협력적 여과기법의 평균과 이웃정보에 관한 연구

김선옥*°, 이경호**, 이석준***, 이희춘****

*한라대학교 정보통신방송공학부, **한라대학교 정보통신방송공학부, ***상지대학교 경영정보학과, **상지대학교 컴퓨터데이터학과

Study of the mean and information of neighbors in NBCFA

Kim, Sun-Ok, Lee, Kyong-Ho, Lee, Seok-Jun, Lee, Hee-Choon

Halla University, Halla University, SangJi University, SangJi University

E-mail :sokim@halla.ac.kr, khlee@halla.ac.kr, digitaldesign@sangji.ac.kr, choolee@sangji.ac.kr

요 약

추천시스템에서 널리 사용되고 있는 협력적 여과기법은 이웃의 정보를 추천대상 고객에게 적용하여 추천에 사용한다. 이 방법을 이용한 추천은 인터넷 사용자에게 알맞은 정보를 제공하여 보다 편리하게 자신이 원하는 정보에 접근하도록 한다. 따라서 추천시스템의 성능향상에 대한 연구가 활발히 진행되고 있으며, 본 논문은 추천시스템의 기능에 대한 정확성을 향상시키기 위한 것이다. 본 논문에서는 먼저, 협력적 여과기법에서 사용되는 고객의 선호도 평가 값에 대한 평균값을 조사하고, 이웃들이 평가한 선호도 평가 값을 분석하였다. 그리고 협력적 여과기법에 두 개의 분석 값을 변수로 적용하여 추천시스템의 예측 정확도를 계산하였다. 본 논문이 제안한 방법과 기존의 알고리즘을 비교한 결과 추천시스템의 성능이 향상됨을 알 수 있다.

1.서론

인터넷의 보급으로 인하여 다양한 정보가 인터넷 사용자들에게 제공되었다. 이러한 다양한 정보는 대량으로 제공되어 필요한 정보를 얻기 위한 많은 시간이 필요하게 되었다. 따라서 정보에 대한 체계적인 추출과 사용자에게 필요한 정보의 제공자가 필요하게 되었다. 이에 따라 추천시스템이 도입되어 인터넷 사용자들에게 필요한 정보를 제공하여 사용자들이 보다 편리하게 정보를 획득할 수 있게 되었다. 또한 기업들은 잠재적인 고객의 확보를 위해 추천시스템을 사용하였으며 이는 기업의 이윤에 도움이 되는 도구로도 이용되었다. 추천시스템은 인터넷의 발달과 더불어 중요한 도구로 인식되고 있으며, 추천시스템의 성능향상을 위한 방법이 지속적으로 연구되고 있다. 본 연구는 추천시스템의 성능향상을 위해 추천시스템 중에서 가장 많이 사용되고 있는 이웃기반 협력적 여과기법을 이용한 방법에 대한 새로운 알고리즘을 제안하여 추천시스템의 성능 향상에 기여하고자 한다.

2. 추천알고리즘

2-1. NBCF 알고리즘

이웃의 정보를 이용하여 추천에 사용되는 협력적 여과기법(Collaborative Filtering)의 가장 일반화 된 알고리즘은 이웃 기반 협력적 여과기법(Neighborhood Based Collaborative Filtering)이다. 이 기법은 미

네소타대학의 GroupLens에서 뉴스 기사 추천을 위해 사용되었으며, MovieLens에서는 영화추천을 위해 사용되는 등 다양한 분야에 적용되고 있다. 이웃기반 협력적 여과기법은 추천을 위해 먼저 추천대상고객의 이웃들을 선정하여야 한다. 따라서 이웃을 선정하는 여러 가지 방법의 연구가 꾸준히 진행되고 있으며, Kim et. al(2007)은 군집화의 방법을 이용한 이웃선정으로 추천시스템의 성능 향상을 연구하였다. Lee et. al(2007)는 인구통계학의 자료를 이용한 이웃선정 방법을 이용하여 추천시스템의 성능 향상에 기여하였다.

