추천 시스템을 위한 고객 클러스터링 방법을 적용한 예측 알고리즘

박지선, 김택헌, 류영석, 양성봉
연세대학교 컴퓨터과학과
{jspark, kimthun, ryu, yang}@mythos.yonsei.ac.kr

A Predictive Algorithm Applying Customer Clustering Method for Recommendation Systems

Ji-Sun Park, Taek-Hun Kim, Young-Suk Ryu, Sung-Bong Yang
Dept. of Computer Science, Yonsei University

요 약

전자상거래에서 최근 대부분의 개인화된 추천 에이전트 시스템들은 협동적 필터링 기술을 적용하고 있다. 이 방법은 고객의 취향에 맞는 상품을 예측하고 추천하기 위하여 비슷한 선호도를 가지는 다른 고객들과의 상관 관계를 구하기 위하여 일반적으로 피어슨 상관 계수를 이용한다. 그러나 이 방법은 오직 두 고객 사이에서 두 고객 모두 평가를 한 상품이 있을 때에만 상관 관계를 구할 수 있으므로 예측의 정확성이 떨어질 수 있다.

본 논문에서는 이러한 이웃 선정 방법에 대한 문제점을 보완하기 위하여 비슷한 선호 패턴을 가지는 고객들을 보다 적절히 군집화 하여 이 군집에 속한 고객들의 평가를 기반으로 협동적 필터링 기술을 수행하는 방법을 제안하고, 기존의 협동적 필터링 기술과의 비교 실험을 통해 성능을 평가하였다. 실험결과 본 논문에서 제안한 방법이 기존의 방법보다 우수함을 확인할 수 있었다.

1. 서론

개인화된 추천 에이전트 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 추천해줄 수 있는 능력이다. 이를 위해서는 데이터 마이닝 기법, 패턴 인식 기술, 정보 필터링 기술 등 다양한 기법들이 적용될 수 있으나 대부분의 추천 에이전트 시스템들은 정보 필터링 기술을 적용한다. 정보 필터링 기술의 대표적인 것으로는 협동적 필터링 기술(collaborative filtering)이 있다[1].

협동적 필터링은 추천 에이전트 시스템에 가장 많이 사용되는 기술로써 Amazone.com, Cdnow.com 등 상업적으로 성공을 거두고 있는 여러 전자상거래 사이트에서 적용하고 있다. 이 방법은 고객이 좋아할 만한 상품을 예측하기 위하여 비슷한 선호도를 가지는 다른 고객들의 상품에 대한 평가에 근거하여 추천하는 방법이므로 높은 예측력과 추천 능력을 가지는 장점이 있다. 그러나 이런 장점에도 불구하고 이 방법은 상품의 속성에 대한 개인의 선호도를 직접적으로 반영하지는 못하는 단점을 가지고 있다.

협동적 필터링 기술에서 지적되는 가장 큰 문제점은 고객의 선호도 간의 유사성을 평가하기 위해 사용하는 피어슨 상관 계수로부터 야기된다[1][4]. 두 고객이 모두 평가를 한 상품이 있어야 하고 오직 두 고객 사이에서만 상관 관계를 구할 수 있으므로 예측의 정확성이 떨어질 가능성이 있다.

본 논문에서는 위에서 언급한 기존의 협동적 필터링 기술의 문제점을 보완하기 위하여 k-means 클러스터링 알고리즘을 사용하여 유사한 선호도를 가지는 고객들을 적절히 군집화 하여 특정 상품에 대한 고객의 선호도를 그 고객이 속한 군집내의 다른 고객들의 평가를 기반으로 예측하여 추천해 주는 새로운 기법을 제안하고 그 성능을 기존의 협동적 필터링 기술과 비교 평가 하였다.

