구매 데이터에 적합한 아이템 기반의 협력적 추천 기법

김완섭[0] 윤찬식 이수원
숭실대학교 컴퓨터학과 인공지능연구실
wskim92@valentine.ssu.ac.kr[0], chansik@valentine.ssu.ac.kr, swlee@computing.ssu.ac.kr

An Item-based Collaborative Recommendation Algorithm for Purchase Data

Wan-Seop Kim[0] Chan-Sik Yune Soowon Lee
Dept. of Computer Science, Soongsil University

요 약

협력적 추천 알고리즘의 성능향상을 위한 많은 연구들이 진행되고 연구 결과로 다양한 협력적 추천 기법들이 제안되고 있다. 이러한 연구에서는 EachMovie, MovieLens등의 선호도(Rating) 값을 기반으로 하는 데이터를 대상으로 추천의 효율을 높이고자 하고 있다. 그러나 실세계에서 우리가 얻을 수 있는 원 거래 데이터(Raw Transaction Data)는 선호도 값을 갖고 있지 않다. 따라서 실세계의 구매 데이터에 효과적인 추천을 하기 위해서는 기존의 선호도 기반 알고리즘이 아닌 구매 정보만을 기반으로 하는 변경된 협력적 추천 알고리즘이 필요하다. 본 논문에서는 연관규칙 탐사 기법에서 사용하는 확신도(confidence)를 유사도식에 사용하고 이를 기반으로 선호도를 예측하는 구매 기반의 협력적 추천 알고리즘을 제안한다.

1. 서 론

인터넷 전자상거래가 활성화되면서 고객의 만족도를 충족시키고 구매율을 높이고자하는 요구와 관심이 커지고 있다. 이를 위해 각 고객의 취향을 분석하여 개인화된 서비스를 제공하여 고객의 만족도를 높이려는 연구가 이루어지고 있다. 개인화된 서비스에서 가장 대표적인 서비스로는 추천 시스템을 들 수 있다.

추천 시스템은 각 고객의 취향에 맞는, 구매 가능성이 높은 상품을 추천하는 시스템으로 다양한 알고리즘이 활용되고 있고 가장 대표적인 추천 알고리즘이 협력적 여과(Collaborative Filtering)에 의한 추천 기법이다. 최근 수년간 협력적 여과 추천 기법의 성능을 높이기 위한 많은 연구가 활발히 진행되고 있으나 기존의 협력적 여과 기법을 실세계의 구매 데이터(예: 쇼핑몰 Transaction DB)에 적용하여 추천하고자 할 경우 적합하지 않는 경우가 많다. 왜냐하면 기존의 추천 알고리즘은 사용자의 각 아이템에 대한 선호도 정보(Rating Value)를 가지고 있는 반면, 실세계의 데이터에는 이러한 정보가 매우 미흡하기 때문이다. 즉, 연구에서 사용하는 데이터(예: EachMovie, MovieLens 등)에는 각 아이템에 대한 고객의 선호도를 구체적인 수치로 가지고 있으나 실제 거래에서 얻을 수 있는 정보는 어떤 상품을 구매했는가, 구매했다면 몇 번 구매했는가에 대한 정보일 뿐이다. 구매횟수로 선호도를 유추할 수도 있으나 실제 데이터에서 구매횟수는 아이템의 성격에 따라 좌우되는 것을 알 수 있다. (예를 들어, 식료품, 화장품 등의 아이템은 구매횟수가 많지만 가전제품의 경우 구매횟수가 적다.) On-line 쇼핑몰에서는 클릭스트림 분석(clickstream analysis)등을 통하여 간접적인 방법으로 고객의 선호도를 유추할 수도 있지만 이 과정에서 불충분한 데이터로 인하여 고객의 선호도를 정확히 이끌어 내지 못하고, 이로 인해 적합한 추천 결과를 제공하지 못한다. 또한 이러한 전처리에 해당하는 작업은 적용하는 데이터에 따라 매우 다양하고 주관적이다. 본 연구에서는 구매 데이터에 대하여 데이터마이닝의 주요 기법인 연관규칙 탐사 기법의 확신도(confidence)를 유사도로 사용하고, 이를 조합하여 선호도를 예측하는 협력적 추천 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 선호도(Rating)에 기반하고 있는 협력적 추천 알고리즘에 대해 설명한다. 3장에서는 구매 정보 DB에 적합한 협력적 추천 시스템을 제안한다. 4장에서는 실제 구매 데이터에 대하여 기존의 알고리즘과 제안하는 알고리즘의 비교 실험을 하였다. 5장에서는 실험평가와 결론을 내린다.

