협동적 필터링을 위한 동시출현빈도 사용의 제한 피어슨 알고리즘

김진상*, 윤병주*
*명지대학교 컴퓨터공학부
e-mail:{jskim72,yoonbj}@mju.ac.kr

A Constrained Pearson Algorithm that uses Co-occurrence for Collaborative Filtering

Jin-Sang Kim*, Byong-Joo Yoon*
*Dept of Computer Engineering, Myong-Ji University

요약

최근 전자상거래 시스템에서 구매 촉진을 위해 사용하고 있는 핵심기술은 고객들로부터 얻어진 구매 정보를 기초로 고객이 좋아할 만한 제품을 예측하여 고객에게 정보를 제공하는 추천시스템이다. 이러한 추천시스템을 위한 추천알고리즘으로서 협동적 필터링(collaborative filtering) 알고리즘이 많이 사용되고 있다. 이 논문에서는 기존의 협동적 필터링 알고리즘의 성능을 향상시킨 동시출현 빈도 개념 알고리즘과 제한 피어슨 알고리즘을 접목시켜서, 사용자 선호도의 예측 정확도를 좀 더 향상시킬 수 있는 새로운 방법을 제안하고, 실험을 통해서 제안한 방법의 예측 정확도의 우수성을 증명하였다.

1. 서론

최근 전자상거래 시스템에서 구매 촉진을 위해 사용하고 있는 핵심기술은 추천시스템(recommender systems)이다. 이러한 추천시스템을 위한 추천알고리즘으로서 제품의 특징에 관계없이 제품을 구입한 고객들 사이의 관계에 기초하여 추천을 제공하는 협동적 필터링(collaborative filtering)을 이용하는 추천알고리즘이 많이 사용되고 있다.

하지만, 기존의 협동적 필터링은 모든 사용자간의 유사도 계산을 기본으로 하기 때문에, 전자상거래 사이트의 데이터 희소성(sparseness)으로 인해서, 실제로 적용하는 데에 시간적으로나 선호도 예측 정확도 면에서 비효율적이다. 따라서, 이러한 문제점을 개선하기 위한 연구로서, 동시출현 빈도 개념을 이용해서 계산 시간을 줄이는 동시에 선호도 예측 정확도를 높이기 위한 동시출현 빈도 개념 알고리즘이 발표되었다[3]. 또한, Social Information Filtering[2]에서는 음악 장르에 대한 사용자 평가값 중에서 사용자의 극단적인 평가자료(extreme value)에 대해서 제한 피어슨 알고리즘이 기존의 피어슨 알고리즘보다 더 높은 선호도 예측 정확도를 보였다

이 논문에서는 기존의 협동적 필터링 알고리즘의 성능을 향상시키기 위해서, 동시출현 빈도 개념 알고리즘과 제한 피어슨 알고리즘의 장점을 접목시킨 동시출현 빈도 개념의 제한 피어슨 알고리즘을 제안하고한다. 또한, 실험을 통해서 제안한 방법의 예측 정확도의 우수성을 증명한다.

2. 관련연구

2.1. 협동적 필터링(collaborative filtering)

협동적 필터링은 다른 사용자의 의견을 참고하여 사용자에게 제품을 추천하는 방법이다. 사용 가능한 아이템에 대한 사용자들의 의견을 데이터베이스로 구축하며, 특정 사용자의 특정 아이템에 대한 평가값을 예측할 때 데이터베이스에서 특정 사용자와 유

사한 사용자(유사그룹)들을 발견하며, 또한 평가 값을 유사 사용자의 의견을 참조하여 예측해 낸다.

최초의 자동화된 협동적 필터링 방법인 GroupLens[1]는 특정 사용자의 제품에 대한 평가를 예측하기 위하여 식(2.1)과 같은 예측 식을 사용한다.

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n} (r_{u,i} - \bar{r}_u) * w_{a,u}}{\sum_{u=1}^{n} w_{a,u}} \qquad (2.1)$$

식(2.1)에서, $w_{a,u}$는 사용자 a와 사용자 u사이의 유사도(similarity weight)를 나타내는데, 식(2.2)와 같이 피어슨 상관 계수로 정의된다.

$$w_{a,u} = \frac{\sum_{i=1}^{m} (r_{a,i} - \bar{r}_a) * (r_{u,i} - \bar{r}_u)}{\sigma_a * \sigma_u} \qquad (2.2)$$

단, σ_a는 사용자 a의 표준편차, σ_u는 사용자 u의 표준편차이다.