본 논문에서는 추천 대상 고객이 선호한 상품 중에서 선호도를 예측하고자 하는 상품에 선호도를 평가한 고객만을 이웃으로 선정한다. 이렇게 선정된 이웃고객들의 평가값과 추천대상 고객의 상품들에 대한 선호도 유사정도를 이용하여 추천시스템에 사용한다. 선호도 유사정도는 추천대상 고객과 이웃 고객이 상품들에 평가한 선호도 평가치들의 상관관계를 이용한다. 다음 식은 두 고객의 유사정도를 나타내는 상관계수 중에서 추천시스템에서 가장 많이 사용되고 있는 피어슨 상관계수이다(Resnick et. al., 1994).

$$r_{uj} = \frac{\sum_{i=1}^{m}(R_{u,i} - \overline{R_u})(R_{j,i} - \overline{R_j})}{\sqrt{\sum_{i=1}^{m}(R_{u,i} - \overline{R_u})^2 \cdot \sum_{i=1}^{m}(R_{j,i} - \overline{R_j})^2}} \quad (1)$$

여기에서, r_{uj}는 추천대상 고객 u와 이웃 고객 j와의 유사정도를 나타내는 가중치이며, $R_{u,i}$는 추천 대상 고객 u가 평가한 상품 i 에 대한 선호도 평가 치이고, $R_{j,i}$는 이웃고객 j가 평가한 상품 i 에 대한 선호도 평가이다. $\overline{R_u}$는 추천대상 고객 u가 평가한 모든 상품들에 대한 평균이고, $\overline{R_j}$는 추천대상 고객의 이웃인 j고객의 상품 선호도평가에 대한 상품들의 선호도 평가 치들에 대한 평균값이다.

추천 대상 고객에게 상품을 추천하기 위해서는 추천 상품에 대한 선호도 예측 값을 계산하여야 한다. 상품에 대한 선호도 예측은 추천대상 고객의 평균과 추천대상 고객의 이웃들이 평가한 평가 값 그리고 이들 이웃고객의 평균값을 사용하며, 식(1)에서 소개된 이웃과의 유사도 가중치를 알아야한다. 다음 식은 협력적 여과기법을 사용한 선호도 예측값을 계산하기 위한 알고리즘이다(Konstan et. al., 1997).

$$\widehat{U}_x = \overline{U} + \frac{\sum_{J \in Raters}(J_x - \overline{J})r_{uj}}{\sum_{J \in Raters}|r_{uj}|} \quad (2)$$

여기에서, $\widehat{U}_x$는 상품 x에 대한 추천 대상 고객 u의 선호도 예측 치이다. $\overline{U}$는 추천 대상 고객 u가 평가한 모든 상품에 대한 평균이다. J_x는 상품 x에 대한 이웃 고객 j의 선호도 평가 치이고, $\overline{J}$는 이웃 고객 j가 평가한 모든 상품에 대한 선호도의 평균이다. $\overline{J}$의 값은 평가치 중에서 상품 x에 대한 평가치는 제외한다. r_{uj}는 추천 대상 고객 u와 추천 대상 고객의 이웃고객인 j의 선호 유사 정도를 나타내는 유사도 가중치이다.

2-2. 알고리즘의 성능 평가

추천시스템에서 선호도 예측의 정확도를 평가하기 위해서는 절대평균오차(Mean Absolute Error)를 이용하며, 추천대상 고객이 실제 평가한 평가 값과 협력적 여과기법에 의해 계산된 선호도 예측 값이 사용된다. 다음 식은 선호도 예측의 정확도를 판정하기 위한 계산식이다(Kim, Lee and Lee, 2008).

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|R_{uj} - \widehat{R}_{uj}| \quad (3)$$

여기에서, R_{uj}는 상품 j에 대한 추천 대상 고객 u의 실제 선호도 평가 치이고, $\widehat{R}_{uj}$는 상품 j의 추천 대상 고객 u을 위한 선호도 예측 값이고, 예측 값은 협력적 여과기법의 알고리즘을 사용한다.