2. 관련 연구

2.1 협력적 추천 기법

협력적 추천 기법은 크게 두 가지로 나뉠 수 있다. 첫째는 사용자 기반 협력적 추천 기법(User-based CF)이며, 둘째는 아이템 기반 협력적 추천 기법(Item-based CF)이다. 사용자 기반 여과 추천 기법은 추천의 대상이 되는 고객에 대하여 그와 비슷한 취향을 갖는 유사 고객군(Neighbors)을 찾고, 이들 유사 고객들이 공통적으로 많이 구매하는 아이템들 중에서 추천 대상 고객이 구매하지 않은 아이템을 추천해 주는 기법이다. 그러나 사용자 기반 기법의 경우에 모델(Model)이 없이 고객별로 항상 모든 고객을 대상으로 가장 유사한 고객을 찾는 과정을 거치므로 전체 고객이 많을 경우 매우 비효율적이다. 또한 고객이 다양한 관심분야를 갖는 경우에 대해서는 적합한 추천을 해 주지 못하므로 한계가 있다. 반면, 아이템 기반 추천 기법은 먼저 아이템들간의 유사도를 계산하여 모델(Model)을 형성하고 이것을 특정 고객에 대한 추천 시에 사용하므로 사용자 기반 추천 기법에 비해 효율적이다. 또한 다양한 사용자 관심분야를 반영하므로 추천 정확도를 높인다[4][5].

2.2 Item-based 협력적 추천에서의 유사도 계산

아이템 기반 협력적 추천 기법은 아이템 간의 유사도를 구하는 부분과 계산된 유사도를 사용하여 추천하는 부분으로 나누어진다. 아이템 간의 유사도를 구하는 대표적인 방식은 코사인식과 피어슨 상관계수식이다.

2.2.1 코사인(Cosine)

두 아이템 i, j 간의 유사도를 벡터 기반의 코사인(Cosine) 방식으로 구하는 식은 아래와 같다.

$$sim(i,j) = cos(\overrightarrow{i,j}) = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\|_2 * \|\vec{j}\|_2} \qquad (식1)$$

위 식에서 $\vec{i}$, $\vec{j}$ 는 두 아이템 i , j를 고객들이 준 선호도를 내용으로 하는 벡터로 표현한 것이다. 따라서 두 벡터는 고객의 수를 차원으로 하고 고객의 선호도(Rating)를 값으로 갖는다.

2.2.2 피어슨 상관계수 (Pearson correlation)

두 아이템 i, j의 유사도를 피어슨 상관계수 (Pearson Correlation)로 구하는 식은 아래와 같다.

$$sim(i,j) = corr(i,j) = \frac{\sum(R_{u,i} - \overline{R_i})(R_{u,j} - \overline{R_j})}{\sqrt{\sum(R_{u,i} - \overline{R_i})^2 \sum(R_{u,j} - \overline{R_j})^2}} \qquad (식2)$$

위 식에서 $R_{u,j}$ 는 고객 u의 아이템 i에 대한 선호도 값이고, $\overline{R_i}$는 아이템 i 에 대해 고객들이 준 선호도 값들의 평균이다.

2.3 Item-based 협력적 추천에서의 예측 기법

기존의 아이템 기반 협력적 추천에서 사용하는 예측식은 아래와 같다. 고객이 입력한 선호도를 위(식1)(식2)에서 얻은 아이템 간의 유사도를 가중치로 하여 더하는 방식이다.

$$P_{u,i} = \frac{\sum_{allSimilarItems,N}(S_{i,N} * R_{u,N})}{\sum_{allSimilarItems,N}(|S_{i,N}|)} \qquad (식3)$$

위 식에서 $P_{u,i}$ 은 고객 u의 아이템 i 에 대한 예측 선호도이다. $S_{i,N}$은 아이템 i 와 다른 아이템 N과의 유사도 값이고, $R_{u,N}$은 고객 u가 아이템 N에 대하여 준 선호도(Rating) 값이다. $|S_{i,N}|$ 는 $S_{i,N}$의 절대값을 의미한다.