본 논문의 구성은 다음과 같다. 2장에서 협동적 필터링 기술에 대한 관련연구를 설명하고, 3장은 본 논문에서 제안하는 알고리즘에 대해서 설명한다. 4장에서 실험을 통한 성능을 분석하며 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

2.1. 협동적 필터링 기술

협동적 필터링 기술은 특정 고객의 상품에 대한 선호도를 예측하기 위하여 대부분의 경우 식(2)에 나타나 있는 피어슨 상관 계수를 이용하여 유사한 선호도를 가지는 이웃들(neighborhood) 을 정하고 식(1)에 의해 예측 선

호도 값을 계산한다[1].

$$U_x = \overline{U} + \frac{\sum_{J \in Raters} (J_x - \overline{J}) r_{UJ}}{\sum_{J \in Raters} |r_{UJ}|} \quad (1)$$

여기서

$$r_{UJ} = \frac{\sum (U - \overline{U})(J - \overline{J})}{\sqrt{\sum (U - \overline{U})^2 \cdot \sum (J - \overline{J})^2}} \quad (2)$$

U_x는 상품 x에 대한 고객 U의 예측된 선호도이고 r_{UJ}는 고객 U와 J의 상관관계를 나타내며 두 사용자 모두 선호도를 표시한 아이템에 대해서만 계산된다. *Raters*는 테스트 상품에 대해 선호도를 표시한 고객들을 나타낸다.

2.2. 협동적 필터링 기술을 적용한 추천 에이전트 시스템

Tapestry[5]는 협동적 필터링 기술을 가장 먼저 적용한 문서 필터링 시스템으로 워크그룹과 같은 공동체 구성원들의 의견에 기반하여 추천을 해주므로 개인화 된 추천 서비스는 제공해 주지 못한다[4]. 개인화된 추천은 추천 대상 고객과 유사한 선호도를 가지는 이웃들을 선택함으로써 제공될 수 있다.

최근 몇 년 동안에는 특히 자동화된 협동적 필터링 시스템이 많이 개발 되었는데 그 중 GroupLens research system[2][3]은 Tapestry의 문제점을 보완 하면서 성능을 인정 받은 시스템으로서 유즈넷 뉴스와 영화에 대한 추천을 수행한다.

GroupLens를 포함한 대부분의 협동적 필터링 기법을 사용하는 추천 에이전트 시스템들은 피어슨 상관 계수를 사용하여 유사 선호도를 가지는 이웃들을 결정한다[1].

2.3. 협동적 필터링 기술의 한계점

GroupLens와 같은 기존의 협동적 필터링 기술에서 사용한 피어슨 상관 계수 기반 예측 기법의 단점은 다음 세가지로 요약할 수 있다[1][4].

첫째, 두 고객사이의 상관관계는 오직 두 고객 모두 선호도를 표시한 상품에 대해서만 계산 되므로 만약 상품의 수가 많으면 일반적으로 같은 상품에 대하여 두 고객 모두 선호도를 표시할 확률은 매우 적게 된다.

둘째, 비록 두 고객이 선호도에 따른 상관관계가 높지 않더라도 다른 고객의 선호도 예측에 좋은 자료가 될 수 있으나 상관관계가 높지 않다는 이유로 이 정보는 활용되지 못한다.

마지막으로, 상관관계가 오직 두 고객 사이에서만 계산 된다는 것이다. 예를 들어 사용자 갑과 을이 아주 높은 상관관계에 있고, 을과 병도 그렇다고 가정하면 갑과 병도 상관관계가 높다고 할 수 있다. 그러나 만약에 갑과 병이 공통된 상품 어느것에도 선호도를 표시하지 않았다면 상관관계를 구할 수 없다.

3. 고객 클러스터링 방법을 이용한 협동적 필터링 기술

3.1. 데이터 셋 (data sets)

본 논문에서는 1997년 digital equipment corporation에 의해서 공개된 EachMovie[7] 데이터 셋을 사용하였다.

이 데이터 셋은 총 72916명의 사용자가 1628개의 영화에 대해 0.0부터 최대 1.0까지 0.2의 차이를 두고 명시적으로 평가한 선호도들로 구성되어 있다. 영화의 장르는 액션, 애니메이션, 외국 예술, 고전, 코미디, 드라마, 가족, 공포, 로맨스, 스릴러의 10가지로 구분되어 있다.