3.실험

3-1. 실험 데이터의 구성

실험 데이터는 GroupLens에서 영화에 관한 선호도를 나타낸 MovieLens 100K dataset

과 MovieLens 1M dataset이다. 100K dataset는 943명의 고객이 1682편의 영화에 1에서 5점사이의 점수로 선호한 평가 값을 나타냈다.

data1은 각 영화에 대해 고객이 선호한 평가치 100,000개의 데이터를 80%의 훈련 데이터(training data)와 20%의 실험 데이터(test data)로 랜덤하게 나누어 실험에 사용하였다. 그리고 100,000개의 데이터 중에서 70%를 훈련 데이터(training data)로 사용하고 20%를 실험 데이터(test data)로 랜덤하게 구성하여 제안한 방법의 정확한 연구를 위해 data2로 사용하였다.

3-2. 제안한 실험방법

NBCFA 알고리즘을 이용한 추천시스템은 이웃에 대한 정보가 추천에 사용되며, 추천 대상 고객의 선호도 평가를 위해서 추천 고객의 평균값과 이웃고객의 선호도 평균값이 중요한 변수로 작용되었다. 본 논문에서는 이들 변수를 분석하여 추천시스템의 예측 정확도를 개선하는 새로운 계수를 탐색적 데이터 분석방법을 이용하여 제시하고자 한다. EDA(Exploratory data analysis) 방법은 데이터의 구조와 특징을 알아보기 위한 데이터 분석법으로 그 중에서 적합 직선을 이용한 방법은 데이터의 지배적인 형태를 알아내는 방법으로 사용되고 있다.

본 논문에서는 데이터의 적합 직선화 방법을 이용한 데이터의 변환으로 추천시스템의 예측 값에 대한 새로운 방법을 제시하고자 한다. 먼저, 두 변수

$$(\hat{U}_x, \frac{\sum_{J\in Raters}(J_x-\bar{J})r_{uj}}{\sum_{J\in Raters}|r_{uj}|}) \qquad (4)$$

을 이용하여 이들 변수사이의 위치를 계산하여 적합 직선의 방정식을 구하였다. 그리고 적합 직선방정식을 이용하여 새로운 보정 직선의 방정식을 정의하였다. 구해진 보정 직선방정식은 각각의 실험 데이터에 따라 다음과 같은 보정함수로 표현될 수 있다.

$$F(a,\ b,\ c) = a\overline{U} + b\frac{\sum_{J\in Raters}(J_x-\bar{J})r_{uj}}{\sum_{J\in Raters}|r_{uj}|} + c \qquad (4)$$

여기에서, 제안한 상수 a, b 그리고 c는 다음과 같다.

<표 1> 실험데이터에 따른 보정함수

실험 데이터	a	b	c
data1	0.83520	1.17559	0.58005
data2	0.83597	1.14728	0.57942

3-3. 실험결과

제안한 상수를 이용하여 보정함수를 사용한 실험데이터에 훈련 데이터를 적용하여 각각의 데이터에 기존의 알고리즘과 예측의 정확도에 대한 값을 비교한 결과는 다음과 같다.

<표 2> 보정함수를 적용한 MAE

실험	MAE평균		개 수 (N)	t	유의확률
	기존 알고리즘	보정함수 알고리즘			
data1	0.75271	0.75093	19973	2.72	0.00
data2	0.75867	0.75736	19969	2.13	0.01

제안한 보정함수를 이용한 data1의 결과를 살펴보면 기존의 알고리즘을 이용한 예측정확도의 값인 0.75271의 결과 값보다 좋아진 0.75093의 결과 값이 나왔다. 그리고 data2의 경우도 제안된 보정함수를 사용한 값이 0.75736으로 기존의 알고리즘을 이용한 값인 0.75867보다 예측 정확도의 값이 좋아졌음을 알 수 있다.

다음은 각각의 실험 데이터를 개인별로 적용하여 보정함수를 사용한 결과를 조사하였다. dataSet1은 0.78065인 기존 값보다 보정함수를 이용한 결과 값인 0.77857으로 향상되었음을 알 수 있다. dataSet2도 보정함수를 이용한 결과 0.78656으로 기존의 알고리즘을 이용한 결과 값인 0.78820보다 좋아졌음을 알 수 있다.

