협동적 필터링 기반 추천 시스템을 위한 향상된 이웃 선정 방법

김택헌, 양성봉
연세대학교 컴퓨터과학·산업시스템공학과
e-mail:kimthun@cs.yonsei.ac.kr

An Improved Neighbor Selection Method for Recommender Systems based on Collaborative Filtering

Taek-Hun Kim, Sung-Bong Yang
*Dept. of Computer Science and Industrial System Engineering,
Yonsei University

요 약

전자상거래에서 추천 시스템은 일반적으로 협동적 필터링이라는 정보 필터링 기술을 사용한다. 협동적 필터링 기술은 유사한 성향을 갖는 다른 고객들이 상품에 대해서 매긴 평가에 기반한다. 협동적 필터링이 유사 선호도를 갖는 이웃 고객들의 평가에 근거하기 때문에 고객에게 가장 적합한 유사 이웃들을 적절히 선정해 내는 것은 추천 시스템에서 예측의 질 향상을 위해 필요하다.

본 논문에서 우리는 ordered clustering을 이용하여 협동적 필터링을 위한 향상된 이웃선정 방법을 제안한다. 이 방법은 탐색 공간을 줄이기 위해 k-means 클러스터링 방법을 사용한다. 그리고 클러스터링에 의해 구성된 고객들에 대해서 threshold 값에 의해 보다 정제된 고객들을 최종 선정함으로써 고객에게 보다 의미 있는 적합한 고객이 최종적인 이웃으로 선정될 수 있도록 한다.

실험은 Compaq Computer Corporation에 의해 제공된 EachMovie 데이터 셋을 사용하였다. 실험 결과로 우리는 제안한 방법이 다른 방법보다 좋은 예측 정확도를 갖는 것을 확인할 수 있었다.

1. 서론

상거래 시장에서 취급되어지는 상품은 오프라인상에서 뿐만 아니라 온라인상에서도 이미 우리의 생각이 미치지 못할 만큼 그 종류가 매우 다양하고 수 또한 셀 수 없을 정도로 많다. 따라서 다양한 선호도를 갖는 고객들에게 그들이 원하는 상품을 찾기 위해 들이는 탐색 비용을 줄일 수 있게 하기 위해 더 좋은 가치를 갖는 양질의 정보를 제공하는 개인화된 추천 시스템의 개발이 필요하다.

추천 시스템은 고객의 선호도를 추출하고 분석하여 고객에게 적합한 상품을 정확하게 예측하여 추천해 줄 수 있어야 하며, 이를 위해 일반적으로 협동적 필터링 (CF: Collaborative Filtering)이라고 하는 정보 필터링 기술을 사용한다. 협동적 필터링은 유사한 선호도를 갖는 다른 고객들의 평가에 기반한 것으로 Amazon.com 및 CDNoow.com 등 상업적으로 성공한 전자상거래 사이트의 추천 시스템에 폭넓게 사용되고 있다[1][2].

협동적 필터링은 상품에 대한 고객들의 선호도 상관관계에 따른 고객들간 선호도의 유사도를 구하고 이를 예측식에 이용하여 상품에 대한 추천 여부를 결정한다. 협동적 필터링이 유사 선호도를 갖는 이웃 고객들의 평가에 근거하기 때문에 고객에게 가장 적합한 유사 이웃들을 적절히 선정해 내는 것은 추천 시스템에서 예측의 질 향상을 위해 필요하다.

이웃선정 방법은 모든 고객을 이웃으로 하여 고객에게 유용하지 않은 고객들 마저 이웃으로 삼는 전통적인 협동적 필터링의 단점을 보완하여 예측의 정확도를 높일 수 있게 하는 방법이다. 이웃선정 방법으로는 k-최대근접 방법과 클러스터링 방법이 많이 쓰이며, 특히 클러스터링을 이용한 방법은 대규모의 데이터 셋으로부터 빠른 추천이 이루어질 수

있도록 하는 방법으로서 큰 의미를 갖고 있다[4][6].