3.2. 기본 알고리즘

임의의 테스트 영화에 대하여 고객의 선호도를 예측하고 추천 리스트를 생성 하기 위해서 다음과 같은 단계를 거친다.

- 단계1 : 영화의 각 장르에 대한 선호도를 각기 다른 차원으로 하여 k-means 클러스터링 알고리즘을 적용하여 기존 고객(training users)을 k개로 군집화 한다.
- 단계2 : k개의 군집들과 고객간의 각 속성에 대한 유사성을 계산하여 가장 유사성이 높은 군집을 선택한다.
- 단계3 : 단계2에서 결정된 군집에 속하는 다른 고객들은 새롭게 구성된 이웃들이며 따라서 이 이웃들에 대해서만 협동적 필터링을 적용하여 테스트 영화에 대한 선호도를 예측 한다.
- 단계4 : 단계3에서 구한 예측 선호도 값을 이용하여 추천 리스트를 생성한다.

3.3. 세부 적용 알고리즘

3.3.1. K-means 클러스터링 알고리즘

K-means 클러스터링 알고리즘은 군집 영역에 속하는 모든 점으로부터 군집 중심까지의 거리의 제곱의 합으로 정의되는 성능 지표를 최소화하는데 바탕을 둔 방법이다[9]. 이 방법은 다음과 같은 단계로 구성 된다.

- 단계 1 : 군집의 수 k를 정한 후, k개의 초기 군집 중심을 선택한다. 일반적으로 주어진 표본 집합의 처음 k개의 표본을 임의로 선택한다.
- 단계 2 : 각 관찰치를 거리 기법을 이용하여 그 중심과 가장 가까운 거리에 있는 군집영역에 분배한다.
- 단계 3 : 단계2의 결과로부터, 모든 군집에 대하여 해당 군집에 포함된 모든 점들로부터 새로운 군집 중심을 계산한다.
- 단계 4 : 모든 군집에 대하여 기존의 중심과 새로운 중심의 차이가 없으면 알고리즘은 수렴하며 종료 되고, 그렇지 않으면 단계2로 간다.

3.3.2. 협동적 필터링에 의한 선호도 예측

3.2절의 단계 2에서 확정된 군집 내에 속하는 고객들이 테스트 고객의 이웃이 된다. 기존의 협동적 필터링 기법이 모든 고객들을 이웃으로 이용한 반면, 본 논문에서 제안하는 기법은 정해진 군집내의 고객들에 대해서만 협동적 필터링 기법을 적용한다. 따라서 식 (1), (2)에서 J는 테스트 영화에 대해 선호도를 입력한 고객들 중 테스트 고객과 같은 군집 내에 있는 고객들을 의미한다.

같은 군집 내에 속한다는 것은 영화에 대해 유사한 선호도를 가진다는 것이므로 이러한 고객들만 협동적 필터링에 적용함으로써 기존의 협동적 필터링 기법에 비하여 선호도 예측의 정확성의 향상과 질 높은 추천 리스트를 생성할 수 있다.

4. 실험 및 성능 평가

4.1. 테스트 데이터 셋

EachMovie 데이터 중 최소 100회 이상 선호도를 입력한 사용자 4788명을 추출하여 모든 장르의 영화에 대해

선호도를 입력한 사용자 3763명을 최종적으로 추출 하였다.

이 중에서 테스트 고객 10명을 랜덤하게 선택하고, 나머지 3753명을 군집을 위한 기존 고객으로 선택하여 실험을 진행하였다.

4.2. 성능 평가 기준

4.2.1. 예측의 정확성에 대한 평가

본 논문에서는 예측 값의 정확성 측면에서 성능을 평가하기 위해 MAE(Mean Absolute Error)를 사용하였으며 식(3)에 나타낸 것과 같이 구할 수 있다[6].

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

(N: 총 예측 회수, ε_i : 예측값과 실제값의 오차, i : 각 예측 단계)

4.2.2. 추천 리스트에 대한 평가

추천 리스트에 대한 성능을 평가 하기 위한 방법으로는 Precision, Precision at Top N, Recall, F-measure가 있다[8].

