협력적 필터링에서 추가정보를 이용한 선호도 예측 정확도 향상에 관한 연구

이희춘*, 이석준**, 김선옥***

*상지대학교 컴퓨터데이터정보학과, **상지대학교 경영정보학과,
***한라대학교 정보통신방송공학부

A Study on improvements of prediction accuracy using additional information in collaborative filtering

Lee, Hee-Choon, Lee, Seok-Jun, Kim, Sun-Ok
Halla University, Sangji University, Sangji University
E-mail :choolee@sangji.ac.kr, digitaldesign@sangji.ac.kr, sokim@halla.ac.kr

요 약

본 연구는 협력적 필터링 기법을 이용한 선호도 예측 과정에서 발생하는 추가 정보를 이용하여 선호도 예측 정확도를 향상시킬 수 있는 방안에 대하여 연구하였다. 본 연구에서는 특정 상품에 대한 목표 고객의 선호도 예측에 선정된 이웃의 수와 선호도 예측 정확도와의 관계를 분석하였다. 분석을 위하여 선호도 예측 과정에 선정된 이웃의 수를 4분위수로 4집단으로 구분하여 구분 집단 간 선호도 예측 정확도에 차이가 나타남을 알 수 있었으며 각 집단의 예측 오차들의 평균들을 이용하여 선형의 보정함수를 제안한다. 본 연구의 결과를 바탕으로 추천시스템에서 이웃 수를 이용한 보정함수를 이용하면 예측 정확도를 높일 수 있다.

1.서론

추천시스템은 전자상거래에서 거래되는 다양한 상품에 대한 정보 중 고객의 선호도 성향과 가장 부합할 수 있는 상품을 자동적으로 예측하여 고객에게 필터링된 정보만을 고객에게 제시할 수 있다. 이를 통하여 고객이 전자상거래 사이트에서 직접 자신의 선호도에 부합하는 상품을 찾기 위해 검색하여야 하는 많은 정보들에게 빼앗기는 시간과 비용을 줄이는 효과를 얻을 수 있으며 또한 고객들이 알지 못하던 새로운 상품 정보의 획득과 같은 효과를 얻을 수 있다. 또한 선호도 예측력이 우수한 추천시스템의 경우 고객의 특별한 선호 성향을 예측할 수 있기 때문에 개인화 서비스를 제공할 수 있다.

추천시스템에서 협력적 필터링 기법은 전자상거래 추천 알고리즘에서 가장 핵심적인 기법으로 알려져 있으며 초기의 내용 기반의 추천시스템의 단점을 보완하고 있다.

협력적 필터링 기법의 가장 일반적인 알고리즘은 이웃 기반의 협력적 필터링 알고리즘으로 이웃 고객들의 상품에 대한 선호 경향을 반영하여 특정 상품에 대한 추천 대상 고객의 선호도를 예측한다. 일반적으로

이웃 기반의 협력적 필터링 알고리즘은 다음과 같은 단계로 추천 대상 고객의 특정 상품에 대한 선호도를 예측한다(Herlocker 등, 1999).

- 1단계: 추천 대상 고객의 선호도 예측을 위한 이웃의 선정과 두 고객 간의 선호도 유사정도 측정.
- 2단계: 선호도 유사정도에 따른 예측 대상 이웃의 선정.
- 3단계: 예측 알고리즘을 이용하여 추천 대상 고객의 선호도를 예측.

2. 선호도 예측 알고리즘

협력적 필터링 선호도 예측 알고리즘은 일반적으로 선호도 예측 대상 고객과 이웃 고객 간 상품들에 대한 선호도 유사 정도를 나타내기 위하여 다양한 형태의 유사도 가중치로 정의될 수 있으며 본 연구에서는 Pearson 상관계수를 이용한다(Breese 등, 1998). 선호도 예측 대상 고객 와 이웃 고객의 선호도 유사 정도를 나타내는 유사도 가중치 는 식(1)과 같이 Pearson 상관계수로 정의한다.

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

식(1)에서 R은 상품에 대한 고객의 선호도 평가치로 5점 척도로 되어 있으며 R_{ui}는 상품 i에 대한 선호도 예측 대상 고객 u의 평가치이며 R_{ji}는 상품 i에 대한 고객 u의 이웃 고객 j의 평가치이다. $\overline{R_u}$와 $\overline{R_j}$는 고객 u와 고객 j가 상품들에 대한 평가치의 평균이다.