본 논문에서 우리는 순서화된 클러스터링 방법을 이용하여 협동적 필터링을 위한 향상된 추천 방법을 제안한다. 이 방법은 탐색 공간을 줄이기 위해 k-means 클러스터링 방법을 사용한다. 그리고 고객에게 가장 적합한 클러스터들을 순서적으로 구성하여 상위 클러스터들을 통합하여 고려함으로써 의미있는 고객들이 후보 이웃으로 제대로 선정될 수 있도록 한다.

그러나 우리가 고객 선호도를 보다 정확히 예측하기 위해서는 고객 각각에 대해서 높거나 혹은 낮은 유사 선호도를 모두 고려해야 한다. 이것은 유사도가 서로 반대인 고객도 선호도를 예측하는데 가치있는 정보를 줄 수 있기 때문이다. 따라서 클러스터링에 의해 구성된 이웃 고객들에 대해서 threshold 값에 의해 보다 정제된 고객들을 이웃으로 최종 선정함으로써 고객에게 보다 의미 있는 적합한 고객이 최종적인 이웃으로 선정될 수 있도록 한다.

실험은 Compaq Computer Corporation에 의해 제공된 EachMovie 데이터 셋을 사용하였다[5]. 실험결과는 제안한 추천 방법이 다른 방법들 보다 더 좋은 예측 정확도를 제공하는 것을 보여준다.

본 논문의 구성은 다음과 같다. 2장과 3장에서 협동적 필터링과 클러스터링 기반 이웃선정 방법에 대해서 설명한다. 4장에서 threshold를 적용한 ordered clustering 방법을 나타내고, 5장에서 실험결과를 보인다. 마지막으로 6장에서 결론을 맺는다.

2. 협동적 필터링

협동적 필터링은 각 아이템에 대한 고객의 선호도로부터 고객의 프로파일을 생성함으로써 아이템을 추천한다. 협동적 필터링에서 선호도는 일반적으로 고객에 의해 평가된 수치 값으로 표현된다. 테스트 고객에게 어떤 새로운 아이템에 대한 선호도를 예측하는 것은 타겟 아이템에 대한 다른 고객(이웃)들의 평가에 기반 한다.

협동적 필터링에서 식1은 고객의 선호도를 예측하기 위해 사용된다. 여기에서 $w_{a,k}$는 피어슨 상관계수를 나타낸 것으로 식2에서 주어진 것처럼 유사 가중치를 말한다[2][3].

$$p_{a,i} = \overline{r_a} + \frac{\sum_k \{w_{a,k} \times (r_{k,i} - \overline{r_k})\}}{\sum_k |w_{a,k}|}. \qquad (1)$$

$$w_{a,k} = \frac{\sum_j (r_{a,j} - \overline{r_a})(r_{k,j} - \overline{r_k})}{\sqrt{\sum_j (r_{a,j} - \overline{r_a})^2 \sum_j (r_{k,j} - \overline{r_k})^2}}. \qquad (2)$$

위 식에서 $p_{a,i}$는 아이템 i에 대한 고객 a의 선호도를 나타낸다. r_a와 r_k는 고객 a의 평가와 고객 k의 평가에 대한 평균을 각각 나타낸다. $r_{k,i}$와 $r_{k,j}$는 아이템 i와 j에 대한 고객 k의 평가를 각각 나타내고, $r_{a,j}$는 아이템 j에 대한 고객 a의 평가를 나타낸다.

만약 고객 a와 k가 한 아이템에 대해서 유사한 평가를 가진다면, $w_{a,k}>0$이다. $|w_{a,k}|$는 두 고객이 이미 평가한 아이템에 대해서 고객 a가 고객 k에 얼마나 동의하는지를 나타낸다고 할 수 있다. 만약 두 고객이 한 아이템에 대해서 반대되는 평가를 내렸다면, $w_{a,k}<0$ 이고 $|w_{a,k}|$는 그들이 같은 아이템에 대해서 얼마나 동의하지 않는지를 나타낸다고 할 수 있다. 그러므로 만약 그들이 상관성이 없다면, $w_{a,k}=0$이다. $w_{a,k}$는 -1에서 1 사이의 값을 가진다.

