한국지능정보시스템학회
02 춘계학술대회 논문집, pp. 265~272

연관 규칙과 협력적 여과 방식을 이용한 추천 시스템

이기현* · 고병진** · 조근식***

Recommender System using Association Rule and Collaborative Filtering

Ki-Hyun Lee · Byung-Jin Ko · Geun-Sik Jo

요 약

기존의 인터넷 웹사이트에서는 사용자의 만족을 극대화시키기 위하여 사용자별로 개인화 된 서비스를 제공하는 협력적 필터링 방식을 적용하고 있다. 협력적 여과 기술은 비슷한 선호도를 가지는 사용자들과의 상관관계를 기반으로 취향에 맞는 아이템을 예측하여 특정 사용자에게 추천하여준다. 그러나 협력적 필터링은 추천을 받기 위해서 특정 수 이상의 아이템에 대한 평가를 요구하며, 또한 전체 사용자에 대해 단지 비슷한 선호도를 가지는 일부 사용자 정보에 의지하여 추천함으로써 나머지 사용자 정보를 무시하는 경향이 있다. 그러나 나머지 사용자 정보에도 추천을 위한 유용한 정보가 숨겨져 있다. 우리는 이러한 숨겨진 유용한 추천 정보를 발견하기 위하여 본 논문에서는 협력적 여과 방식과 함께 데이터 마이닝(Data Mining)에서 사용되는 연관 규칙(Association Rule)을 추천에 사용한다. 연관 규칙은 한 항목 그룹과 다른 항목 그룹 사이에 존재하는 연관성을 규칙(Rule)의 형태로 표현한 것이다. 이와 같이 생성된 연관 규칙은 개인 구매도 분석, 상품의 교차 매매(Cross-Marketing), 카탈로그 디자인, 염가 매출품(Loss Leader)분석, 상품 진열, 구매 성향에 따른 고객 분류 다양하게 사용되고 있다. 그러나 이런 연관 규칙은 추천 시스템에서 잘 응용되지 못하고 있는 실정이다. 본 논문에서 우리는 연관 규칙을 추천 시스템에 적용해, 항목 그룹 사이에 연관성을 유도함으로써 추천에 효율적으로 사용할 수 있음을 보였다. 즉 전체 사용자의 히스토리(History) 정보를 기반으로 아이템 사이의 연관 규칙을 유도하고 협력적 여과 방식과 함께 보조적으로 연관 규칙을 추천을 위해 사용함으로써 추천 시스템에 효율성을 높였다.

Key words: Recommender System, Association Rule, Collaborative Filterining

1. 서론

개인화 된 추천 시스템은 자동화된 정보 필터링 기술을 적용하여 고객의 취향에 맞는 상품을 추천해 주는 시스템이다. 추천 시스템에서 가장 중요한 것은 고객의 선호도를 정확하게 분석하고 정제하여 정확한 예측력으로 고객이 원하는 가장 적절한 상품을 추천해 줄 수 있는 능력이다.

이러한 추천을 위한 방법으로 많은 알고리즘들이 연구되어 왔으며 그 대표적인 예가 내용 기반 여과 방식이다. 내용 기반 여과는 사용자의 선호도와 아이템들의 내용과의 상관도에 기초하여 사용자가 선호할 만한 아이템들은 선택한다. 그러나, 내용 기반 여과는 다음과 같은 한계점을 가지고 있다. 첫 번째로 아이템은 컴퓨터가 이해할 수 있는 형식이어야 하는 어려움이 있다. 두 번째로 사용자들은 알고 있지는 않지만 원하는 새로운 아이템을 찾고 싶어한다. 그러나 이전의 개인적인 취향이 이런 새로운 아이템을 발견할 수 있는 어떤 암시도 제공해 주지 않는다. 내용 기반 여과 방식의 이러한 문제점을 해결하기 위하여 협력적 여과(collaborative filtering), 혹은 사회적인 정보 여과(social information filtering) 이라는 방법이 제안되었다.[7][8][9]

협력적 여과는 전체 아이템의 일부분에 대한 추천 대상자(Target User)의 평가를 기반으로 추천 대상자(Target User)는 전체 사용자로 이루어진 n 차원 공간에서 한 점으로 매핑(mapping)된다. 그 후 n 차원 공간 내에서 가장 가까운 유한 명의 이웃을 찾고 이웃이 가지고 있는 아이템 중 추천 대상자(Target User)에게 새로운 특정 아이템을 추천하여준다. 즉 달리 말해 A라는 사용자가 B라는 사용자와 유사하다면 B가 높은 선호도를 보인 아이

