카테고리 정보를 이용한 추천 성능의 향상

김춘호^{O} 김준태
동국대학교 컴퓨터공학과
{chkim^{O}, jkim}@dongguk.edu

Improvements of Recommendation Performance with Categorical Information

Choonho Kim^{0} Juntae Kim
Dept. of Computer Engineering, Dongguk University

요 약

추천 시스템은 사용자의 아이템에 대한 선호도를 예측함으로써, 사용자에게 적합한 아이템을 추천한다. 이러한 추천 시스템은 희소성과 확장성의 문제를 안고 있다. 희소성이란 사용자의 선호도 예측의 토대가 되는 정보의 부족으로 인하여 추천 아이템의 범위가 제한되는 것이고, 확장성이란 사용자나 아이템의 수가 증가함에 따라 추천 시간이 증가하는 것이다. 본 논문에서는 아이템의 카테고리 정보를 이용한 다중 레벨 연관규칙을 선호도 예측에 적용하여 희소성과 확장성의 문제를 완화하고자 하였다. 연관규칙을 이용하여 선호도 예측을 위한 모델을 구축하여 확장성을 해결하고, 다중 레벨 연관규칙을 이용하여 추천 아이템의 범위를 확장할 수 있었다. 단일 레벨만을 사용한 방법과 비교한 결과, 다중 레벨을 사용한 방법이 좋은 성능을 보임을 확인할 수 있었다.

1. 서론

현재 인터넷 쇼핑몰에서 판매하는 아이템의 수는 일반 백화점과 맞먹으리만큼 다양하고 사용자가 원하는 아이템을 쉽게 찾을 수 있도록 카테고리별로 분류되어 있다. 인터넷 쇼핑몰을 이용하는 사용자의 수가 기하급수적으로 증가함에 따라, 사용자들의 구매정보를 분석하여 구매하지 않은 다른 아이템을 사용자에게 추천할 수 있는 추천 시스템이 많이 활용되고 있다. 하지만, 방대한 아이템의 수에 비해 한 사용자가 구매하는 아이템의 수는 극히 일부에 불과하여 사용자의 선호도(구매) 정보는 희소하게 되고 추천할 수 있는 아이템의 범위가 제한되는 희소성의 문제가 발생된다. 또한, 사용자 사이의 유사도를 계산하는 추천 방식은 사용자의 수가 증가함에 따라 계산 시간이 증가하여 실시간 추천이 어렵게 되는 확장성의 문제가 발생된다[1].

본 논문에서는 추천 시스템에서의 이러한 문제를 해결하기 위한 모델기반(model-based) 방법을 제안한다. 본 논문에서 제안하는 방법은 아이템간의 연관규칙을 찾아내어 추천을 위한 모델을 구성하고 이 모델을 이용하여 아이템을 추천함으로써 확장성 문제를 해결하고, 아이템의 다중레벨 카테고리 정보로부터 상위레벨의 연관규칙도 추천에 이용함으로써 희소성 문제를 해결하는 것이다. 이러한 다중레벨 연관규칙을 이용한 추천 알고리즘의 성능 향상을 알아보기 위하여 아이템 사이의 연관규칙만을 이용한 방법과의 비교 실험을 수행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 제안하는 다중레벨 카테고리 정보를 이용한 모델기반 추천 알고리즘에 대하여 설명하며, 3장에서는 성능 측정을 위한 실험 방법 및 결과를 분석하고, 4장에서 결론을 맺는다.

2. 다중레벨 연관규칙을 이용한 추천

본 장에서는 연관규칙 마이닝을 이용한 추천 방식과, 아이템의 계층적 구조를 통하여 상위 계층의 카테고리 사이의 관련도를 이용함으로써 추천 정확도를 높이는 방법을 설명한다.

2.1 연관규칙 마이닝

연관규칙 마이닝(association rule mining)은 트랜잭션 데이터베이스로부터 동시에 출현하는 아이템들을 찾아내어 아이템들 사이의 관련성을 찾아내는 것이다. 아이템 A가 출현할 때 아이템 B가 동시에 출현하는 빈도가 높다면 연관 규칙은 $A \Rightarrow B$ (if A, then B)와 같은 형식으로 표현되며, 연관규칙에 대한 정량적 평가는 다음과 같은 지지도(support)와 신뢰도(confidence)로 나타낼 수 있다[2].

$$support(A \Rightarrow B) = p(A \cup B)$$

$$confidence(A \Rightarrow B) = p(B \mid A)$$

Support는 전체 트랜잭션중 아이템 A, B가 동시에 출현한 빈도로써 연관 규칙의 유용성을 나타내고, confidence는 아이템 A가 출현한 경우 중에서 아이템 B가 출현한 빈도로써 연관규칙의 유효성을 나타낸다. 유용한 연관규칙이란 사용자에 의해 정의된 최소 지지도($minSup$)와 최소 신뢰도($minCon$)을 만족하는 연관규칙들을 말한다.