- Precision : 추천 리스트 중에서 몇 개의 영화를 고객이 실제로 좋아했는지를 나타내는 평가 방법.
- Precision at Top N : Top N 추천 리스트 중에서 몇 개의 영화를 고객이 실제로 좋아했는지를 나타내는 평가 방법.
- Recall : 고객이 좋아하는 영화 중에서 얼마나 많은 영화가 추천이 되었는지 나타내는 평가 방법.
- F-measure : precision과 recall에 동등한 중요도를 부여하여 하나의 평가방법으로 사용 하는 것으로 식(4)와 같이 구할 수 있다.

$$F = \frac{2 * precision * recall}{precision + recall} \qquad (4)$$

4.3. 실험 결과 및 분석

표1과 2는 본 논문에서 제안한 방법과 GroupLens 방법과의 비교 실험 결과를 나타낸 것이다. 여기에서 k값은 반복 실험을 통해 얻은 군집의 수이며 이 실험에서는 27을 사용하였다.

평가방법	GroupLens	제안한 방법
MAE	0.213436	0.204670

표1. 예측의 정확성에 따른 비교 실험 결과

평가방법	GroupLens	제안한 방법
Precision at Top 7	68 %	81 %
Precision	60 %	79 %
Recall	51 %	54 %
F-measure	55 %	64 %

표2. 추천 리스트에 대한 비교 실험 결과

표1과 표2의 실험 결과를 보면 본 논문에서 제안한 고객 클러스터링 방법을 적용한 협동적 필터링 방법이 GroupLens 방법보다 모든 평가 항목에서 우수한 성능을 보임을 알 수 있다.

특히 Precision at Top 7의 평가 항목에서는 GroupLens 방법에 비해 아주 우수한 성능을 보였다.

5. 결론 및 향후 연구

본 논문에서는 유사한 선호 패턴을 가지는 고객들을 k-means 클러스터링 기법을 통해 적절히 군집화 하고 이 군집에 속한 고객들의 평가를 기반으로 협동적 필터링 기술을 수행하는 방법을 제안 하였다.

제안한 방법에 대한 성능을 기존의 협동적 필터링 방법과 비교 실험한 결과 예측의 정확성, 추천 리스트에 대한 평가 모두 성능 향상이 있었다.

유사한 선호 패턴을 가지는 고객들을 군집화 할 때 영화의 장르에 대한 선호도를 가지고 수행하였으나 향후 연구에서는 장르뿐만 아니라 여러 가지 속성에 대한 선호도를 이용한다면 보다 신뢰성 있고 향상된 결과를 기대할 수 있을 것이다.

6. 참고 문헌

[1] Daniel Billsus, Michael J. Pazzani, "Learning Collaborative Information Filters," *Proceedings of ICML*, pp.46-53, 1998.

[2] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J., "GroupLens: Applying Collaborative Filtering to Usenet News," *Communications of the ACM*, Vol. 40, No. 3, pp.77-87, 1997.

[3] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," *Proceedings of ACM CSCW'94 Conference on Computer Supported Cooperative Work*, pp.175-186, 1994.

[4] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Analysis of Recommendation Algorithms for E-Commerce," *The ACM E-Commerce 2000 Conference*, 2000.

[5] Goldberg, D., Nichols, D., Oki, B.M., and Terry, D., "Using Collaborative Filtering to Weave an Information Tapestry," *Communications of the ACM*, Vol. 35, No. 12, pp.61-70, 1992.

[6] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl, "An Algorithm Framework for Performing Collaborative Filtering," *Proceedings of the 1999 Conference on Research and Development in Information Retrieval*, 1999.

[7] P. McJones, EachMovie collaborative filtering data set, URL:http://www.research.digital.com/SRC/eachmovie/, 1997.

[8] Raymond J. Mooney, Loriene Roy, "Content-Based Book Recommending Using Learning for Text Categorization," *Proceedings of the fifth ACM Conference on ACM 2000 digital libraries*, pp.195 - 204, 2000.

[9] 이성환, 패턴인식의 원리 1권, p.96-100, 홍릉과학출판사, 1994.