협력적 필터링 추천시스템에서 이웃의 수를 이용한 선호도 예측보정 방법

이석준* 김선옥** 이희춘***
(Seok-Jun Lee, Sun-Ok Kim, Hee-Choon Lee)

요 약 본 연구는 웹상에서 거래되는 아이템을 고객에게 추천하는 추천시스템에서 추천대상 고객의 정보와 이웃 고객의 정보를 이용한 협력적 필터링 추천기법에서 선호도 예측을 위해 필요한 이웃의 수가 선호도 예측 정확도에 영향을 주고 있음을 제시하고 이를 이용한 선호도 예측치의 보정 방법에 대하여 제안한다. 본 연구의 제안을 위하여 이웃 기반의 협력적 필터링 알고리즘과 대응평균 알고리즘을 이용하여 MovieLens 1 million dataset에 대하여 선호도 예측 정확도를 분석하고 분석결과를 토대로 개별 선호도 예측에 소요된 이웃의 수와 예측 정확도의 관계를 분석하였다. 분석결과를 이용하여 이웃 수에 따라 선호도 예측 결과를 다수의 집단으로 구분하여 각 집단에서 이웃의 수를 이용한 선호도 예측 정확도 향상에 대한 방법을 제안한다. 본 연구의 제안을 통하여 기존 선호도 예측 알고리즘으로 생성된 예측 결과에 선호도 예측 과정에서 부가적으로 발생한 정보를 추가하여 최종 예측 결과를 향상시킬 수 있을 것으로 기대한다.

핵심주제어 : 부가정보, 추천시스템, 협력적 필터링

Key Words : additiomal information, recommender system, collaborative filtering

Ⅰ. 서 론

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[2][3].

추천시스템은 전자상거래에서 고객 자신의 선호도에 부합하는 상품정보 수집뿐만 아니라 가장 적합한 상품까지 자동적으로 고객에게 제시하는 시스템이다. 추천시스템은 인터넷에서 고객의 정보탐색 비용을 절감시켜주며 고객의 구매 만족도를 높여줄 수 있는 마케팅 도구이다. 또한 우수한 품질의 추천이 고객에게 제공되면 웹사이트에 대한 고객의 충성도를 높일 수 있다. 고객에 대한 추천 접근법에 따라 추천시스템은 내용기반(content-based)과 협력적 필터링(collaborative filtering) 기법으로 크게 나눌 수 있다[7][8][9]. 내용기반 추천기법은 고객과 상품에 대한 정보를 문자화시킨 프로파일을 구성하여야 하기 때문에 고객과 상품에 대한 정보를 충분히 반영하기 어려운 문제점을 가지고 있다. 또한 고객과 상품의 관계에서 고객의 과거 구매이력만을 고려하기 때문에 추천 상품에 대한 과도한 특성화가 이루어지는 문제점도 가지고 있다 (김용수, 2006)[2]. 협력적 필터링 추천기법은 고객과 상품의 세부적인 내용은 의도적으로 무시하고 이들의 관계만을 이용하는데 일반적으로 상품에 대한 고객의 선호도 평가와 같은 명시적 수치자료를 이용하여 특정 상품에 대한 선호도를 예측하고 결과에서 높은 예측치를 부여받

* 상지대학교 경상대학 경영정보학과 교수
** 한라대학교 정보통신방송공학부 교수
*** 상지대학교 이공과대학 컴퓨터데이터정보학과 교수

은 상품을 고객에게 추천하는 방법을 취하고 있다. 협력적 필터링 추천기법은 고객 자신의 선호 경향과 선호도가 유사한 이웃 고객을 선정하고 이들의 관계를 이용하여 선호도 예측을 한다. 협력적 추천 시스템은 시스템 구축 초기에 고객의 수가 적어 유사 선호도를 가진 이웃 고객을 선정할 수 없어 추천이 불가능한 문제점을 가지고 있으나 현재 상업적으로 가장 성공적인 추천기법으로 알려져 있으며 Amazon.com과 같은 거대 전자상거래 업체들이 이용하고 있으며 학문적으로도 많은 연구가 이루어지고 있는 분야이다[4][6][8]. 본 연구는 협력적 필터링 알고리즘을 이용하여 고객의 선호도를 예측하는 과정에서 생성되는 추가정보들의 활용방안에 대하여 연구하였다.

