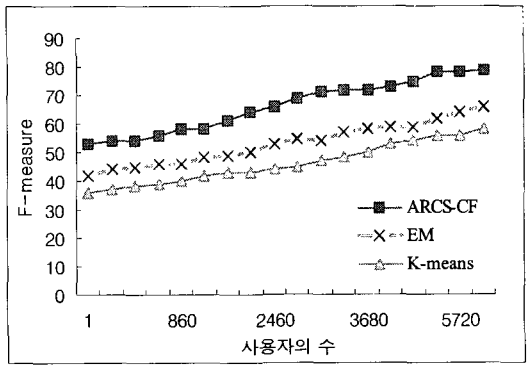

그림 6. 사용자의 수에 따른 F-measure의 성능평가

[그림 5]는 ARCS-CF와 IbCF에서 사용자의 수에 따른 MAE의 성능 평가이다. [그림 6]에서 평가한 사용자의 수가 0에서 2,500명인 지점에서 본 논문에서 제안한 알고리즘이 더 좋은 성능을 보임을 알 수 있다. 아이템

들이 군집 속에 들어가는 수를 제한하고 그 수보다 적을 경우 본 논문에서 제안한 방법을 사용하면 정확도 면에서 좋은 성능이 나타난다.

[그림 6]은 K-means와 EM을 이용하여 사용자 수의 증가에 따른 F-measure의 성능변화를 나타낸 것이다. EM[10]은 사용자를 군집하기 위한 편리한 방법이나 같은 군집에 속한 사용자들이 검색한 문서는 같은 클래스에 있어야 한다는 제한을 가지며 K-means는 군집 방법은 빠르나 정확도가 낮은 단점이 있다. 세 가지 방법의 사용자 수가 증가함에 따라 F-measure에 의한 성능이 점차 향상됨을 보이며 특히, ARCS-CF는 K-means보다 23.2%, EM보다는 26.7%의 높은 성능을 나타내고 EM과 K-means는 사용자수가 적은 경우에는 전체적으로 낮은 성능을 보인다.

5. 결 론

본 논문에서 제안한 연관관계 군집 분할 방법을 이용한 아이템 필터링 시스템은 아이템들을 군집하고 군집 내에 있는 아이템들에 대해서 평가한 사용자들의 선호도를 기반으로 아이템 선호도를 예측하였다. 또한 군집을 적용하여 아이템에 대한 평가가 적더라도 아이템이 속한 군집에 데이터가 많다는 점을 이용하여 협력적 필터링의 희박성 문제를 해결하였다. 연관관계 군집 분할 방법은 기존의 협력적 필터링의 단점을 보완하는 역할을 하며, 특정 사용자의 선호도에 대한 정보가 없더라도 아이템간의 연관관계에 의해 추천이 가능하다. 그리고 사용자에게 몇 가지 아이템만을 추천한다고 가정할 때 그 사용자가 평가하지 않은 모든 아이템에 대하여 예측을 하는 것이 아니라 사용자가 선호할만한 아이템 군집을 추출함으로서 사용 시간과 컴퓨팅 횟수를 감소할 수 있다. 아이템의 속성을 유사도에 반영하기 위해 적용하였던 가중치를 단순히 아이템의 분류 속성에 관한 유사관계가 아닌 복합적인 속성들을 반영하도록 한다면 실험의 일부분에서 나타난 가중치의 오류를 줄일 수 있으며, 정확성 측면에서 더욱 향상된 추천 시스템을 구현할 수 있을 것이다. 평가 데이터가 많아지면 군집의 개수가 엄청난 숫자로 늘어나게 되고, 결과적으로 이를 처리하는 연산 시간이 늘어나게 되는 단점은 향후 연구이다.

참 고 문 헌

[1] 이희정, 홍태호, "클러스터링 기반 사례기반추론을이용한 추천시스템 개발", 한국경영과학회 춘계학술대회논문집, pp.506-509, 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] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based Collaborative Filtering Recommendation Algorithms," Proc. of the 10th Int. Conf. on WWW. pp.285-295, 2001.

[4] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New Algorithms for Fast Discovery of Association Rules," Proc. of the 3rd IEEE Conference on Knowledge Discovery and Data Mining, pp.283-286, 1997.

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

[6] 이준규, *인터넷 개인화 아이템 추천 알고리즘에 대한 연구*, 연세대학교 석사학위논문, 2000.

[7] http://www.cs.umn.edu/research/GroupLens/

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

[9] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl, "An Algorithm Framework for Performing Collaborative Filtering," Proc. of the Int. Conf. on Management of Data, 1999.

[10] G. J. McLachlan and T. Krishnan, *The EM*

Algorithm and Extensions, New York: John Wiley and Sons, 1997.

저 자 소 개

조 동 주(Dong-Ju Cho) 준회원

- 2004년 2월 : 상지대학교 행정학과 (행정학사)
- 2007년 3월 ~ 현재 : 상지대학교 컴퓨터정보공학과 석사과정

<관심분야> : 지능시스템, 시맨틱 웹, 인공지능

정 경 용(Kyung-Yong Jung) 정회원

- 2000년 2월 : 인하대학교 전자계 산공학과(공학사)
- 2002년 2월 : 인하대학교 컴퓨터 정보공학과(공학석사)
- 2005년 2월 : 인하대학교 컴퓨터 정보공학과(공학박사)
- 2006년 3월 ~ 현재 : 상지대학교 컴퓨터정보공학부 교수

<관심분야> : 데이터마이닝, 지능시스템, 인공지능

박 양 재(Yang-Jae Park) 정회원

- 1983년 : 인하대학교 공과대학 전자공학과(공학사)
- 1990년 : 인하대학교 정보공학과(공학석사)
- 2003년 : 인하대학교 전자계산공학과(공학박사)
- 1993년 ~ 현재 : 가천의과학대학교 의료공학부 정보기술학과 유비쿼터스 컴퓨팅전공교수

<관심분야> : 이동 컴퓨팅, U-헬스케어, HCI