<표 3> 보정함수를 적용한 개인별 MAE

데이터	알고리즘	MAE평균	N	t	유의확률
dataSet1	기존 알고리즘	0.78065	943	2.05	0.02
	보정함수 알고리즘	0.77857			
dataSet2	기존 알고리즘	0.78820	943	1.71	0.04
	보정함수 알고리즘	0.78656			

4.결론

협력적 여과기법을 이용한 추천시스템에서 추천 대상 고객의 평균값과 이웃이 평가한 선호도 평균값을 조사하였다. 그리고 이 두 개의 결과 값을 변수로 사용한 보정함수를 이용하여 기존의 알고리즘을 이용한 결과 값과 비교분석하였다. 전체데이터에서 보정함수를 이용한 결과 값이 기존 알고리즘을 이용한 결과 값보다 향상됨이 조사되었다. 또한 개인별로 보정함수를 이용한 결과 값도 기존의 알고리즘을 이용한 결과보다 좋아짐을 알 수 있었다. 따라서 협력적 여과기법을 이용하여 각각의 데이터에 보정함수를 적용하면 추천시스템의 예측 정확도가 개선됨을 기대할 수 있다.

[참고문헌]

[1]김선옥, 이석준, 이희춘(2008). 임계값이 표준편차에 미치는 영향에 관한 연구,2008 한국IT서비스춘계학술대회, pp.511-515, 2008.

[2]김재경, 오희영, 권오병(2007). 유비쿼터스 환경에서 협업필터링을 이용한 상품그룹 추천, 한국IT서비스학회지, Vol.6, No.2, pp.113-123 2007.

[3]이석준, 김선옥 (2007). 협업필터링에서 고객의 평가치를 이용한 선호도 예측의 사전평가에 관한 연구, 경영정보학연구, Vol. 17, No. 42, pp. 187-206, 2007.

[4]이석준, 김선옥, 이희춘 (2007). Pre-Evaluation for Detecting Abnormal Users in Recommender System, Journal of the Korean Data & Information Science Society, Vol. 18, No. 3, pp. 619-628, 2007.

[5]이희춘 (2006). On the Effect of Significanceof Correlation Coefficientfor Recommender System, Journal of the Korean Data & Information Science Society, Vol. 17, No. 4, pp. 1129-1139, 2006.

[6] 이희춘, 이석준, 정영준 (2006). The Effect of Co-rating on the Recommender System of User Base, Journal of the Korean Data & Information Science Society, Vol. 17, No. 3, pp. 775-784.

[7] Kim, S. O. and Lee, S. J. (2007). The Effect of Data Sparsity on Prediction Accuracy in Recommender System, Journal of the Korean Society for Internet Information, Vol. 8, No. 6, pp. 9-15.

[8] Kim, S. O., Lee, S. J. and Lee, H. C. (2008). A study on Improvement of Prediction Accuracy by Critical Value, Journal of the Korean Data Analysis Society, Vol. 10, No. 1, pp. 591-601.

[9] Konstan, B., Miller, D., Maltz, J., Herlocker, L., Gordon, J. and Riedl, J. (1997).GroupLens:Applying Collaborative FilteringtoUsenetNews, Communications of the ACM, Vol. 40, No. 3, pp. 77-87.

[10] Melville, P., Mooney, R. and Nagarajan, R. (2002). Content-Boosted Collaborative Filtering for Improved Recommendations, Proceedings of the eighteenth national Conference on Artificial Intelligence, pp. 187-192, 2002.

[11] Pazzani, M. J. (1999). Framework for Collaborative, Content_Based and Demograghic Filtering, Artificial Intelligent Review, pp. 394-408, 1999.

[12]Resnick, P. N., Iacovou, M., Suchak, P., Bergstrom, J. and Riedl, J. (1994).GroupLens: an open architecture for collaborative filtering of netnews, In Proceedings of the 1994 ACM conference on Computer supported cooperative work, pp. 175-186, 1994.