연관 관계 군집에 의한 협력적 여과 방법

김진현[0] 정경용 김태용 이정현
인하대학교 전자계산공학과
(ador[0], dragon, tykim)@nlsun.inha.ac.kr, jhlee@inha.ac.kr

A New Collaborative Filtering Using Associative Relation Clustering

Jin-Hyun Kim[0] Kyung-Yong Jung Tae-Yong Kim Jung-Hyun Lee
Dept. of Computer Science & Engineering, Inha University

요 약

협력적 여과 방법은 사용자의 평가 데이터를 이용하므로, 항상 초기 평가 문제(First-Rating Problem)와 희박성 문제(Sparsity Problem)가 발생한다. 최근 이러한 문제를 해결하기 위해 많은 연구가 진행되고 있는데, 본 논문에서는 연관 규칙을 이용하여 이러한 문제를 해결하고자 한다. 사용자의 평가 데이터를 이용하여 아이템간의 연관성을 산출하고, 연관성이 높은 아이템끼리 군집한다. 사용자와 군집간에 피어슨 상관 계수(Pearson Correlation Coefficient)를 이용하여 가중치를 구하고, 이것으로 선호도를 예측한다. 이러한 방법을 기존의 협력적 여과 방법과 함께 속성에 의한 군집 방식과 비교 평가하였다. 또한, 효율적인 군집을 위한 Split Cluster Method를 제안하고, 기존의 트리 방식의 군집과 비교 평가하였다.

1. 서론

유사한 아이템들을 군집(Clustering)하여 성능을 향상시키는 방법은 정보 검색 분야에서 많이 쓰인다. 기존의 협력적 여과 방법에서도 군집을 이용한 방법을 사용하였다. 사용자들의 경우, 프로파일(Profile)이나 나이, 성별 등의 속성을 이용하여 군집 하였고, 아이템들의 경우는 아이템간의 유사도가 높은 것끼리 군집하거나, 내용기반 여과(Context-based Filtering)를 병합하여 비슷한 내용을 군집하였다. 또한, 선호도에 가장 영향을 미치는 대표 속성을 이용하여 군집을 하였다[1][2][3].

군집의 이점은 아이템에 대한 평가 데이터가 적더라도 그 아이템이 속한 군집에 대한 평가 데이터는 많다는 점이다. 본 논문에서는 연관 규칙을 협력적 필터링 시스템에 적용할 경우, 평가 데이터가 적은 아이템에 대해 다른 아이템과의 향상성(Lift)이 높다는 것을 발견하였다. 이를 이용하면 위 아이템은 많은 군집에 속하게 되므로 희박성 문제를 해결할 수 있다. 제안한 방법의 성능을 평가하기 위해서 기존의 협력적 여과 시스템과 대표속성을 이용한 군집에 의한 협력적 여과 방법과 비교 평가하였다.

2. 관련 연구

2.1 협력적 필터링

협력적 필터링은 사용자와 유사한 선호도를 가지는 이웃을 찾아내고 사용자간에 선호도를 표시한 아이템의 선호도를 예측하기 위해서 사용된다. 대표적인 유사도 기준 값으로는 Correlation, Vector based similarity 등이 있다[4]. 이러한 방법을 응용하여 아이템 간의 유사도를 측정하는데도 이용되는데, Correlation-based Similarity, Cosine-based Similarity 등이 사용된다[2].

피어슨 상관 계수는 협력적 여과 방법에서 대표적으로 쓰이는 사용자 유사도 가중치를 계산할 때의 방법으로, 사용자가 평가한 값을 이용하여 사용자간의 오차와 표준편차를 구함으로써 가중치를 계산한다. 그러므로, 사용자의 평가 데이터에 많은 영향을 받는다. 데이터가 너무 적으면 가중치가 한 사용자에게 편중되므로 예측 값이 부정확하게 나오고 너무 많으면 가중치의 계산량이 많아진다. 특히, 예측하려는 아이템을 평가한 사용자의 개수가 많아지면, 그 계산량은 제곱 승으로 증가하게 된다[4].

2.2 Association Rule

연관 규칙은 "사용자의 행동 패턴은 일정한 규칙을 가진다."는 가정 하에 유용한 규칙을 찾아내는 방법이다. 이 방법은 대체로 장바구니 분석과 같은 사용자의 구매 경향을 파악하려는 곳에서 많이 쓰이는 방법이다. 대표적으로 Apriori 알고리즘, ARHP(Association Rule Hypergraph Partitioning) 알고리즘, FP-Tree 알고리즘 등이 있다[5][6].