협동적 필터링은 고객과 비슷한 선호도를 갖는 다른 고객들의 평가를 기반으로 하기 때문에 추천 시스템에 적합하다. 그러나 비록 협동적 필터링이 추천 시스템을 위한 좋은 선택이라고 여겨질 수 있지만, 예측의 질을 향상시키기 위한 여지가 여전히 많이 남아 있다. 이를 위해서 협동적 필터링은 유용한 이웃선정 방법이 필요하다.

3. 클러스터링 기반 이웃선정 방법

k-means 클러스터링 방법은 서로 유사한 선호도를 갖는 고객들로 구성된 k개의 클러스터를 만든다[4][6]. 이 방법은 먼저 임의로 k명의 고객들을 k개의 클러스터에 대한 초기 중심점으로 선택한다. 그런 후에 모든 고객은 클러스터의 중심과 고객 사이의 거리가 최소가 되는 하나의 클러스터에 할당한다. 거리는 Euclidean distance를 사용하여 계산하게 되는데 이것은 고객과 각 중심점 사이의 각 속성의 차에 대한 제곱의 합에 대한 제곱근으로 구할 수 있다.

그 후에 각 클러스터에 대해서 우리는 클러스터에 현재 속한 고객들을 기반으로 클러스터의 평균을 다시 계산한다. 이 평균은 클러스터의 새로운 중심으로 고려된다. 새로운 중심을 찾은 후에 우리는 고객이 속해야 하는 클러스터를 찾기 위해 각 고객에 대한 거리를 계산한다. 평균을 재계산하고 거리를 계산하는 것은 종료조건을 만나게 될 때까지 반복된

다. 종료조건은 일반적으로 모든 새로운 중심이 이전 중심으로부터 각각 얼마나 움직였는가 하는 것이다. 만약 모든 새로운 중심이 어떤 한계 거리 내에서 움직였다면 우리는 반복을 종료한다.

클러스터링에 의해 최종적으로 k개의 클러스터와 각 클러스터에 속하는 고객 집합이 결정되면 각 클러스터에 속한 고객 집합은 모두 후보 이웃이 된다. 이후 선호도를 예측하고자 하는 고객의 상품 속성에 대한 선호도와 각 클러스터의 대표 값이 지니는 속성 선호도 사이의 거리를 계산하여 가장 최소의 값을 갖는 클러스터를 선정한다. 이렇게 결정된 클러스터에 속하는 고객 집합이 최종적인 이웃 고객으로 선정된다.

클러스터링 기반 이웃선정 방법은 클러스터링 이후 예측하고자 하는 고객의 선호도와 가장 유사한 클러스터내의 고객들만 예측을 위한 이웃으로 결정하기 때문에 대용량 데이터 셋의 경우 빠른 예측을 통한 추천이 가능하게 되는 장점이 있다.

3. Ordered clustering을 이용한 이웃선정 방법

이웃 고객을 찾는 탐색 공간을 다변화하기 위한 방법으로 클러스터를 순서적으로 구성하는 방법을 생각할 수 있다. 우리는 클러스터링을 함으로써 상품의 속성들에 대한 고객의 선호도를 유사 성향에 따라 알맞게 분류할 수 있다. 그런 다음 테스트 고객에게 가장 적합한 "*best*" 클러스터를 선택함으로써 탐색 공간을 줄이며 의미 있는 유사 고객들을 이웃 고객으로 구성할 수 있다.

하지만 우리가 클러스터를 하나만 선택했을 경우, 비록 그것이 테스트 고객에게 가장 적합한 클러스터라 할지라도 의미 있는 고객이 해당 클러스터에 모두 포함될 것을 의미하는 것은 아니다. 따라서 테스트 고객과 그 다음으로 적합한 차선의 클러스터들에 대해서도 탐색 공간을 확장할 필요가 있다. 이렇게 클러스터들을 테스트 고객과의 유사 정도에 따라 순서적으로 구성하여 상위 클러스터들을 묶어 탐색 공간을 확장한다면 의미 있는 고객들을 이웃 고객 집합에 추가 시킬 수 있게 된다.