3. 구매 데이터에 적합한 추천 기법

3.1 선호도 데이터와 구매 데이터의 차이점

EachMovie, MovieLens등의 Rating 기반의 Data의 경우 고객의 아이템에 대한 선호도(like), 비선호도(dislike) 정보를 수치적으로 가지고 있다. 반면 실세계에서 발생하는 Data를 통해서는 구매 정보를 통해 고객의 아이템에 대한 선호가 있음을 간접적으로 파악할 수 있으나 얼마나 좋아하는지의 정보를 알 수 없다. 또한 비선호에 대한 정보는 매우 미흡하다. 구매한 아이템에 대해서는 고객이 선호도를 가지고 있다고 볼 수 있지만, 구매하지 않은 것에 대해서 고객이 명시적으로 싫어한다고 표현한 것이 아니기 때문에 비선호도를 가진다고 볼 수 없다. 단지 비선호의 가능성이 있을 뿐이며, 이와 동일하게 선호의 가능성도 가지고 있다.

3.2 기존 협력적 추천 기법의 한계

	상품A	상품B	상품C	상품D
고객1	구매	구매		구매
고객2		구매	구매	
고객3	구매	구매		
고객4			구매	
고객5		구매		구매
고객6			구매	
고객7		구매		

[표1]

	상품A	상품B	상품C	상품D
고객n1	구매	$?_1$		
고객n2	$?_2$	구매	구매	
고객n3		$?_3$		구매

[표2]

기존 협력적 추천 기법의 한계를 설명하기 위해 간단한 구매데이터의 예를 들었다. [표1]은 7명의 고객이 4개의 아이템에 대하여 구매한 내역에 대한 데이터 예이고, [표2]는 추천을 하고자 하는 고객의 구매내역 예이다. 물음표(?) 부분에 대하여 선호도를 예측하고자 한다.

앞에서 언급한 협력적 추천의 대표적인 유사도 계산식인 코사인(식1)과 상관계수(식2) 방식은 선호도 기반의 데이터에서 좋은 예측률을 보인다[4]. 그러나 위의 [표1]의 구매 데이터에 적용한다면 좋은 예측을 보이지 못한다. 위의 [표1]은 모델형성에 사용할 구매 데이터이고 [표2]는 추천하고자 하는 고객들의 구매 데이터이다. 추천 대상 고객의 데이터에서 물음표(?) 표시부분은 [표1]을 통해 유추해 보면 모두 구매 가능성이 높은 것들이다. 기존의 방식을 적용하여 [표1]의 구매 데이터로 아이템간의 유사도를 구하면 코사인 유사도식을 통해서는 [표3], 피어슨 상관계수 유사도식을 통해서는 [표4]의 아이템간의 유사도값을 얻게된다.

	상품A	상품B	상품C	상품D
상품A	1	0.63	0	0.5
상품B	0.63	1	0.26	0.63
상품C	0	0.26	1	0
상품D	0.5	0.63	0	1

[표3]코사인 유사도

	상품A	상품B	상품C	상품D
상품A	1	0.54	-0.53	0.31
상품B	0.54	1	-0.32	0.54
상품C	-0.53	-0.32	1	-0.53
상품D	0.31	0.540	-0.53	1

[표4]피어슨 상관계수 유사도

[표1]에서 상품 A를 구매한 고객은 상품 B를 구입하는 성향을 강하게 보이고, 반면 상품 B를 구입했을 때 상품 A를 구매하는 성향은 약하다. 그러나 위 결과[표2] 및 [표4]에서는 Sim(A,B) = Sim(B,A) = 0.632, Sim(A,B) = Sim(B,A) = 0.64의 결과로 이러한 구매의 패턴을 나타내지 못한다. 즉, 기존 유사도 계산식은 위[표1]과 같은 구매 데이터에 반영할 경우 구매 패턴의 성격을 잘 반영하지 못하는 한계가 있다.