* 인하대학교 전자계산학과 석사 과정
** 인하대학교 전자계산학과 석사 과정
*** 인하대학교 전자계산학과 교수

템에 대해 아직 A가 가지고 있지 않은 아이템을 추천해주는 기법이다. 그러나 이러한 모델링(Modeling)의 문제점으로 사용자가 평가한 아이템의 집합은 전체 아이템 집합에 비해서 극히 작으므로 유사한 사용자를 찾기 위해 서는 강제적으로 임계치 이상의 아이템에 대해서 평가해야 한다는 것이다. 즉 임계치 이상의 아이템에 대해 평가하지 않은 사용자는 정확한 추천이 이루어 질 수 없다. 또 다른 문제점으로 협력적 여과 방식은 추천 대상자(Target User)와 유사한 히스토리(History)를 가지는 특정 n명의 이웃이 가진 정보를 바탕으로 추천을 해주므로 국소적 추천(local recommendation)에 머물고 나머지 이웃들에게서 이끌어 낼 수 있는 전역적 추천(global recommendation)을 놓칠 수 있다. 예를 들어 영화를 추천 받을 때 자신과 유사한 취향을 가지는 몇 명의 사용자로부터 추천을 받을 수도 있지만 또한 자신과 유사한 취향은 가지지 않으나 전체 사용자 집단에 대해 확률적으로 많은 사용자들이 본 영화를 추천 받을 수도 있을 것이다.

따라서 본 논문에서는 이러한 문제점을 해결하고 성능을 향상시키기 위해 데이터 마이닝(Data Mining)에서 사용되는 연관 규칙(Association Rule)과 기존에 협력적 여과(Collaborative Filtering)을 혼합한 시스템을 제안한다.

2. 관련 연구

2.1 협력적 여과(Collaborative Filterng)

협력적 여과 방식(Collaborative Filtering)은 오늘날 웹 상에 대부분의 성공적인 추천 시스템에 쓰이는 대표적인 추천 기술이다. 협력적 여과 방식은 다른 사용자들의 선호도 정보를 바탕으로 해서 유사한 성향을 가지는 이웃 사용자를 찾고, 그 이웃 사용자에 의해 높은 선호도를 보인 구매 아이템 을 사용자에게 추천하는 방식이다. 이들 시스템은 추천 대상자(Target User)와 유사한 히스토리 정보를 가지는 이웃들의 집합을 찾기 위해 통계적 방법을 사용한다. 우선 이런 유사한 이웃들이 찾아지면 특정 아이템을 추천 대상자(Target User)에게 추천해 주기 위하여 여러 알고리즘들을 사용하게 된다.[13]

2.1.1 이웃 선택(Nearest Neighborhood)

협력적 여과 방식에서 가장 중요한 단계로써 추천 대상자(Target User)와 유사한 히스토리 정보를 가지는 이웃들의 집합을 찾는 과정이다. 유사도를 계산하는 알고리즘으로는 대표적으로 Correlation Measure 또는 Cosine Measure 가 사용된다.

Correlation : 이 경우 두 사용자 a 와 b 사이에 유사도는 Pearson Correlation을 구함으로써 측정되어 질 수 있고 공식은 식 (1)과 같다.

$$corr_{ab} = \frac{\sum_i (r_{ai} - r_a)(r_{bi} - r_b)}{\sqrt{\sum_i (r_{ai} - r_a)^2 \sum_i (r_{bi} - r_b)^2}} \quad \text{.....(1)}$$

여기서 r_{ai} 는 사용자 a 에 아이템 i 에 대한 선호도 값이고, r_a 는 사용자 a 의 선호도 평균값이다.

Cosine : 이 경우 두 사용자 a 와 b는 M 차원 좌표 공간상에 두 벡터로 표현되어 지고 두 사용자 사이에 유사도는 두 벡터사이에 코사인 각을 계산함으로써 측정되어 질 수 있고 공식은 식(2)와 같다.

$$Cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|_2 * \|\vec{b}\|_2} \quad \text{.....식(2)}$$

여기서 " · " 는 두 벡터사이에 벡터곱을 나타낸다.

위와 같이 사용자 사이에 유사도가 구해지면 다음은 예측에 사용될 이웃의 수를 결정하는 것이다. 예측에 사용될 이웃의 수를 결정하기 위해서 Thresholding 기법과 Best-n-neighborhood 기법을 사용한다. Thresholding은 사용자간에 유사도 가중치가 어느 정도의 값 이상인 이웃들만을 사용해서 예측하도록 제안하는 방법이고 Best-n-neighborhood는 특정 사용자와 유사한 n명의 이웃을 사용해서 예측하도록 제안하는 방법이다. 한편 더 좋은 예측을 위해 두 방법을 조합해서 유사도 가중치가 어느 정도 값 이상인 이웃들 중 n명을 사용하는 방법도 있다.[4][13]

2.1.2 추천 아이템 선택 (Generation of Recommendation)

협력적 여과 방식에 최종 단계로써 위에서 구해진 이웃으로부터 추천 대상자(Target User)에게 추천되어질 아이템을 선택하는 단계이다. 이웃 선택 알고리즘으로 구해진 M 개의 이웃으로부터 출현 빈도가 가장 높은 아이템을 빈도순으로 정렬한다. 그후 추천 대상자(Target User)에는 아직 새롭거나 사지 않은 베스트 아이템 N개(Best-N)를 골라 추천해준다.[13]

2.2 연관 규칙(Association Rule)