2.2. 동시출현 빈도 개념 알고리즘

기존의 협동적 필터링 알고리즘을 사용하는 추천 시스템은 모든 사용자간의 유사도 계산을 기본으로 하기 때문에 전자상거래 사이트의 데이터 희소성(sparseness)으로 인해 실제로 적용하는 데에 중요한 문제점이 있다. 이런 문제점을 해결하기 위해서 동시출현 빈도에 기반한 협동추천시스템의 성능 향상[3]에서는 동시출현 빈도를 이용해 공통으로 관심있는 항목이 많은 사용자들의 그룹을 형성하고, 그 그룹 안에서 우선적으로 유사도를 계산하여 계산 시간을 줄이며, 동시에 선호도 예측 정확도를 높이기 위해서 동시출현 빈도 개념 알고리즘을 제안하였다.

2.3. 제한 피어슨 알고리즘

Social Information Filtering[2]에서는 사용자들이 특별히 좋아하거나 또는 특별히 싫어하는 음악만을 평가하고, 자신이 좋아하는 음악을 다른 사용자들에게 추천하는 것과 관심이 있다는 것에 초점을 맞추어, 1에서 7사이의 평가점수에서 극단적인 평가자료(extreme value)이다. 즉, 사용자 평가 값이 사용자가 싫어하는 음악을 의미하는 2이하이거나 좋아하는 음악을 의미하는 6이상의 평가 값에 대해서 각 알고리즘을 평균에러를 기준으로 선호도 예측 정확도를 비교하였다.

실험 결과, 사용자의 극단적인 평가자료에 대해서 기존의 피어슨 상관계수 식을 수정한 새로운 상관계수 식인 식(2.3)을 사용하는 제한 피어슨 알고리즘이 다른 알고리즘보다 예측 정확도가 더 높은 것으로 나왔다

$$\beta_{xy} = \frac{\sum (Ux-4)(Uy-4)}{\sqrt{\sum (Ux-4)^2 \times \sum (Uy-4)^2}} \qquad (2.3)$$

여기서, Ux는 사용자 x의 평가값, Uy는 사용자 y의 평가값이며, 4는 평가 범위의 중간값이다.

3. 예측 정확도 향상을 위한 방법 제안

3.1 기존 방법들의 문제점

기존의 협동적 필터링 알고리즘의 문제점을 개선한 동시출현 빈도 개념 알고리즘에서 공통으로 평가한 항목이 많은 사용자들은 비슷한 취향을 가졌으므로, 상관 관계가 높을 것이라고 가정하였다. 하지만, 반드시 그렇지는 않을 것이다. 공통으로 평가한 항목이 많은 사용자들의 유사도가 실제로 높지 않아도, 동시출현 빈도 개념 알고리즘에 의해서 비슷한 취향을 가진 유사 사용자 그룹을 형성해서, 잘못된 추천을 초래할 수 있다.

또한, Social Information Filtering의 제한 피어슨 알고리즘에서 사용자가 공통으로 평가한 항목 수에 관계없이, 많은 항목을 공통으로 평가한 사용자와 공통으로 평가한 항목이 적은 사용자에게 같은 가중치를 주어서, 공통으로 평가한 항목의 수가 다른 사용자들을 차별화 하지 못했다.

3.2 새로운 방법 제안

이 논문에서는 3.1절에서 제기한 동시출현 빈도 개념 알고리즘과 제한 피어슨 알고리즘 각각의 문제점을 해결하기 위해서, 동시출현 빈도 개념 알고리즘에서 제안한 공통으로 평가한 항목이 많은 사용자들의 평가데이터만을 사용하는 방법과, 제한 피어슨 알고리즘에서 제안한 사용자 사이에 높은 양의 상관관계를 가지는 사용자의 평가데이터만을 선호도 예측에 사용하는 방법을 접목시킨, 동시출현 빈도 개념의 제한 피어슨 알고리즘을 제안한다.

제안하는 방법은 동시출현 빈도 개념 알고리즘을 적용하며, 사용자가 평가한 항목 중에서 두 사용자가 동시에 평가한 항목이 전체 평가항목 수의 70% 이상인 유사 사용자 그룹을 형성하고, 공통으로 평가한 항목이 많은 유사 사용자 그룹에 속한 사용자들을 대상으로 제한 피어슨 알고리즘의 새로운 상관계수 식을 사용하며, 사용자 사이의 상관계수를 계산하고, 새로운 상관계수의 절대치가 특정 임계치(threshold)이상인 사용자의 평가데이터만을 사용해서 선호도를 예측하는 것이다.