연관 규칙의 평가 기준으로는 지지도, 신뢰도, 향상도 그리고, α-related가 있다. 이 중, 지지도와 α-related는 상대적으로 적은 평가 데이터에 대해서 부정확한 값을 나타내고 신뢰도는 방향성을 가지므로 사용할 수 없다. 그러므로, 본 논문에서 사용될 연관 규칙의 척도는 향상도로 한다.

3. 연관 규칙 군집에 의한 선호도 예측 시스템

아래 [그림 1]은 본 논문에서 제안하는 연관 관계 군집에 의한 선호도 예측 시스템의 전체적인 구성도이다.

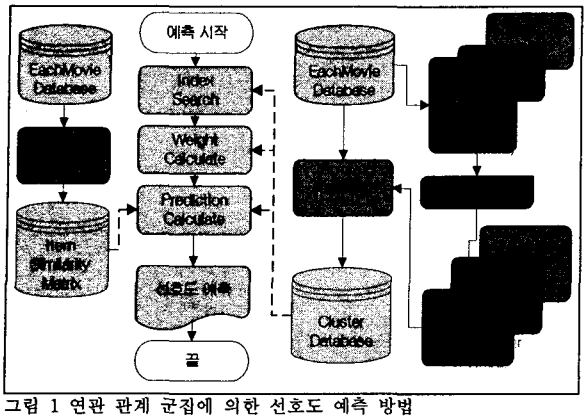

그림 1 연관 관계 군집에 의한 선호도 예측 방법

3.1 연관 규칙을 이용한 아이템의 향상도 측정

아이템간의 연관 규칙은 사용자의 아이템에 대한 평가 값을 이용하여 찾아낸다. 여기서, 기존의 논문에서는 지지도나 신뢰도를 사용하여 규칙을 찾아냈지만, 본 논문에서는 향상도를 사용한다. 향상도는 실제의 신뢰도를 독립 가정 하에서의 신뢰도로 나눈 값으로 식(1)과 같이 계산한다.

$$L(A,B) = \frac{\Pr(B|A)}{\Pr(B)} = \frac{\Pr(A \cap B)}{\Pr(A)\Pr(B)} \qquad \text{식(1)}$$

이 식은 실제의 지지도를 독립 가정 하에서의 지지도로 나눈 값과 동일하다. 또한 상호 대칭적이므로 방향성이 없다. L(A,B)값이 1에 가까우면 아이템 A와 B는 서로 독립에 가깝고, 1보다 작으면 음의 연관 관계, 1보다 크면 양의 연관 관계를 가진다. 그러므로, 아래 3.2절에서 사용할 α값은 1 이상으로 정한다.

우선 아이템들을 식(1)을 사용하여 Item Similarity Matrix를 만든다. 이 메트릭스는 두 아이템간에 방향성을 가지고 있지 않으므로, 직각 삼각형 구조가 된다.

3.2 α-cut

α-cut은 소속 함수의 [0,1]사이의 값에서 임의의 α(0≤α≤1)값이 되는 함수 값에 대한 퍼지 상태 변수의 구간을 나타낸다. 이 α-cut은 퍼지 집합의 원소들에 대해 집합에 속할 기준을 정의할 때 사용되는 방법이다. 임의의 X을 원소로 하는 퍼지 집합 A에 대해서 임의의 α∈[0,1]값을 가진 α-cut을 적용한 퍼지 집합 A_α는 다음과 같이 정의한다.

$$A_\alpha = \{x \mid A(x) \geq \alpha\}$$

따라서, 퍼지 집합 A_α는 퍼지 집합에 속할 소속정도의 값이 α값 이상으로 이루어진 집합이다. 본 논문에서는 Item Similarity Matrix의 값을 Boolean 값으로 변환시켜서 아이템간의 노드를 만드는 역할을 한다. 또한, 자기 자신과의 관계는 1보다 높으므로 제외시킨다. ISM에 α-cut을 적용시켜 아이템 군집에 필요한 노드를 생성한다.

3.3 Split Cluster Method

α-cut을 적용하여 만들어진 노드들을 서로 연결하여 연관 관계 그래프를 만든다. 이 Graph를 이용하여 연관성을 가진 군집을 찾아내야 한다. 기존 알고리즘 중, 가장 빠른 군집 알고리즘은 Hypergraph Clique Clustering의 Bottom-up 알고리즘이다[7]. 그러나, 위의 Tree 알고리즘도 O(n!)이라는 복잡도로 인해서 본 논문에서 실험할 수 없었다. 그러므로, 본 논문에서는 Split Cluster Algorithm을 제안한다.