그림1은 클러스터를 순서적으로 구성하여 테스트 고객에게 가장 적합한 두 개의 클러스터를 통합한 후 이웃 고객 집합을 다시 한 번 threshold에 따라 구성해 나가는 것을 보여주고 있다. 이 방법은 최선의 클러스터만 탐색하는 방법에 비해서 탐색 공간이 조금 늘어나는 점이 있지만, 드러나지 않은 의미 있는 고객들을 찾아낼 수 있다는 점에서 예측 정확도의 향상을 가져올 수 있게 된다.

Threshold τ의 적용은 양의 상관 고객("*positive*" 고객)과 음의 상관 고객("*negative*" 고객)을 함께 고려한다. 즉, 클러스터링 이후 선정된 클러스터에 속한 고객들 중에서 negative τ 값 보다 작거나 같은 고객들이나 positive τ 값 보다 크거나 같은 고객들을 이웃 고객으로 최종 선정하게 된다. 그림에서 실선으로 연결된 고객은 positive 고객을, 점선으로 연결된 고객은 negative 고객을 나타낸다.

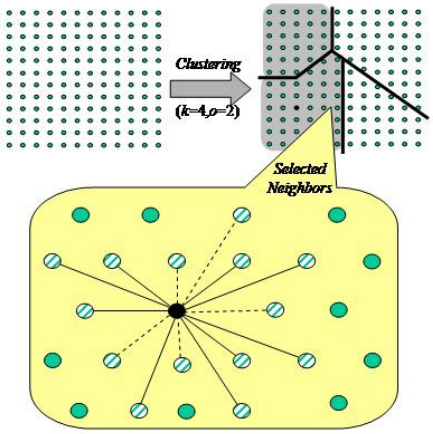

그림1. Ordered clustering (k=4, o=2)

4. 실험 결과

4.1. 실험 환경

제안한 추천 방법의 예측 정확도를 평가하기 위해 우리는 Compaq Computer Corporation에 의해 제공된 EachMovie 데이터 셋을 사용하였다[5]. EachMovie 데이터 셋은 72,916명의 고객들이 명시적으로 1,628개의 영화에 대해서 평가한 2,811,983개의 선호도들로 구성되어 있다. 고객 선호도는 0.0에서 1.0까지 0.2 간격의 수치 값으로 표현되어 있다.

실험을 위해서 우리는 데이터 셋으로부터 모든 고객들 중에서 최소 100회 이상 선호도를 입력한 3,763 명의 고객들을 추출하였다. 우리는 이 중에서 임의적으로 10명의 고객들을 테스트 고객으로 선정하였고, 나머지 고객들을 트레이닝 고객으로 하였다. 또한 이렇게 구성된 테스트 셋에 있는 각 테스트 고객에 대해서 테스트 고객에 의해 실제 평가된 5개의 영화를 임의의 테스트 아이템으로 선정하였다. 그리고 최종 실험 결과는 서로 다른 5개의 테스트 셋으로 실험한 결과에 대한 평균을 구하여 얻었다.

4.2. 실험 평가

추천 시스템의 통계적 예측 정확도 평가 방법 중

하나가 개별 예측에 대해서 예측된 평가 값에 대한 실제 고객의 평가 값의 오차에 대한 평균을 나타내는 MAE(mean absolute error)이다[1][3][6]. MAE는 식(3)에 의해서 계산된다. 이 식에서 N은 예측의 총 횟수이고 ε_i는 아이템 i에 대한 예측된 평가 값과 실제 평가 값 사이의 오차를 뜻한다. 그러므로 낮은 MAE는 고객의 수치적 평가에 대한 더 정확한 예측 정확도를 갖는다.

$$|E| = \frac{\sum_{i=0}^{N} |\varepsilon_i|}{N}. \quad (3)$$

4.3. 실험 결과

우리는 k-means 클러스터링에 의한 이웃선정 방법(*KMCF*)과 본 논문에서 제안한 ordered clustering에 threshold를 적용한 이웃 선정 방법(*KMTCF_o*)을 상호 비교하였다. 우리는 KMTCF_o에서 best 클러스터 하나만 선정할 경우를 따로 "*KMTCF*"로 구분하여 나타내었다.