선호도를 예측하려는 사용자와 공통으로 평가한 항목이 적은 사용자를 유사도 계산에서 제외시키고, 공통으로 평가한 항목이 많은 사용자 중에서도 높은 양의 상관관계를 가진 사용자들의 평가데이터만을 사용하여 선호도를 예측하므로, 선호도 예측 정확도가 향상될 것으로 기대된다.

동시출현 빈도 개념의 제한 피어슨 알고리즘은 다음과 같다.

1. 선호도를 예측하려는 사용자와 전체 항목 수의 70% 이상의 항목을 공통으로 평가한 사용자 그룹인 G_u를 형성한다.
2. G_u의 사용자를 대상으로 식(3.1)을 사용해서 선호도를 예측하려는 사용자와 다른 모든 사용자들 사이에 새로운 상관 계수 β_{iu}를 구한다.

$$\beta_{iu}=\frac{\sum(U_i-3)(U_u-3)}{\sqrt{\sum(U_i-3)^2\times\sum(U_u-3)^2}} \qquad (3.1)$$

여기서, β_{iu}는 사용자 i와 다른 사용자들 u사이의 상관계수이고, 3은 1에서 5사이의 평가 범위의 중간 값이다.

3. $|\beta_{iu}|$가 임계치(threshold) L이상인 모든 사용자들을 구한다. 이 사용자들은 선호도를 예측하려는 사용자와 유사한 사용자들의 그룹인 N_u를 구성한다.
4. 식(3.2)를 사용해서 각 사용자의 가중치(weight) w_{iu}를 구한다.

$$w_{iu}=\left(\frac{\beta_{iu}-L}{1-L}\right)^2,\ where\ |w_{iu}|\le 1 \qquad (3.2)$$

여기서, β_{iu}는 사용자 i와 u사이의 상관계수이며, L은 임계치이다.

5. w_{iu}를 각 사용자의 가중치로 사용해서 식(3.3)을 이용해서 선호도를 예측한다.

$$P_{ij}=\overline{U_i}+\frac{\sum_k^{N_i} r_{ku}\times s_{kj}}{\sum_k^{N_i} r_{ku}} \qquad (3.3)$$

여기서, $\overline{U_i}$는 사용자 i의 평가 평균값, $\sum_k^{N_i} r_{ku}$는 사용자 k와 다른 사용자들 u와의 가중치의 합이며, s_{kj}는 사용자 k의 항목 j에 대한 평가값이다.

3.3 실험 및 토의

3.3.1 실험 환경과 test data sets

이 논문에서 제안한 방법은 Java언어로 구현되어 운영체제에 관계없이 실제 전자상거래 시스템에서도 수행 가능하도록 하였으며, PentiumIII 433Mhz, 128MB RAM의 컴퓨터 환경에서 실험하였다.

실험에 사용된 data set은 GroupLens Research Project[4]에서 제공한 MovieLens data set에서 일부를 뽑아 사용하였다. MovieLens data set은 총 943명의 사용자가 1682개의 영화에 대해서 1에서 5점 사이의 점수로 평가한 100,000개 데이터이다. Data set은 최소한 20개 이상의 영화에 대해서 평가한 사용자 데이터만으로 구성되며, 영화는 19개 장르로 구분된다. 총 100,000개의 data set중에서 80,000개를 뽑아 training data로 사용하고 20,000개를 test data로 사용하도록 하고 있다.

3.3.2 예측 정확도 비교

<표 3.1>은 기존의 동시출현 빈도 개념 알고리즘과 제한 피어슨 알고리즘 그리고 이 논문에서 제안한 동시출현 빈도 개념의 제한 피어슨 알고리즘을 각각 5번 실험한 결과의 평균을 보여준다.