위에서 얻은 유사도 값을 가지고 [표2]의 데이터에 (식3)의 예측식으로 예측 선호도 값을 구해보면 코사인식에서는 $<?_1, ?_2, ?_3> = <0.42, 0.56, 0.42>$의 예측 선호도값을 얻게된고, 피어슨 상관계수식을 통해서는 $<?_1, ?_2, ?_3> = <0.39, 0.01, 0.39>$의 예측 선호도 값을 얻는다. 즉 예측 선호도 값이 높지 않음을 볼 수 있다.

3.3 구매 확률에 기반한 유사도 측정 방법

구매 데이터에서 방향성이 분명히 존재하지만 기존의 협력적 여과 추천 기법에서 사용하는 유사도 계산방법 즉, 벡터 기반의 Cosine 계산식(식5)과, Pearson Correlation 계산식(식6)에서는 Sim(A,B) = Sim(B,A)이므로 이러한 구매의 방향성 정보를 표현하지 못한다. 이러한 한계는 데이터마이닝의 연관규칙 기법에서 사용하는 확신도(confidence)를 사용하여 해결할 수 있다.

$$sim(A,\ B) = confidence(A, B) = \frac{purchase(A \cap B)}{purchase(A)} \quad (식4)$$

(식4)에서 $Purchase(A)$는 아이템 A를 구매한 고객의 수를 의미하고, $Purchase(A \cap B)$는 아이템 A도 구매하고 B도 구매한 고객의 수를 의미한다. 이 유사도 계산식에 의하여 $sim(A,B) \neq sim(B,A)$인 A,B 간의 유사도가 생성되어, 기존의 유사도 계산 방법에서 얻을 수 없는 방향성 정보를 얻게 됨으로써 더욱 의미 있는 추천이 가능하다.

3.4 구매 확률 기반의 예측 기법

(식4)에서 얻은 유사도 값을 활용하여 추천을 하고자 할 경우 이에 적합한 새로운 예측 기법이 요구된다. 본 연구에서는 아래와 같은 새로운 예측식을 제안한다.

$$P_{a,j} = \sqrt{\frac{\sum_{i} confidence(i,j)^2}{N}} \quad (식5)$$

위 식에서 Pa,j는 고객 a가 아이템 j를 구매할 가능성을 의미한다. $confidence(i,j)$은 위에서 유사도식으로 제안한 확신도 값으로 아이템 i를 구매한 경우 아이템 j를 구매할 확률을 의미한다. 고객 a가 구매한 모든 아이템 i에 대해서 $confidence(i,j)$를 계산에 사용한다. N은 고객이 구매한 아이템의 개수이다.

(식5)는 고객 a가 구매한 아이템(i)들에 대하여 $confidence(i,j)$를 큰 값에 가중치를 주어 평균하는 계산식이다.

(식5)에서 큰 확신도 값에 가중치를 주는 이유는, 어떤 고객에게 T라는 아이템을 추천하고자 할 경우 이 고객이 confidence(A,T) = 1(100%)인 아이템 A를 구매했다고 하면 그 고객에게 T를 높은 예측값으로 추천할 수 있다. 만약 이 고객이 A 외에도 confidence(B,T) = 0.1, confidence(C,T)=0.1인 아이템, B, C를 구매했다고 하더라도 여전히 T를 구매할 가능성이 높음을 예측할 수 있다. 따라서 예측하고자 하는 아이템에 대하여 구매 확신도 값이 여러 개일 경우, 이 값들 중 큰 값에 가중치를 주어 평균하는 것이 필요하다.

3.2절의 데이터[표1],[표2]에 대하여 유사도 계산식으로 (식4)를, 예측식으로 (식5)를 적용하면 <?1 , ?2 , ?3> = <1, 0.71, 1>의 높은 예측 선호도 값을 얻을 수 있다.

4. 실험

4.1 실험 데이터

실험을 위한 데이터로 백화점 거래 데이터를 사용하였다. 이 데이터는 1999년1월1일부터 2001년 12월 31일까지의 3년간의 거래 내역을 가지고 있다. 총 508개의 아이템에 대한 3194명의 고객의 구매 내역으로 구성되어 있다. 전체 데이터 중 70%를 훈련 데이터로 모델 형성에 사용하였고 30%를 검증 데이터로 사용하였다.