연관 규칙(Association Rule)은 한 항목 그룹과 다른 항목 그룹 사이에 존재하는 연관성을 규칙의 형태로 표현한 것이다.[1][2] 연관 규칙 탐사는 사용자에 의해 적절하게 입력된 지지도(Support), 신뢰도(Confidence)라는 척도를 이용하여 데이터 상호간의 연관성을 파악할 수 있다. 연관 규칙은 일반적으로 다음과 같이 기술되어질 수 있다. I= { $i_1, i_2, i_3, i_4 \ldots$ }를 아이템(Items)들의 집합이라고 하고 D를 트랜잭션(Transaction)들의 집합이라 하자. 그러면 각 레코드는 I 에 속하는 아이템들의 부분집합으로 구성되어 진다. 이때 연관 규칙은 다음과 같은 형태로 기술되어 질 수 있다.

$$X \rightarrow Y \ (X \subset I, Y \subset I \ , \text{ AND } X \cap Y = \emptyset)$$

여기서 X, Y 는 각각 연관 규칙의 바디(Body), 헤드(Head) 라고 불린다. 연관 규칙이 내포하고 있는 의미는 트랜잭션 내에 아이템 X가 존재하면 어

느 정도에 확률을 가지고 Y 역시 트랜잭션 내에 존재함을 의미한다. 이들 연관 규칙은 주어진 트랜잭션들의 집합과 관련해 지지도(Support)와 신뢰도(Confidence) 라는 두 가지 측정값이 존재한다. 신뢰도(Confidence)는 아이템집합 X를 포함하고 있는 트랜잭션 중 아이템집합 Y 역시 포함하고 있는 트랜잭션의 비율이며, 지지도(Support)는 모든 트랜잭션에 대해 아이템집합 X, Y를 둘 다 포함하고 있는 트랜잭션의 비율을 나타낸다. 다른 말로 말해서 연관 규칙의 신뢰도는 아이템들 사이에 연관 정도(Correlation)를 나타내며 지지도는 아이템 사이에 연관 정도(Correlation)의 중요도를 나타낸다.

이러한 연관 규칙은 주로 장바구니 분석(Market Basket Analysis)에 주로 사용되며 구체적인 예는 다음과 같다. "분유를 구매하는 사람의 30%는 기저귀를 구매하며 전체 트랜잭션의 2%는 분유와 기저귀를 포함하고 있다." 여기서 30%는 이 규칙의 신뢰도(confidence)라고 불리며 2%는 이 규칙의 지지도라고 불린다.

이와 같이 생성된 연관 규칙은 개인 구매도 분석, 상품의 교차 매매(Cross-Marketing), 카탈로그 디자인, 염가 매출품(Loss Leader)분석, 상품 진열, 구매 성향에 따른 고객 분류 등등 다양하게 사용된다.[10]

2.2.1 Apriori 알고리즘

연관 규칙 발견 알고리즘 중의 하나인 Apriori는 지지도를 이용하여 동시에 자주 나타나는 항목(빈발 항목 집합)들을 정제하고 빈발 항목 집합에서 생성된 규칙들은 신뢰도를 이용하여 정제하는 방식이다. Apriori는 후보 항목 집합에서 각각의 지지도를 계산한 후 사용자가 정의한 지지도보다 크거나 같은 조건을 만족하는 데이터로 빈발 항목 집합을 구성한다. 그리고 후보 항목 집합은 전 단계의 빈발 항목 집합의 조인연산을 통해 구성된다. Apriori 알고리즘은 효율을 높이기 위해 AprioriTid, AprioriHybrid 등으로 확장되어 연구되고 있다.[10]

3. 추천 시스템

3.1 추천전략 (Recommendation Strategy)

일반적으로 많은 사용자들은 특정 아이템을 추천 받을 시 많은 다양한 경로를 통해 추천을 받을 수 있다.

첫째 : 자신과 비슷한 취향의 사용자로부터 특정 아이템을 추천을 받을 수 있으며 이것은 사용자 집합에 대해 단지 추천 대상자(Target User)와 유사한 취향을 가지는 n 명의 이웃에 정보를 기반으로 추천이 이루어지는 협력적 여과 방식과 같다.[7]

둘째 : 특정 분야에 전문가를 통해 추천을 받는 형식으로 이것은 지식 베이스와 협력적 여과 방식을 혼합한 추천 시스템과 유사하다.[6]

셋째 : 마지막으로 우리의 시스템에서 사용되는 추천 전략으로써 기존에 협력적 여과 방식과 함께 추천 대상자(Target User) 와는 취향에 있어 유사성은 없는 이웃이지만, 이들 중 많은 사람들이 인정하는 아이템을 추가로 추천하는 방법이다.

기존에 협력적 여과 방식의 경우 사용자 집합에 대해 단지 추천 대상자(Target User)와 유사한 취향을 가지는 n 명의 이웃에 정보를 기반으로 추천이 이루어졌기 때문에 n 명을 제외한 나머지 사용자 정보는 모두 무시되었다. 그러나 우리는 모든 사용자를 대상으로 추천 아이템에 대한 연관 규칙(association rule)을 동시에 유도함으로써 기존에 협력적 여과 방식이 가지고 있는 약점을 많은 부분 해결하였다.