II. 협력적 필터링

2-1. 선호도 예측 알고리즘

추천시스템에서 협력적 필터링 기법을 이용한 선호도 예측 접근법은 상업용 추천시스템에서 성공적으로 적용되고 있으며 주요 전자상거래 사이트에서 적용되고 있다. 협력적 필터링의 개념은 특정 상품에 대한 추천대상 고객의 선호도 예측을 위하여 이웃 고객의 선호도를 이용하는 것이며 고객의 선호도를 이용할 경우인 사용자 기반의 접근법과 상품의 선호도 유사정도를 이용하는 아이템 기반의 접근법으로 나눌 수 있다. 일반적으로 대규모의 상품이 거래되는 전자상거래에서는 고객 간의 선호도 정보를 이용하는 것 보다 아이템 간의 선호도 정보를 이용하는 경우가 많다.

협력적 필터링 기법의 적용은 다음 <그림1>과 같이 이웃 선정 과정에서 시작된다.

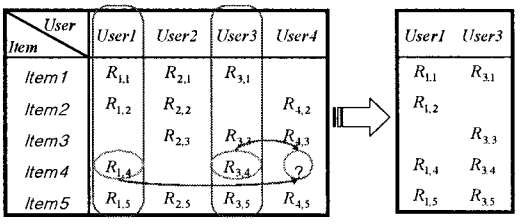

그림 1. 이웃 선정 과정

<그림1>에서 item4에 대한 user4의 선호도를 예측하기 위하여 먼저 예측대상 상품인 item4에 선호도를 평가한 다른 고객들의 선호도가 필요하다. 이때 선호도를 평가한 고객이 협력적 필터링 알고리즘 적용의 이웃으로 선정된다.

이웃 선정 과정이 이루어지면 선정된 이웃의 정보를 이용하여 item4에 대한 user4의 선호도를 예측하게 되는데 이웃 기반의 협력적 필터링 알고리즘(Neighbor Based Collaborative Filtering Algorithm)이 널리 이용되고 있다[7]. 다음 수식(1)과 <그림2>는 NBCFA의 정의와 적용 과정을 보여준다.

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

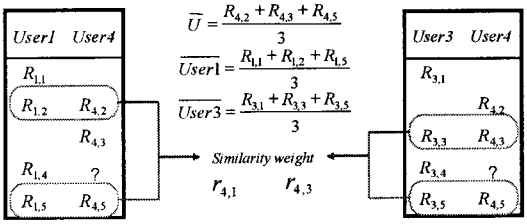

그림 2. NBCFA를 이용한 선호도 예측 과정

수식(1)에서 $\overline{U}$ 는 추천대상 고객인 user4의 선호도 평가치의 평균이고 $\overline{J}$ 는 이웃으로 선정된 user1과 user3의 선호도 평가치의 평균이다. J_x 는 예측대상 상품 x 에 대한 이웃의 선호도 평가치로 item4에 대한 user1과 user3의 선호도 평가치이다. 또한 추천대상 고객인 user4와 각 이웃 고객과의 선호도 유사정도를 평가하기 위하여 유사도 가중치를 정의하는데 본 연구에서는 피어슨 상관계수를 이용하여 고객 간 선호도 유사정도를 정의하였다[9][10].

NBCFA의 적용에서 추천대상 고객의 선호도 평균과 이웃 고객의 선호도 평균은 각각 추천대상 고객과 이웃 고객의 선호도를 대표한다. 하지만 유사도 가중치인 피어슨 상관계수는 추천대상 고객과 이웃 고객이 서로 공통으로 평가한 상품에 대해서만 고려된다. 이러한 문제는 각각의 평균이 과도한 자신의 성향을 반영하는 문제를 가질 수 있기 때문에 이를 개선한 대응평균 알고리즘(Correspondence Mean Algorithm)이 제안되었다[5][6].

다음 수식(2)와 <그림3>은 NBCFA를 개선한 CMA를 설명한다.

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

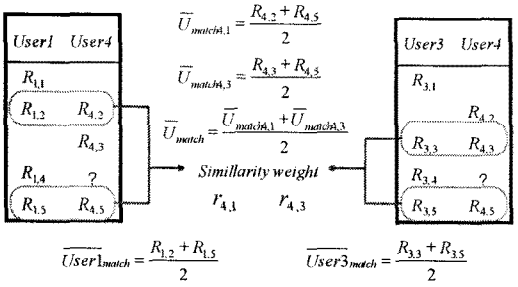

그림 3. CMA를 이용한 선호도 예측 과정

수식(2)에서 $\overline{U}_{match}$ 는 추천대상 고객인 user4와 이웃 고객인 user1, user3과 공통으로 평가한 평가치의 평균을 다시 평균하여 이웃 간의 관계를 보정한다. 또한 $\overline{J}_{match}$ 는 이웃으로 선정된 user1과 user3가 추천대상 고객과 공통으로 평가한 선호도 평가치의 평균이다. 이로써 이웃과의 공통성을 부가한 평균을 이용하기 때문에 과도하게 반영된 선호경향을 보정할 수 있다.