알고리즘	극단적인 평가자료		
	평균에러	예측범위 (%)	예측시간 합계(ms)
동시출현 빈도 개념 알고리즘	1.559634	89.96	345586
제한 피어슨 알고리즘	1.327746	90.7	349547
동시출현 빈도 개념의 제한 피어슨 알고리즘	1.290772	88.786	338349.4

<표 3.1> 기존 알고리즘과 제안한 알고리즘의 실험 결과 비교
평균에러(MAE : Mean Abstract Error)

<표 3.1>의 극단적인 평가자료에 대한 실험 결과에서, 평균에러는 동시출현 빈도 개념 알고리즘의 평균에러가 1.559이며 제한 피어슨 알고리즘의 평균에러는 1.327로, 제한 피어슨 알고리즘이 훨씬 낮은 평균에러를 보였고, 제안한 동치출현 빈도 개념의 제한 피어슨 알고리즘은 기존의 제한 피어슨 알고리즘보다 더 낮은 1.290의 평균에러를 보였다.

예측범위는 제안한 방법이 기존 알고리즘보다 예측에 사용되는 평가데이터를 축소시켜서, 기존 알고리즘보다 더 낮은 예측범위를 보였고, 제안한 방법이 기존 알고리즘 보다 계산 량이 감소되어서, 예측시간이 짧아진 것을 알 수 있다.

이 논문에서 제안한 동시출현 빈도 개념의 제한 피어슨 알고리즘은 선호도를 예측하려는 사용자의 영화와 같은 장르에 속하는 영화수의 70%이상의 영화를 공통으로 평가한 비슷한 취향을 가진 사용자 그룹에서도 상관계수의 절대치가 임계치 이상인 유사 사용자의 평가데이터만 사용하기 때문에, 예상대로 기존의 두 알고리즘보다 더 낮은 평균에러를 보여서, 사용자가 아직 평가하지 않은 특정한 영화를 얼마나 좋아할 것인가를 예측하는 선호도 예측 정확도가 향상된 것을 알 수 있다.

4. 결론

기존의 협동적 필터링(automated collaborative filtering) 알고리즘의 문제점을 해결하기 위해서 많은 연구가 진행되어 왔다.

동시출현 빈도에 기반한 협동추천시스템의 성능 향상에서는 동시출현 빈도 개념을 이용해 공통으로 관심있는 항목이 많은 사용자들의 그룹을 형성하고 그 그룹 안에서 우선적으로 유사도를 계산하여 계산시간을 단축하려고 하였으며, Social Information Filtering에서는 사용자 사이의 상관계수에 임계치를 적용해서 선호도를 예측하려는 사용자와 양의 상관관계를 가지는 사용자 그룹의 평가데이터만 사용해서 선호도를 예측하려고 하였다.

이 논문에서는 동시출현 빈도 개념 알고리즘과 제한 피어슨 알고리즘의 평균에러 측면의 장점을 접목시켜서, 사용자가 평가한 항목 중에서 두 사용자가 동시에 평가한 항목이 전체 평가항목수의 70%이상인 사용자 그룹을 대상으로 유사도를 계산하고, 공통으로 평가한 항목이 많은 사용자 그룹에서도 사용자 사이의 상관계수의 절대치가 임계치(threshold)이상인 양의 상관관계를 가지는 사용자의 평가데이터만을 사용해서 선호도를 예측하는 동시출현 빈도 개념의 제한 피어슨 알고리즘을 제안하였다.

실험을 통해서 기존의 알고리즘과 제안한 알고리즘의 선호도 예측 정확도를 비교한 결과, 제안한 알고리즘이 기존의 알고리즘보다 예측시간을 단축시킬 수 있었고, 특히 평균에러를 많이 줄일 수 있어서, 선호도 예측 정확도는 많이 향상되었지만, 예측범위는 다소 떨어지는 경향을 보였다.

앞으로는 제안한 알고리즘의 낮은 평균에러와 빠른 예측시간을 유지하면서 예측범위도 높일 수 있는 방법에 대한 연구가 필요하다.

참고문헌

[1] Rssnick P., Iacovou N., Sushak M., ergstrom P., and Riedl J., GroupLens : An peon architecture for collaborative filtering of Netnews, Proceedings of the 1994 ComputerSupported Collaborative Work Conference, 1994.

[2] Upendra Shardanand, Social Information Filtering for Music Recomendation, MIT EECS, TR-94-04, Learning and Common Sense Group, MIT Media Laboratory, 1994.

[3] 박지연, 박윤심, 유견아, 「동시출현 빈도에 기반한 협동추천시스템의 성능 향상」, 2000년 한국정보처리학회 추계 학술발표논문집, 제 7권 제 2호, pp. 333-336, 2000.

[4] http://www.grouplens.org/