즉 기존에 협력적 여과 방식은 추천 대상자(Target User)가 아이템을 추천 받기 위해서는 특정 수 이상의 아이템에 대해 평가를 요구하였고 그 평가를 기반으로 유사한 취향을 가지는 사용자를 찾아 아이템을 추천하였다. 그러나 우리의 추천 시스템은 추천 대상자(Target User)가 특정 수미만의 아이템에 대해 평가를 하여도 평가된 아이템과 연관이 많은 아이템을 추천하거나 아이템과는 연관이 없지만 많은 사람에 의해 인정되어지는 아이템을 추천하는 것이다. 또한 이웃이 아닌 나머지 대부분의 사용자 정보를 대상으로 유용한 연관 규칙을 유도함으로써 추가에 유용한 추천 정보를 제공하여 준다.

예제) <표1>과 같은 트랜잭션 테이블이 있다. 다음 트랜잭션 테이블은 사용자 {1, 2, 3, 4, 5, 6, 7, 8, 9} 에 대한 아이템 {A, B, C, D, E, F}의 선호도를 나타내고 있다. 즉 동그라미가 처진 셀은 아이템에 대해 사용자가 높은 선호도를 보임을 뜻한다. 이 테이블에 대해 연관 규칙을 찾기 위해 우리는 지지도(support),와 신뢰도(confidence)를 각각 50% 이상으로 잡는다고 가정하고 협력적 여과 방식을 사용하여 추천을 받기 위해서 추천 대상자(Target User)는 3개 이상의 아이템에 대해 평가를 해야한다고 가정한다.

<표 1> 사용자 트랜잭션 테이블

user	A	B	C	D	E	F
1	O		O		O	O
2	O		O		O	O
3	O	O	O	O		
4		O	O	O	O	
5		O				O
6		O	O	O	O	
7		O	O	O		
8	O	O	O			
9		O				

1) 추천 대상자(Target User)가 특정 임계값(3) 이상의 아이템에 대해 평가를 한 경우

예를 들어 사용자가 아이템 {A, C, E} 에 대해 높은 선호도를 보였다. 이 경우 추천 대상자(Target User)와 사용자{1,2} 사이에 거리가 가장 가까우므로 사용자{1,2}가 높은 선호도를 보인 아이템 중 아직 추천 대상자(Target User)에게 알려지지 않은 아이템 F를 추천한다. 또한 아이템 C에 대해 조건에 맞는 다음과 같은 연관 규칙 한 개를 발견할 수 있다

C -> B (support: 55%, confidence: 71%)

그러므로 아이템 B 도 추천한다.

2) 추천 대상자(Target User)가 특정 임계값(3) 미만의 아이템에 대해 평가를 한 경우

예를 들어 사용자가 아이템 {B, F} 에 대해 높은 선호도를 보였다. 이 경우 임계값(3) 미만에 아이템에 대해 평가를 했으므로 협력적 여과 방식이 사용되지 못한다. 그러므로 연관 규칙의 body 부분에 아이템 B 또는 F 가 존재하고 특정 지지도와 신뢰도 이상 을 가지는 연관 규칙 모두를 발견함으로써 추천 대상자(Target User)에게 특정 아이템을 추천한다. 즉 연관 규칙의 body에 B 또는 F를 가지면서 (support: 50% 이상 , confidence: 50% 이상)을 가지는 연관 규칙을 발견한다고 가정하면 다음과 같은 연관 규칙 한 개를 발견할 수 있다.

B -> C (support: 55% , confidence: 71%)

그러므로 아이템 C를 추천한다.

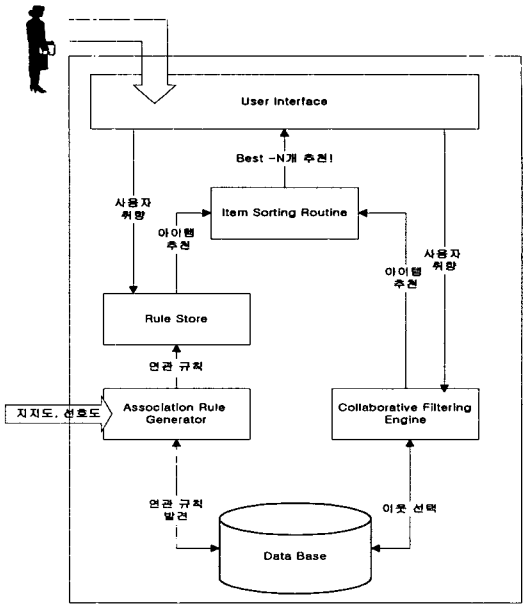
3.2 시스템 구조 (System architecture)

우리의 시스템은 크게 연관 규칙 생성 모듈과 협력적 여과 모듈로 나뉜다. 연관 규칙 생성 모듈은 많은 사람들에 의해 관심이 보여지는 아이템에 대한 연관 규칙들을 생성하며 협력적 여과 모듈은 추천 대상자(Target User)와 비슷한 취향을 가지는 이웃들에 의해 추천되어지는 아이템을 생성한다. 한편 이렇게 양쪽 모듈에서 생성된 아이템들에 대해 병합하고 우선 순위를 책정하는 모듈이 있다. 시스템 구조는 <그림 1> 과 같다.