4.2 실험

4.2.1 실험 평가 방법

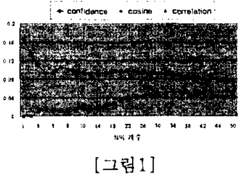
실험의 평가 방법으로는 정확도(Precision)과 재현률(Recall)의 두 평가치의 값을 하나로 평가하는 F1 측정방법을 사용하였다. 정확도는 topN에 의해 추천된 아이템 중 실제 구매된 아이템의 비율을 의미하고, 재현률은 고객이 실제로 구매한 아이템들 중에서 topN에 의해 추천된 아이템의 비율을 의미한다. 정확도(식6)와 재현률(식7), F1측정치(식8)의 수식은 아래와 같다.

$$precision = \frac{|test \cap topN|}{} \quad (식6) \qquad recall = \frac{|test \cap topN|}{} \quad (식7)$$

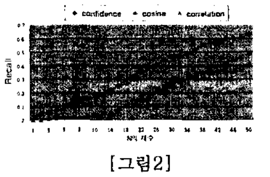
$$F1 = \frac{2*recall*precision}{} \quad (식8)$$

실험에서는 테스트 데이터의 각 고객에 대해서 F1값을 계산하였고, 모든 고객의 F1값을 평균하여 최종적인 F1값을 구하였다. 실험에서는 topN 추천에서 N을 1부터 50까지 변경시키며 추천하였을 때의 정확도와 재현률을 구하고 이를 이용하여 F1값을 구하였다.

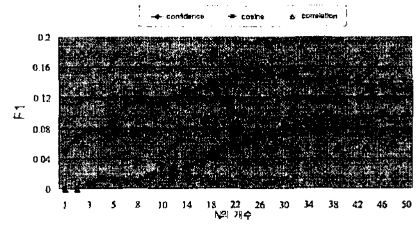
4.2.2 알고리즘 별 실험 결과



[그림1]



[그림2]



[그림3]

본 연구에서는 세가지의 알고리즘을 비교하였다. 첫 번째 cosine 방식은 유사도로 코사인식을 사용하고 (식3)을 예측식으로 사용하였고, 두 번째 correlation방식은 유사도로 피어슨 상관계수식을 (식3)을 예측식으로 사용한 협력적 추천 방식을 사용하였다. 마지막 세 번째로 본 논문에서 제안한 confidence 방식은 유사도로 연관규칙 기법의 확신도(식4)를 사용하고 예측식으로 (식5)를 사용하는 협력적 추천 방식의 세가지 방식을 실험에서 비교하였다. [그림1],[그림2],[그림3]은 각각 추천에 대한 정확도, 재현률, F1 값에 대한 비교 결과이다. 이 결과는 코사인 방식과 상관계수방식의 경우에 비해 제안하는 확률기반의 협력적 추천기법이 높은 F1값을 보이므로 구매 데이터에서 좋은 예측 성능을 보임을 알 수 있다.

5. 결론

기존 아이템 기반 협력적 여과 추천 기법은 선호도를 기반으로 하고 있고, 선호도 정보 데이터에 적용할 경우 의미있는 추천 결과를 얻을 수 있다. 그러나 선호도 정보가 충분하지 않은 구매 데이터에 이 방식을 사용할 때 추천 효율이 좋지 않다. 본 논문에서는 유사도 계산에서 아이템 간의 확신도(confidence)를 사용하였고 이를 바탕으로 새로운 예측식을 제안하여 기존의 알고리즘에서 나타나는 한계를 극복하여 구매 데이터에서의 추천 효율을 향상시켰다.

6. 참고문헌

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

[2] D. Billdle and M. J. Pazzni. "Learning Collaborative Information filters", In Proceedings of ICML, pages46-53, 1998.

[3] P. Resnick, et. al., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews", Proc. of ACM CSCW, 1994.

[4] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl., "Item-Based Collaborative Filtering Recommendation Algorithms", 2000

[5] G. Karyis "Evaluation of Item-Based Top-N Recommendation Algorithms", 2000.

[6] M. Nichols "Implicit Rating and Filtering", In Proceeding of the 5th DELOS Workshop on Filtering and Collaborative Filtering, 1997

[7] B. Sarwar, G. Karypis, J. Riedl "Recommender Systems in e-Commerce", In proceedings of ACM E-Commerce, 2000.