본 논문에서는 사용자-아이템 선호도 행렬에서 각 사용자가

특정 아이템을 선호한다고 알려진 경우를 1, 그렇지 않은 경우를 0으로 하고, 각 사용자의 정보를 하나의 트랜잭션으로 보고 아이템들 사이의 연관규칙을 구한다.

2.2 연관규칙을 이용한 추천

연관규칙을 이용한 추천 방법은 사용자(active user)가 선호하는 아이템 *i, j*가 있을 때 아이템 *k*에 대한 선호도 예측값을 유용한 연관규칙 중 조건부가 *i* 또는 *j*이고 결과부가 *k*인 연관규칙들이 있는지를 찾아서 이러한 연관규칙들의 confidence의 합으로 계산하는 것이다. 연관규칙을 사용하는 방법으로서 적용 가능한 연관규칙들의 confidence 중 최대값을 적용하는 방법도 생각할 수 있으나[3], 본 논문에서는 더 많은 연관규칙이 적용되는 아이템에 가중치 주기 위하여 합을 사용하여 정규화 하였으며 실험적으로 confidence의 합을 적용하는 경우가 정확도가 높음을 확인하였다.

n 명의 사용자와 m 개의 아이템 사이의 선호도 행렬(preference matrix)을 P라 하고, P로부터 계산된 연관규칙들의 confidence를 나타내는 연관 행렬(association matrix)을 A라고 하자. 즉, A의 원소 a_{ij}는 연관규칙 $i \Rightarrow j$의 confidence 이다. Active user의 선호도 벡터를 u라고 할 때, 연관규칙의 confidence 합을 계산한 active user에 대한 추천벡터 r은 아래와 같이 나타낼 수 있다. 아이템의 연관규칙을 이용한 추천은 이렇게 계산한 추천 벡터 r에서 예측값이 높은 상위 N 개를 추천하는 방식이다.

$$\mathbf{r} = \mathbf{u} \cdot \mathbf{A}$$

각 아이템의 예측값은 sigmoid 함수를 통하여 다음과 같이 [0.5, 1]로 정규화하여 사용한다.

$$\mathrm{Norm}(\mathbf{r}_i) = \frac{1}{1+e^{-\mathbf{r}_i}}$$

2.3 다중레벨 연관규칙을 이용한 추천

아이템 사이의 연관규칙을 이용한 추천에서, 사용자들의 선호도 정보가 많지 않은 경우에는 유용한 연관규칙 수가 많지 않으므로 연관 행렬은 매우 희소한 행렬(sparse matrix)이 된다. 실제로 사용자의 구매 데이터를 선호도로 사용하는 경우 각 사용자는 수많은 아이템들 중 극히 일부만을 구매하게 되므로 연관규칙의 수는 매우 적게 된다. 연관 행렬이 희소 행렬인 경우 대부분의 아이템에 대한 선호도 예측이 불가능하여 사용자에게 적절한 아이템을 추천하지 못하는 문제가 발생하게 된다. 본 논문에서는 이러한 문제를 해결하기 위하여 아이템의 상위 레벨 카테고리 사이의 연관규칙을 추천에 이용하는 방법을 제안한다.

본 논문에서 제안하는 다중레벨 연관규칙을 이용한 추천 방법은 사용자가 선호하는 아이템이 속하는 카테고리 c_i를 찾고, 유용한 카테고리 연관규칙 중 조건부가 일치하는 $c_i \Rightarrow c_j$가 있다면 카테고리 c_j에 포함되는 모든 아이템들에 일정 비율의 선호도를 부여하는 것이다.