3.2.1 연관 규칙 생성(Rule Generation)

추천 시스템에서 우리는 아이템에 대한 사용자의 선호도를 저장하고 있는 데이터 베이스로부터 연관 규칙을 유도하기 위해 연관 규칙 생성기(Association Rule Generator) 모듈을 사용한다. 연관 규칙 생성기는 룰을 생성하기 위해 Apriori Algorithm[2] 을 사용하였으며 외부로부터 적당한 지지도(Support)와 신뢰도(Confidence)값을 입력으로 받는다. 이때 연관 규칙 생성기는 입력받은 지지도와 신뢰도 이상의 값을 가지는 모든 연관 규칙들을 데이터 베이스로부터 생성하게 된다. 이렇게 생성된 연관 규칙들은 룰 저장소에 저장되며 사용자에게 아이템 추천을 위해 사용되어진다.

추천 시스템에서 연관 규칙을 생성하는데 있어서 우리는 사용자 입력에 대해 실시간으로 룰을 생



그림<1> 추천 시스템의 전체 구조

성하지 않는다. 대신 우리는 주기적으로 데이터 베이스로부터 연관 규칙을 생성하고 생성된 연관 규칙들을 룰 저장소로 업데이트한다. 사용자에 대한 실시간 반응속도를 높이기 위하여 협력적 여과(Collaborative Filtering) 와 연관 규칙(Association Rule) 생성 알고리즘을 동시에 실시간 적용하는 것이 아니라 추천에 있어 보조적인 역할을 하는 연관 규칙의 경우 미리 데이터 베이스부터 생성하여 룰 저장소에 저장하고 주기적으로 연관 규칙들을 업데이트한다.

3.2.2 협력적 여과(Collaborative Filtering)

협력적 여과 엔진(Collaborative Filtering Engine) 에서는 사용자로부터 부분적으로 아이템에 대한 선호도를 입력으로 받고 데이터 베이스로부터 유사한 취향을 가지는 이웃들을 Correlation Measure 또는 Cosine Measure 방법을 사용하여 M명 선택 한 후 , 선택된 이웃들이 가진 아이템 중에서 출현 빈도가 가장 높은 아이템을 빈도순으로 정렬한다. 이렇게 정렬된 아이템은 Item Sorting Routine으로 전달되게 된다.

3.2.3 추천 아이템 정렬(Item Sorting)

추천되어질 아이템들은 중요도에 따라 정렬되어져 추천 대상자(Target User)에게 추천되어야 한다. 이런 목적으로 아이템 정렬 루틴은 룰 저장소와 협력적 여과 엔진으로부터 추천 대상자(Target User)에게 추천되어질 아이템을 전달받는다. 전달받은 아이템들은 우선 순위를 체크해 오름차순으로 정렬한 후 추천 대상자(Target User)에게 베스트 아이템 N(Best-N)개를 골라 추천되어진다. 이때 아이템 사이의 우선 순위는 다음과 같이 정한다. 룰 저장소로부터 전달받은 아이템들의 집합을 A라고 하고 협력적 여과 엔진으로부터 전달받은 아이템들의 집합을 B라고 하자. 그러면 아이템 정렬

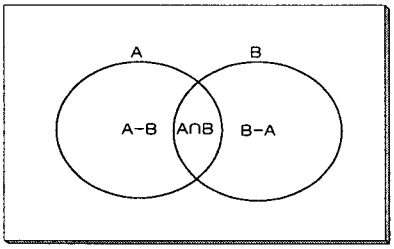
<표 2> 선호도 0.6 이상인 영화에 대한 지지도 변경에 따른 연관 규칙들과 발견된 연관규칙들의 갯수

지지도	연관 규칙	룰 갯수
5% 이상	Toy Story-> Grumpier old man, Jumanji-> Appolo13, Mr,Hollands Opus	1714
10% 이상	Toy Story-> 12Monkeys, Goldeneye-> Braveheart , Heat->12 monkey	387
15% 이상	Crimsom Tide-> The Fugitive, Fulp Fiction-> Aladdin, Brave Heart-> Appolo13.....	82
20% 이상	Appolo13-> Dances with Wolves, Appolo13->Batman, Dance with Wolves-> Batman..	17
30% 이상	ϕ	0

<표 3> 선호도 0.4 이상인 영화에 대한 지지도 변경에 따른 연관 규칙들과 발견된 연관규칙들의 갯수

지지도	연관 규칙	룰 갯수
5% 이상	Braveheart-> Appolo13, Die Hard-> Pulp Fiction, Sense and Sensibility->Batman.....	2790
10% 이상	Toy Story-> 12Monkeys, Goldeneye-> Braveheart , Heat->12 monkey	740
15% 이상	Toy Story-> Independence Day, Golden Eye-> Appolo13,Brave Heart-> Appolo13....	208
20% 이상	clear and present danger-> true lies , forrest gump -> the fugitive	55
30% 이상	Appolo13-> Dances with Wolves, Appolo13->Batman, Dance with Wolves-> Batman..	3
40% 이상	ϕ	0

루틴에 존재하는 아이템들의 집합은 <그림 2> 과 같이 나타낼 수 있고 추천을 위한 아이템들의 집합들은 $A \cap B$, $B-A$, $A-B$ 영역으로 나누어진다.



<그림 2> 집합 A 는 룰 저장소에서 넘어온 아이템의 집합이고 집합 B 는 협력적 여과 엔진으로부터 받은 아이템의 집합이다.