실험 결과는 표1에 나타내었다. 우리는 많은 실험을 통해 파라미터 값을 MAE가 가장 낮게 나오는 값으로 정하였다. 모든 방법에서 k 값은 클러스터링에 의한 클러스터의 수를 말한다. KMTCF와 KMTCF_o에서 파라미터 τ 값은 positive 값과 negative 값으로 구분된 threshold 값이다. 마지막으로 KMTCF_o에서 파라미터 o의 값은 클러스터를 순서화한 후에 이웃 고객 선정을 위한 탐색 공간으로 몇 개의 클러스터를 결정할 것인지를 나타내는 선정된 클러스터의 수를 뜻한다.

표1. 실험 결과

방법	MAE	파라미터
KMCF	0.172146	k=21
KMTCF	0.166095	k=21, τ=(-0.6, 0.4)
KMTCF_o	0.164841	k=21, τ=(-0.6, 0.5), o=2

실험 결과는 KMTCF가 KMCF 보다 우수한 예측 정확도를 가진다는 것을 보여주며, 이 때 KMTCF의 threshold 값은 τ = (-0.6, 0.4) 이다. 또한 결과로부터 우리는 KMTCF_o가 순서화된 클러스터 수가 2 일 때, 다른 방법들 중에서 가장 좋은 예측 정확도를 갖는 것을 알 수 있다. 이것은 클러스터를 순서적으로 구성하지 않았을 때 제외될 가능성이 있는 우수 고객들을 이웃으로 적절히 선정할 수 있었기 때문이라 할 수 있다.

5. 결론

추천 시스템은 고객의 선호도를 추출하고 분석함으로써 정확한 예측을 수행하는 능력을 갖는 것이 매우 중요하다. 협동적 필터링이 비록 추천 시스템에 폭넓게 사용되고 있지만, 이의 단점을 극복하기 위한 노력들이 예측의 질을 향상시키기 위해 수행되어야 한다.

본 논문에서 우리는 ordered clustering 방법에 threshold를 적용한 새로운 이웃 선정 방법을 제안하였다. 이것은 클러스터링을 통해 대규모 데이터 셋에서 탐색 공간을 줄일 수 있으며, 순서적 클러스터링과 threshold의 적용으로 예측 정확도의 향상을 가져올 수 있다. 실험 결과로부터 제안된 방법의 예측 정확도 향상을 확인할 수 있었다.

참고문헌

[1] Badrul M. Sarwar, George Karypis, Joseph A. Konstan, John T. Riedle, “Application of Dimensionality Reduction in Recommender System - A Case Study,” Proceedings of the ACM WebKDD 2000 Web Mining for E-Commerce Workshop, 2000.

[2] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, “Analysis of Recommendation Algorithms for E-Commerce,” Proceedings of the ACM E-Commerce, 2000.

[3] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl, “An Algorithmic Framework for Performing Collaborative Filtering,” Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 1999.

[4] B.M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Recommender Systems for Large-Scale E-Commerce: Scalable Neighborhood Formation Using Clustering,” Proceedings of the Fifth International Conference on Computer and Information Technology, 2002.

[5] EachMovie Collaborative Filtering Data Set. Compaq Computer Corporation, url: http://www.research.compaq.com/SRC/eachmovie/.

[6] O’Connor M., and Herlocker J., “Clustering Items for Collaborative Filtering,” Proceedings of the ACM SIGIR Workshop on Recommender Systems, 1999.