다중레벨 연관규칙을 이용한 추천 과정은 다음과 같이 나타낼 수 있다. 먼저, 하위 레벨(k-1)의 카테고리(레벨 0는 각 아이템)가 상위 레벨(k)의 카테고리중 어디에 포함되는지를 나타내는 카테고리 관계 행렬(category relation matrix)을 C_k라 하자. C_k의 원소 $C_k(i,j)$는 레벨 k-1의 카테고리 i가 레벨 k의 카테고리 j에 포함되면 1, 아니면 0으로 표현된다. 그러면 레벨 k 카테고리에 대한 선호도 행렬 P_k와 active user의 레벨 k 카테고리에 대한 선호도 벡터 u_k는 다음과 같이 계산된다.

$$\mathbf{P}_k = \mathbf{P}_0 \cdot \mathbf{C}_1 \cdot \mathbf{C}_2 \cdot \ldots \cdot \mathbf{C}_k$$

$$\mathbf{u}_k = \mathbf{u}_0 \cdot \mathbf{C}_1 \cdot \mathbf{C}_2 \cdot \ldots \cdot \mathbf{C}_k$$

레벨 k 카테고리 사이의 연관 행렬 A_k는 P_k로부터 찾아지는 카테고리 사이의 연관규칙들의 confidence로 구성된다. 이때 사용자가 특정 카테고리에 해당되는 아이템을 선호한 경우 동일 카테고리의 다른 아이템들에도 선호도를 주기 위하여 A_k의 대각 요소는 모두 1로 한다. 즉 모든 레벨 1 이상의 카테고리 c_i에 대하여 $c_i \Rightarrow c_i$의 confidence는 모두 1이다. 이렇게 구성된 정보로부터 모든 레벨의 연관규칙을 이용하는 active user에 대한 추천 벡터 r은 아래와 같이 나타낼 수 있다. 식에서 α_k는 레벨 k 연관규칙에 대한 가중치를 나타내는 상수로서 합은 1이다.

$$\mathbf{r} = \alpha_0 \cdot \mathbf{u}_0 \cdot \mathbf{A}_0 + \alpha_1 \cdot \mathbf{u}_1 \cdot \mathbf{A}_1 \cdot \mathbf{C}_1^{\mathrm{T}} + \ldots + \alpha_k \cdot \mathbf{u}_k \cdot \mathbf{A}_k \cdot \mathbf{C}_k^{\mathrm{T}} \cdot \ldots \cdot \mathbf{C}_1^{\mathrm{T}} \quad \left(\sum \alpha_k = 1\right)$$

3. 실험 및 결과

본 논문에서 제시한 다중레벨 연관규칙을 이용한 추천 알고리즘의 성능을 평가하기 위하여 단일레벨 연관규칙만을 이용한 추천 방법과의 비교 실험을 수행하였다. 모든 실험은 Pentium-4 1.8GHz, 1Gbytes 메모리, Windows 2000 운영체제하에서 수행되었다.

3.1 데이터 집합

실험을 위한 데이터는 MovieLens Dataset을 사용하였다. MovieLens Dataset은 사용자들의 영화에 대한 선호도를 6개의 등급(0, 1, 2, 3, 4, 5)으로 평가한 명시적인 dataset으로 943명의 사용자의 1682개의 영화에 대한 100,000개의 평가값으로 구성되어 있다. MovieLens 아이템들은 2개의 레벨로 구성되어 있으며, 영화에 대한 상위 레벨로서 18개의 영화 장르가 있고 각 아이템은 1개 이상의 장르에 포함되어 있다[5]. 본 논문에서는 100명의 사용자를 랜덤 샘플링하여 사용하였고, 아이템에 대한 연관규칙 마이닝을 적용하기위해 각 사용자의 특정 아이템에 대한 선호도를 1과 0의 형태로 변환하였다. 변환 방법은 각 사용자의 평가 스케일을 반영하기 위하여 각 사용자의 평가값의 평균보다 높은 아이템의 평가값을 1, 그렇지 않은 평가값은 0으로 변환하였다.

3.2 실험방법 및 평가방법

실험은 data set을 평가값의 개수를 기준으로 하여 80%를 training set으로 하고 20%를 test set으로 하여 5번의 실험결과를 평균하는 5-fold cross validation을 사용하였다. 실험 방법은 training set의 평가값으로 연관규칙을 생성하고, 생성된 규칙을

이용하여 N개의 아이템을 추천하여 숨겨진 아이템(test set의 아이템)이 top-N set에 포함되는 비율을 계산하였다. 추천 성능은 recall과 precision을 각각 같은 가중치로 혼합하는 micro-averaged F1 measure[4]를 사용하였다.

$$recall = \frac{|test \cap top-N|}{|test|}$$

$$precision = \frac{|test \cap top-N|}{N}$$

$$F1 = \frac{2 \times recall \times precision}{recall + precision}$$

우선 연관규칙 적용방법에서 여러 개의 적용 가능한 연관규칙이 있을 경우 최대 confidence 규칙을 적용하는 방법과 confidence의 합을 적용하는 방법과의 차이를 실험하였고, 본 논문에서 제시한 다중레벨 연관규칙을 이용한 방법의 성능을 단일레벨 연관규칙 추천 방법과 비교하였다. Top-N 추천에서 N은 10, 20, 30, 40, 그리고 50 까지로 변화시키면서 F1을 측정하였다.