1) 1순위($A \cap B$)

이 경우 협력적 여과 방식에 의해 추천되어지고 동시에 연관 규칙에 의해 추천되어지는 아이템은 우선 순위가 높다고 할 수 있다. 즉 추천 대상자(Target User)와 비슷한 취향에 사용자에게 추천 받은 동시에 다수에 여러 사람에 의해서도 추천되어진 아이템이기 때문이다.

2) 2순위($B-A$)

그 다음으로 우선 순위를 가지는 아이템은 추천 대상자(Target User)와 비슷한 취향에 사용자에게 추천 받은 아이템들의 집합으로 주(Primary) 알고리즘인 협력적 여과 엔진에 의해서만 추천되어지는 아이템들의 집합이다. 이때 아이템의 선호도가 0.6 미만이라면 추천 대상에서 제외되고 다음 단계에 따른다.

3) 3순위($A-B$)

마지막으로 다수에 여러 사람에 의해 추천되어지는 아이템들에 집합으로 보조(Secondary) 알고리즘인 연관 규칙에 의해 추천되어지는 아이템들의 집합이다. 만일 추천되어지는 아이템이 임계치 이상의 지지도 또는 신뢰도를 넘으면 추천한다.

3.3 추천 절차

우리의 추천 시스템의 추천 대상자(Target User)에 대한 추천 절차는 전체적으로 다음과 같다.

1) 사용자 입력

추천 대상자(Target User) 는 사용자 인터페이스를 통해 N 개의 아이템에 대한 선호도를 입력한다.

2) 추천 아이템 선택

ㄱ) 만일 사용자가 임계치 이상의 아이템에 대해 평가값을 입력하였다면 Collaborative Filtering Engine 과 룰 저장소(Rule Store)로부터 추천되어질 아이템을 아이템 정렬 루틴으로 전달한다.

ㄴ) 사용자가 임계치 미만의 아이템에 대해 평

입력하였을 경우 룰 저장소로부터 추천되어질 아이템을 선택한 후 아이템 정렬 루틴으로 전달한다.

3) 추천 아이템 정렬

아이템 정렬 루틴은 전달받은 아이템들을 우선순위에 따라 정렬한 후 사용자에게 전달하여 준다.

4. 실험 및 평가

본 논문에서 제안한 추천시스템은 Java 2 로 구현하였으며, 실제 실험에 쓰인 데이터 집합은 컴팩 연구소에서 18개월 동안 협력적 여과 알고리즘을 연구하기 위해서 영화에 대한 사용자의 선호도를 조사한 EachMovie 데이터를 사용하였다. 이 데이터는 총 72916명의 사용자와 1628종류의 영화에 대해서 0.0에서부터 1.0까지 0.2 간격으로 명시적으로 평가한 선호도로 구성되어 있다.[3]

우선 데이터 집합으로부터 연관 규칙을 발견하기 위해 선호도와 지지도와 값을 조정하면서 연관 규칙을 발견하였다. 연관 규칙으로 발견된 아이템과 협력적 여과 방식으로 발견된 아이템은 하나로 합쳐져 추천시스템의 성능 평가를 위해 사용되어졌다.

4.1 연관 규칙 발견

본 논문에서는 사용자에 의해 선호도 0.4 미만의 평가를 받은 모든 영화들을 흥미가 없는 영화로 간주하고 연관 규칙을 발견하는데 있어 제외시켰다. 그리고 연관 규칙을 발견하기 위하여 우선 선호도 0.4 이상인 영화를 대상으로 지지도를 조절해 가며 연관 규칙들을 발견하였으며 다음으로 선호도 0.6 이상인 영화를 대상으로 지지도를 조절해 가며 연관 규칙들을 발견하였다. 결과는 다음과 같다.

1) 선호도 0.6 이상

전체 사용자 72916명에 대해 지지도(Support) 5%(3645명) 이상을 가지는 연관 규칙들은 전체적으로 1714개가 나왔으며 모든 연관 규칙들이 body와 head에 각각 하나에 아이템을 가지고 있었다. 즉 A, B-> C 와 같은 body나 head에 복수에 아이템을 가지는 연관 규칙들은 발견할 수 없었다. 지지도 30% 이상에서 발견되는 연관 규칙들은 없었고 지지도 15% 이상에 대해서는 연관 규칙들이 매우 희박하게 발견되었다. 또한 선호도 0.4 이상에서 발견되는 연관 규칙들의 개수에 비해 절반 수준에 머물렀다. <표2>

2) 선호도 0.4 이상

전체 사용자 72916명에 대해 지지도(Support) 5%(3645명) 이상을 가지는 연관 규칙들은 전체적으로 1714개가 나왔으며 선호도 0.6 이상인 연관 규칙들과 마찬가지로 모든 연관 규칙들이 body와 head에 각각 하나에 아이템을 가지고 있었다. 즉 A, B-> C 와 같은 body나 head에 복수에 아이템을 가지는 연관 규칙들은 발견할 수 없었다. 또한 지지도 20% 이상에 대해서는 연관 규칙들이 매우 희박하게 발견되었다. <표3>

4.2 실험 및 평가 방법

4.2.1 평가 방법

실험을 위해 우선 전체 실험 자료를 80%, 20%로 임의로 분류하여 80%에 속하는 자료를 학습자료(Training Data)로 사용하고 20%에 속하는 자료를 검사 데이터(Test Data)로 사용하였다. 우리는 기존에 협력적 여과 방식의 추천 시스템에 연관 규칙을 추가하였을 때의 성능 향상을 알아보는 것이므로 우선 협력적 여과 방식만을 사용한 추천한 시스템과 연관 규칙을 추가했을 때 의 시스템의 성능 차이를 비교하였다. 협력적 여과 방식을 위한 이웃 선택은 10부터 시작해서 20만큼씩 늘려나갔고 추천되어질 아이템 수(Best -N)를 20개의 아이템으로 제한하였다.