알고리즘 1 Split Cluster Algorithm

```
DC←#Delete Node Classes;
EC←#Equivalence Classes;
While(EC[i] in Each EC) {
    pNode[0]←#All Unique Item in EC[i];
    while(DC[j] in Each DC) {
        while(pNode[k] in Each pNode) {
            if (pNode[k] ⊃ DC[j])
                Split pNode[k];
        }
        while(pNode[k] in Each Splited pNode) {
            while(pNode[l] in Each Unsplited pNode)
                if (pNode[l] ⊃ pNode[k])
                    Delete pNode[k];
        }
    }
    pCluster[] ← #pNode[] 삽입;
    Clear pNode[];
}
Assign(pCluster[]);
```

이 알고리즘은 삭제된 노드들을 이용하여 전체 하나의 군집에서 차례차례 분리해 나가는 방식을 사용한다. 분리된 군집들은 기존 군집들과 비교하여 포함관계가 성립하면, 삭제한다. 이렇게 삭제된 노드들을 모두 이용하여 분할된 최종 군집을 본 논문에서는 이용한다. 복잡도는 위 [알고리즘1]에서 보는 것과 같이 $O(n^3)$이 된다.

3.4 아이템 간의 유사도 계산

선호도를 예측하기 위해 군집에 속한 아이템을 평가한 사용자와의 유사도를 이용한다. 그러나, 아이템 간에는 차이가 있으므로, 그대로 선호도 예측을 할 경우 많은 오류가 발생한다. 그러므로, 아이템 간에 유사도를 측정하여, 그 값을 가중치로 하여 식(3)에 적용시킨다. 아이템 간의 유사도 계산은 Correlation-based Similarity를 적용한다[2].

3.5 새로운 사용자의 선호도 예측

새로운 사용자의 선호도 예측은 먼저 예측하려는 아이템이 속한 군집들과 그 군집에 속한 아이템에 의해서 이루어진다. 우선 기존의 피어슨 상관 계수를 변형하여 식(2)과 같이 만든다.

$$w(a,i) = \frac{\sum_j (r_{a,j} - \overline{r_a})(r_{i,j} - \overline{r_i})}{\sqrt{\sum_j (r_{a,j} - \overline{r_a})^2 \sum_j (r_{i,j} - \overline{r_i})^2}} \qquad \text{식(2)}$$

여기서 j는 사용자 a와 i가 공통으로 평가한 군집들의 수가 되고, $\overline{r_a}$ 는 사용자 a의 평균 선호도이다. 식(2)를 식(3)에 대입하여 예상 선호도를 구하게 된다.

$$P_{a,k} = \overline{r_a} + \frac{\sum_i^n sim(i,k)\, w(a,i)(r_{i,k} - \overline{r_i})}{\sum_i^n sim(i,k)\, w(a,i)} \qquad \text{식(3)}$$

한 아이템에 대해 평가한 사용자가 적을 경우는 대체로 상관 관계가 높게 나타나는데, 이것은 그 아이템을 평가한 사용자가 평가한 다른 아이템에 대해 모두 높은 관계를 가지

기 때문이다. 이러한 경우 이 아이템이 속한 군집의 크기는 매우 큰 반면, 이 아이템이 속한 군집의 수는 매우 적다. 그래서, 이 군집을 평가한 사용자들이 많아지게 되고, 예측 값의 정확도는 높아진다. 반면, 많은 사용자들이 평가한 아이템에 대해서는 많은 군집에 속하게 되어 평가의 정확도가 낮아진다. 그러므로, 한 아이템이 군집 속에 들어가는 수를 제한하고, 그 수보다 많아질 경우 기존의 피어슨 상관 계수를 이용하여 예측하고, 그 수보다 적을 경우 본 논문에서 제한하는 방법을 사용한다면 더 높은 적중률을 나타낼 것이다.

4. 실험 및 성능 평가

본 논문에서는 EachMovie[8] 데이터를 이용하였다. 이 중에서 데이터의 관계가 모호한 레코드를 삭제하여 영화를 100씩 10개의 그룹으로 나누어 실험하였다. 또한, 전체 평가 수는 1,428,362개이다(T10I100D1395k).