2-1. NBCFA

이웃 기반의 협력적 필터링 알고리즘(Neighbor Based Collaborative Filtering Algorithm)은 추천 대상 고객의 선호도 평가치와 이웃으로 선정된 고객의 선호도 평가치를 이용하여 다음 식(2)와 같이 정의된다(Resnick 등, 1994).

$$\widehat{U_x}=\overline{U}+\frac{\sum_{J\in Raters}(J_x-\overline{J})r_{uj}}{\sum_{J\in Raters}|r_{uj}|},\quad where\ \ \overline{J}=\frac{\sum_{i=1}^{n}J_i}{n},\ i\neq x \quad (2)$$

여기서,

- $\widehat{U_x}$: 선호도 예측 대상 상품 x에 대한 선호도 예측 대상 고객 u의 선호도 예측치
- $\overline{U}$: 선호도 예측 대상 고객 u가 평가한 모든 상품에 대한 선호도 평가치 평균
- J_x : 선호도 예측 대상 상품 x에 대한 이웃 고객 j의 선호도 평가치
- $\overline{J}$: 이웃 고객 j가 평가한 모든 상품에서 선호도 예측 대상 상품 x에 대한 평가치를 제외한 선호도의 평균
- r_{uj} : 선호도 예측 대상 고객 x와 이웃 고객 j의 선호도 유사 정도를 나타내는 유사도 가중치

2-2. CMA

NBCFA는 선호도 예측에서 자신의 선호도 경향을 나타내는 $\overline{U}$와 이웃의 선호 경향을 나타내는 $\overline{J}$가 너무 과도하게 자신의 경향을 반영하기 때문에 이를 조정할 필요성이 있으며 이를 위하여 예측 대상 고객 u와 이웃 고객 j가 동시에 선호도를 평가한 상품만을 이용한 $\overline{U}_{match}$와 $\overline{J}_{match}$를 이용하는 대응평균 알고리즘(Correspondence Mean Algorithm)이 제안되었다(Lee, 2006).

$$\widehat{U_x}=\overline{U}_{match}+\frac{\sum_{J\in Raters}(J_x-\overline{J}_{match})r_{uj}}{\sum_{J\in Raters}|r_{uj}|} \quad (3)$$

2-3. 평가척도

선호도 예측 알고리즘의 예측 정확도는 test dataset의 실제 선호도 평가치와 이에 대한 선호도 예측치의 절대 오차 평균인 MAE(mean absolute error)를 이용하여 평가하며 다음 식(4)와 같이 정의한다.

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|R_{uj} - \widehat{R_{uj}}| \qquad (4)$$

3.실험

3-1. 실험 데이터의 구성

본 연구의 실험을 위하여 GroupLens에서 공개한 MovieLens 100K dataset을 이용하였다. 100K dataset는 943명의 고객이 1682편의 영화에 1에서 5점사이의 선호도가 평가된 총 100,000개의 평가치로 구성되어 있다. 보정함수를 이용한 선호도 예측 정확도의 향상을 비교하기 위하여 80%의 훈련 데이터(training data)와 20%의 실험 데이터(test data)로 랜덤하게 나누어 실험에 사용하였다. 또한 선호도 예측에서 데이터의 희소성에 따라 예측 성능의 효과도 분석하기 위하여 80%의 훈련 데이터에서 랜덤하게 10%씩 추출하여 구성한 70%, 60%의 훈련 데이터를 구성하여 희소성이 증가한 데이터에서의 보정함수의 성능을 측정하였다.

3-2. 실험방법

본 연구에서 제안하는 보정함수의 성능을 평가하기 위하여 NBCFA와 CMA를 이용한 선호도 예측 결과와 실험 데이터의 실제 선호도 예측 결과의 오차를 구하였다. 그리고 선호도 예측 과정에서 얻어진 이웃 수의 정보를 이용하여 4집단으로 구분한 후 각 집단의 오차의 특성을 파악하였다. 파악된 특성을 이용하여 원래의 선호도 예측 결과에 오차를 줄일 수 있는 선형 보정함수를 도출하였다.

3-3. 실험결과

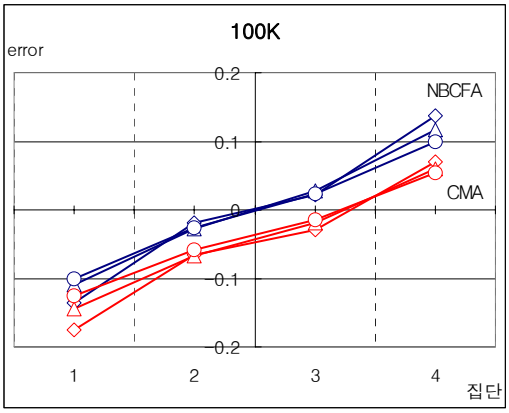
다음 <표 1>은 훈련 데이터의 희소성이 상대적으로 증가된 6:2의 dataset과 희소성이 상대적으로 감소된 8:2의 dataset에 대한 NBCFA(NB)와 CMA(CM)의 예측 오차에 대한 4개의 구분 집단 간 평균 차이에 대한 분산 분석 결과이다.