우리는 추천되어진 베스트 아이템 N개(Best-N)가 검사 자료 내에서 얼마나 정확하게 예측되어졌나를 알아보기 위해 정보 검색(Information Retrieval)에서 사용되어지는 Recall, Precision을 사용하였다.[13][14]

Recall: 검사 데이터(test set)에 대한 매칭 데이터(hit set)의 비율로 다음과 같다.

$$Recall = \frac{|test \cap best - N|}{|test|} \quad \text{.....식(3)}$$

Precision: 추천 대상자에게 추천되어진 베스트 아이템 N개에 대한 매칭 데이터의 비율로 다음과 같다.

$$\Pr ecision = \frac{|test \cap best - N|}{N} \quad \text{.....식(4)}$$

그러나 위 측정 방법은 서로 반비례 관계가 존재한다. 예를 들면 N 이 증가하면 Recall 은 증가하고 Precision은 감소하는 경향이 있다. 그러므로 Recall 과 Precision에 동일한 가중치를 주어 하나로 통합한 standard F1 metric 방식을 사용한다.[13]

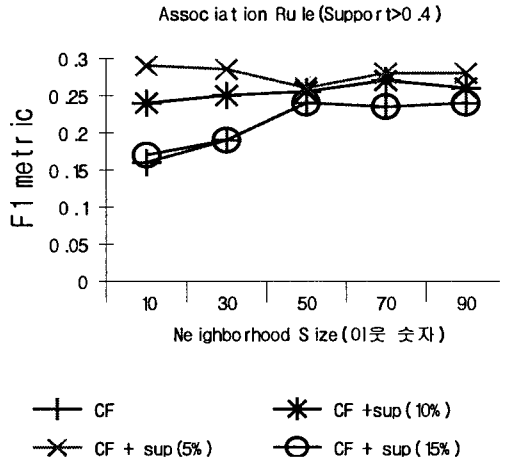
$$F1 = \frac{2 * Recall * \Pr ecision}{Recall + \Pr ecision} \quad \text{.....식(5)}$$

4.2.2 성능 평가

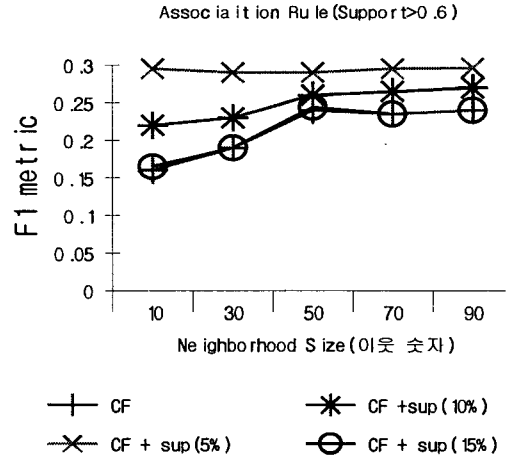
순수 협력적 여과 방식만을 사용한 시스템과 협력적 여과 방식과 연관 규칙을 같이 사용하였을 때의 성능 비교 결과는 다음과 같다.

<그림 3>은 협력적 여과 방식(CF)만 사용한 결과와 사용자 선호도 0,4 이상인 아이템으로 이루어진 연관 규칙들을 지지도(sup)를 바꾸어가면서 협력적 여과 방식과(CF) 와 같이 사용하였을 때의 결과를 비교한 것이다. 그림에서 협력적 여과 방식만 사용한 경우는 이웃의 숫자(Neighborhood Size)가 어느 수준까지 증가하면 성능향상이 더딘 것을 볼 수 있다. 그러나 연관 규칙을 같이 사용한 경우 이웃의 숫자가 적을 경우에도 좋은 결과를 나타내고 있다. 특히 지지도 5% 이상인 연관 규칙을 협력적 여과 방식과 같이 사용한 경우(CF+ sup(5%)) 에는 이웃의 숫자에 상관없이 매우 안정된 성능을

보여주고 있었다. 그러나 연관 규칙의 수가 비교적 적게 발견된 지지도(15%)이상인 연관 규칙을 협력적 여과 방식과 같이 사용한 경우(CF+sup(15%)) 순수 협력적 여과 방식(CF)와 결과 차이가 거의 없었다. 일반적으로 연관 규칙의 개수가 많을수록 성능이 더 뛰어난 것을 볼 수 있었고 특히 협력적 여과 방식에서 이웃의 숫자를 적게 잡을 경우 연관 규칙을 같이 사용하면 성능이 매우 좋음을 알 수 있었다.