3.3 실험 결과

그림 1은 두 가지 연관규칙의 적용 방법에 대한 실험 결과로써 AR-MAX는 적용 가능한 연관규칙들의 confidence 중 최대값을 사용한 방법이고, AR+는 연관규칙들의 confidence의 합을 사용한 방법이다. 실험 결과에서 본 논문에서 제시한 AR+가 평균 20%정도의 좋은 성능을 보임을 볼 수 있다.

그림 2는 카테고리 정보를 이용한 추천 알고리즘의 실험 결과이다. AR+는 레벨 0 아이템의 연관규칙을 이용한 방법으로, 본 실험에서는 가능한 모든 연관규칙을 사용하여 실험하였다. 데이터 집합이 희소하므로 *minSup*와 *minCon*의 적용은 오히려 유용한 연관규칙의 수를 급격하게 감소시켜 추천 성능의 저하를 초래한다. MAR은 본 논문에서 제시한 다중레벨 연관규칙을 이용한 방법으로, 데이터 집합에 있는 정보인 레벨 1의 카테고리 정보(장르)까지 이용하였다. 실험적인 결과로서 상위 카테고리 레벨의 연관규칙에는 비교적 높은 *minSup*와 *minCon*을 적용할수록 좋은 성능을 나타냄을 알 수 있었으며, 위의 결과는 레벨 1의 연관규칙에 대하여 *minSup* = 80%, *minCon* = 80% 를 적용한 결과이다. 추천 백터를 계산할 때 각 레벨에 대한 적용상수로서 α_0는 0.7, α_1은 0.3 을 사용하였다.

그림 2에서 볼 수 있듯이 본 논문에서 제시한 MAR 방법이 AR+ 방법보다 top-N 추천에서 항상 1.0%이상의 성능 향상을 보였다.

4. 결론

일반적으로 추천 시스템은 사용자의 수가 증가함에 따라 추천 시간이 증가하는 확장성의 문제와, 사용자 선호도 정보가 희소할 경우 추천 성능이 저하되는 데이터 희소성 문제를 안고 있다. 본 논문에서는 이러한 문제들을 해결하기 위하여 아이템간의 연관규칙을 아이템의 선호도 예측을 위한 모델로 이용함으로써

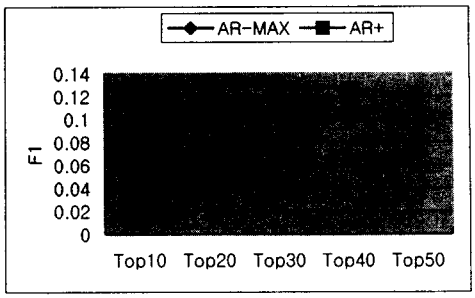

그림 1 연관규칙 적용 방법의 성능 비교

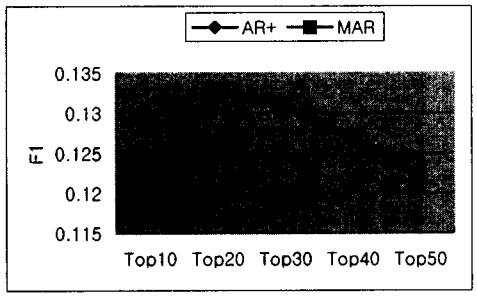

그림 2 카테고리 정보를 적용한 방법의 성능 비교

확장성 문제를 해결하였으며, 아이템의 카테고리 정보를 이용하여 다중레벨 연관규칙을 추천에 적용함으로써 추천 아이템의 범위를 확장하였다.

MovieLens 데이터 집합을 이용한 성능 실험을 통하여, 본 논문에서 제시한 다중레벨 연관규칙을 이용하는 방법이 단일레벨의 연관규칙만을 사용하는 것보다 효과적임을 알 수 있었다.

5. 참고문헌

[1] Breese, J.S., Heckerman, D., and Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 43-52, (1998)

[2] Han, J., and Kamber, M.: Data Mining: Concepts and Techniques, Morgan Kaurmann Publishers, (2000)

[3] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J.T.: Analysis of Recommendation Algorithms for E-Commerce. *ACM Conference on Electronic Commerce*, pages 158-167, (2000)

[4] Yang, Y., and Liu, X.: A re-examination of text categorization methods. *Proceeding of SIGIR-99*, (1999)

[5] http://www.cs.umn.edu/Research/GroupLens/index.html