예측의 정확성을 평가하는데 가장 많이 사용하는 기준은 MAE(Mean Absolute Error)이다. MAE는 식(4)와 같이 나타낸다. 여기서 N은 총 예측 횟수이고, ε는 예측된 선호도와 실제 선호간의 오차를 나타낸다.

$$|\overline{E}| = \frac{\sum_{i=1}^{N} |\varepsilon_i|}{N} \qquad \text{식(4)}$$

아래 표는 두 알고리즘이 군집을 하는데 걸리는 시간을 비교한 것이다. 아이템의 개수가 증가할수록 Split Cluster Method가 더 좋은 성능을 보임을 알 수 있다.

[표 1] 두 알고리즘의 실행 시간(단위 : 초)

Algorithm \ Item	30개	40개	50개	60개	70개
Bottom-up Tree	2	6	11	21	52
Split Cluster	2	3	3	5	11

다음 그림 2는 아이템에 대해서 평가한 사용자의 수와 아이템이 군집에 속한 개수와 관계를 그래프로 표현한 것이다. X축은 사용자의 수이고, Y축은 군집의 개수를 나타낸다. 또한, α값을 각각 3, 6, 9로 나누어 아래 그래프에 표현하였다.

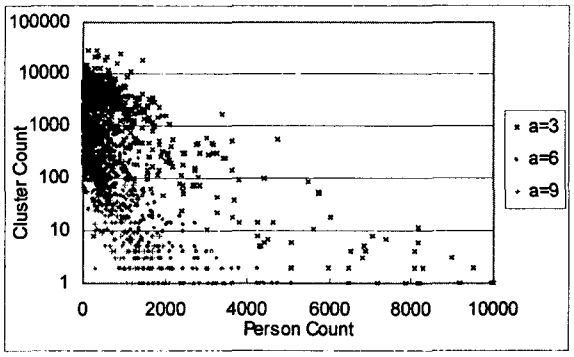

그림 2 아이템에 대한 사용자와 군집간의 관계

이 그래프에서 나타났듯이 사용자의 수가 적을수록 군집의 수가 많음을 알 수 있었다.

다음 그래프는 기존의 PCC(Pearson Correlation Coefficient)와의 MAE를 비교한 것이다.

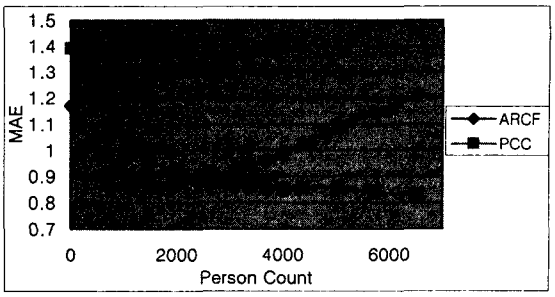

그림 3 사용자 수에 대한 MAE 평가(α=6)

그림 3을 보면, 사용자의 평가 수가 0에서 2500명인 지점에서 본 논문에서 제안한 알고리즘이 더 좋은 성능을 보임을 알 수 있었다. 그러나, 사용자 수가 증가할수록 기존 알고리즘보다 성능이 저하되었다.

5. 결론

본 논문에서는 연관 규칙을 이용하여 아이템을 군집하고, 이 군집을 평가한 사용자들의 선호도를 기반으로 사용자들의 선호도를 예측해 보았다. 이 방법이 모든 방면에 명확한 성능 개선을 이루지는 못했지만, 상대적으로 적은 로그에 대해 정확도가 높은 예측 값을 구할 수 있었다.

향후에 본 논문에서 발생하는 클러스터간의 아이템 편중 문제, 많은 군집에 한 아이템이 들어가는 문제, 클러스터링 알고리즘의 수행시간이 오래 걸리는 문제 등을 해결한다면, 제안한 알고리즘이 더 좋은 결과를 낼 수 있을 것이다.

참고 문헌

[1] Derry O' Sullivan, David Wilson, "Using Collaborative Filtering Data in Case-Based Recommendation", The 15th International FLAIRS Conference, 2002.

[2] Badrul Sarwar, George Karypis, "Item-based Collaborative Filtering Recommendation Algorithms", Accepted for publication at the WWW10 Conference, 2001.

[3] 정경용, 김진현, 이정현, "연관 사용자 군집과 베이지안 분류를 이용한 사용자 선호도 예측 방법", 한국정보과학회, 2001.

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

[5] E. H. Han, et. al., "Clustering Based On Association Rule Hypergraphs", DMKD, 1997.

[6] Jaiwei Han, Jian Pei, Yiwen Yin, "Mining Frequent Patterns without Candidate Generation", Proceedings of the ACM SIGMOD, 2000.

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

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