<그림 3> 순수 협력적 여과 방식과 연관 규칙 (support>0.4)을 같이 사용한 협력적 여과 방식과의 비교



<그림 4> 순수 협력적 여과 방식과 연관 규칙 (support>0.6)을 같이 사용한 협력적 여과 방식과의 비교

<그림 4>는 협력적 여과 방식(CF)만을 사용한 결과와 사용자 선호도 0.6 이상인 아이템으로 이루어진 연관 규칙들을 지지도(sup)를 바꾸어가면서 협력적 여과 방식과(CF)와 같이 사용하였을 때의 결과를 비교한 것이다. 앞에서 비교한 사용자 선호도 0.4 이상인 연관 규칙을 적용하였을 때와 많은 차이를 보이지는 않았다. 역시 지지도 15% 이상인 연관 규칙을 적용하였을 때는 순수 협력적 여과 방식과 차이가 거의 없었다. 이유는 연관 규칙의 개수가 적기 때문에 시스템에 영향을 많이 못 주는 것 같다.

5. 결론 및 전망

본 논문에서는 기존에 사용자에게 아이템 추천을 위하여 협력적 여과(Collaborative Filtering)와 연관 규칙(Association Rule)을 통합한 시스템을 제시하였다. 본래 연관 규칙은 데이터 마이닝(Data Mining)에서 지식 발견을 위해 사용되는 알고리즘에 일종이지만 추천 시스템에서도 효율적으로 사용될 수 있음을 보았다. 연관 규칙(Association Rule)을 사용함으로써 기존에 협력적 여과로는 찾아질 수 없는 유용한 아이템을 발견할 수 있었고 이 아이템을 협력적 여과를 통해 발견되어지는 아이템과 함께 추천함으로써 추천의 효율을 높일 수 있었다.

그리고 이와 함께 우리의 시스템에 대한 발전 방향을 몇 가지 언급할 수 있다.. 첫째로 연관 규칙을 발견하는데 있어 어느 정도에 지지도와 신뢰도를 주어 룰을 발견해야 하는 가이다. 둘째로 실시간으로 룰을 생성할 수 있는 방법론을 찾는 것이다. 이 문제에 대해서는 추천 대상자(Target User)가 직접 지지도와 신뢰도를 입력하고 실시간으로 해당하는 룰만을 축출할 수 있는 제약 연관 규칙(Constraint Association Rule)을 생각 할 수 있다. 셋째로 연관 규칙의 지지도 와 신뢰도뿐만 아니라 흥미도(Interestingness)도 고려하여 어떻게 연관 규칙의 품질을 높일 수 있는 가이다. 마지막으로, 협력적 여과, 연관 규칙과 함께 지식 베이스를 추천 시스템에 추가할 때 시스템의 성능을 평가하는 것이다.

참고 문헌

[1] W. Lin, C. Ruiz, and S. A. Alvarez. "A new adaptive-support algorithm for association rule mining." Technical Report WPI-CS-TR-00-13, Department of Computer Science, Worcester Polytechnic Institute, May 2000

[2] R. Agrawal, H. Mannila, R. Srikant. "Fast algorithm for mining association rules in large databases. In Proceedings of the 20th International Conference on Very Large Data Bases, pages 487-499, September 1994

[3] P. McJones. "Eachmovie collaborative filtering data set. http://www.research.digital.com/SRC/eachmovi " 1997 DEC System Research

[4] Anuja Gokhale and Mark Claypool. "Thresholds for more accurate collaborative filtering " In Proceedings of the IASTED International Conference on Artificial Intelligence and Soft Computing, Honolulu, Hawaii, USA, August 9-12 1999

[5] Donglass W, Jinmook Kim, "Implicit Feedback for Recommender System" In the Proceedings of the AAAI Workshop on Recommender Systems, Madision,WI, July, 1998
[6] Robin Burke. "Integrating Knowledge-based and Collaborative-filtering Recommender Systems" In Workshop on AI and Electronic Commerce, AAAI 1999
[7] Paul, R, Neophytos, I, Mitesh, S. Peter, B, Jonh, R, GroupLens " an open architecture for collaborative filtering of netnews" In Proceedings of ACM CSCW'94 Conference on Computer Supported Cooperative Work, pages 175-186, 1994
[8] Terveen, L, hill, W, Amento, B. McDonald, D. and Creter, J ,"Social Information Filtering Algorithms for Automating "Word of Mouth", In Proceedings of the CHI'95. Denver, CO. May 1995
[9] Patrick Baudisch, "Joining Collaborative And Content-based Filtering" CHI'99 Workshop Interacting with Recommender System 15/16 May 1999 Pittsbur호,Pennsylvania, USA
[10] Jiawei Han, Micheline Kamber "Data Mining Concepts and Techniques" Morgan Kaufmann Publishers 2000
[11] Marko Balabanovic ,Shoham,"Content-Based, collaborative Recommendation"Communication of the ACM, 40(3), March 1997
[12] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl " Analysis of Recommendation Algorithms for E-Commerce" Minnesota, 2000
[13] Herlocker, J, Konstan," An Algorithmic Framework for Performing Collaborative Filtering" In Proceedings of ACM SIGIR'99 ACM Press