<표 1> 분산분석 결과

data set	알고리즘	구분	제곱합	자유도	평균 제곱	F	유의확률
6:2	NB	간	189.6	3	63.20	67.79	0.000**
		내	18601.5	19955	0.93		
		합계	18791.1	19958			
	CM	간	153.9	3	51.29	56.98	0.000**
		내	17961.6	19955	0.90		
		합계	18115.5	19958			
7:2	NB	간	135.3	3	45.12	49.38	0.000**
		내	18240.1	19965	0.91		
		합계	18375.4	19968			
	CM	간	109.6	3	36.55	41.51	0.000**
		내	17577.3	19965	0.88		
		합계	17686.9	19968			
8:2	NB	간	105.9	3	35.29	39.23	0.000**
		내	17963.0	19969	0.90		
		합계	18068.9	19972			
	CM	간	85.3	3	28.44	32.86	0.000**
		내	17279.0	19969	0.87		
		합계	17364.3	19972			

* : p<0.05, ** : p<0.01

4개의 분할 집단의 평균을 보면 다음 <그림 1>과 같이 선형의 관계가 나타남을 알 수 있으며 이를 이용하여 식(5)와 같은 보정함수를 이용한 새로운 예측 알고리즘을 제안할 수 있다.

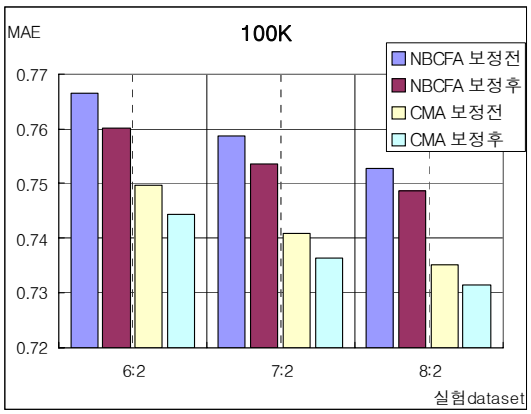


<그림 1> 구분 집단의 평균

$$\widetilde{R}_x = \widehat{R}_x + f(N_x) \quad (5)$$

식(5)는 식(2)와 (3)에서 제시된 NBCFA와 CMA에 의해 생성된 특정 사용자 x의 상품들에 대한 개별 예측치에 선호도 예측 과정에서 선정된 특정 사용자 x의 이웃 수인 N_x를 이용한 선형보정함수 $f(N_x)$을 적용한 새로운 예측치 $\widetilde{R}_x$를 제안하고 있다.

보정함수 적용 전 선호도 예측 결과와 보정함수 적용 후 선호도 예측 결과의 성능을 통계적으로 비교한 결과로 다음 <표 2>와 같은 결과를 얻었으며 훈련 데이터의 희소 정도에 따라 예측 정확도가 달라지고 있지만 보정함수에 의한 예측치의 보정이 기존의 정확도보다 우수함을 알 수 있다.



<그림 2> 보정함수 적용 전,후의 예측 성능 비교

4.결 론

본 연구는 협력적 필터링 기법을 이용한 선호도 예측 과정에서 이웃의 수와 선호도 예측 정확도와의 관계를 분석하였다. 선호도 예측 과정에 선정된 이웃의 수를 4분위수로 4집단으로 구분하여 구분한 집단 간 선호도 예측 정확도에 차이가 나타남을 알 수 있었으며 각 집단의 예측 오차들의 평균들을 이용하여 선형의 보정함수를 제안할 수 있었다. 제안한 보정함수를 통하여 100K의 6:2, 7:2, 8:2 실험 dataset 모두에서 보정함수를 적용하여 선호도 예측 정확도를 향상시킬 수 있었다.

[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]이희춘 (2006). Improved algorithm for user based recommender system. Journal of Korean Data & Information Science Society, Vol. 17, No. 3, pp. 717-726.

[4] 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.

[5] Breese, J., Heckerman, D. and Kadie, C. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering, In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence, pp. 43-52.

[6] Herlocker, J., Konstan, J., Borchers, A. and Riedl, J. (1999). An algorithm framework for performing collaborative filtering. In Proceedings of the 1999 Conference on Research and Development in Information Retrieval, pp. 230-237.

[7] Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P. and Riedl, J(1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews, In Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, pp. 175-186.