2-2. 사전평가 방법

협력적 필터링 알고리즘을 이용한 선호도 예측 정확도의 경우 이웃의 선정 과정과 유사도 가중치 계산과정에 많은 자원이 소요된다. 그렇기 때문에 선호도 예측 알고리즘 적용 전 주어진 data에서 개략적으로 선호도 예측 정확도에 대한 평가의 가능성이 연구되었다[4].

다음 <그림4>는 사용자 기반에서 개별 고객이 평가한 선호도 평가치를 이용하여 선호도 예측 이전에 정확도를 평가할 수 있는 방법들을 제시하고 있다.

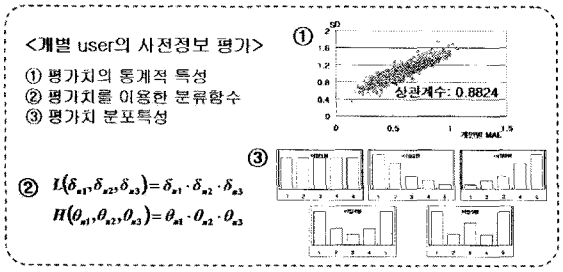

그림 4. 사전평가 방법

사전평가는 개별 고객이 평가한 선호도 평가치의 분산을 이용한 방법, 평가치의 패턴을 이용한 분류함수의 적용, 개별 고객의 선호도 평가치 분포에 대한 분포적합도 검정의 적용의 방법이 제안되었다.

2-3. 부가정보의 활용

본 연구에서는 선호도 예측 과정에서 생성된 부가정보의 활용에 대해 연구하였다. 부가정보란 사전정보와 달리 선호도 예측 과정 중 이웃의 선정과정과 유사도 가중치 계산과정에서 발생하는 정보를 말한다. 여기서 이웃 선정과정에서는 선호도 예측에 사용되는 이웃의 수가 부가정보로 생성되며 유사도 가중치의 계산과정에서는 추천대상 고객과 이웃 고객의 평가치들에서 공통으로 평가한 응답쌍의 수가 발생한다. 본 연구에서는 이웃 선정과정에서 생성된 이웃 수를 이용하여 선호도 예측 정확도를 향상시키기 위한 방법을 제시한다.

III. 실험

3-1. 실험 dataset

본 연구에 사용된 실험 dataset은 GroupLens에서 공개한 MovieLens 1million dataset이다. 1million dataset은 6040명의 고객이 3952편의 영화에 1에서 5점사이의 점수로 선호도가 평가되어 있으며 각 고객들은 최소 20편의 영화에 평가하도록 되어 있다.

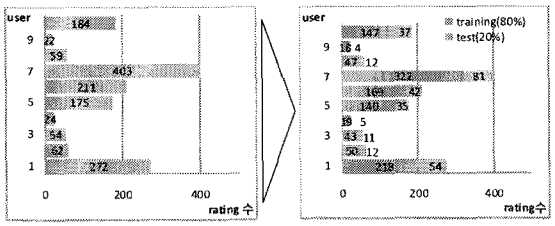

그림 5. 실험 dataset의 구성

실험을 위하여 80%의 training dataset과 20%의 test dataset을 구성하였다. 이때 개별 고객이 평가한 선호도 평가치에서 각각 80%와 20%로 랜덤하게 추출하여 전체 고객의 80%를 취합하여 training dataset을 구성하였고 20%를 취합하여 test dataset을 구성하였다.

3-2. 실험 방법

선호도 예측 과정에서 생성되는 부가정보를 이

용하여 선호도 예측 정확도를 향상시키기 위한 방법을 제안하기 위하여 실험 dataset에 대하여 NBCFA와 CMA를 이용하여 선호도를 예측하였다. 일반적으로 선호도 예측 정확도를 평가하기 위하여 절대 편차의 평균인 MAE를 이용하지만 본 연구에서는 test datset의 실제 선호도 평가치와 예측 선호도 평가치의 오차를 이용하여 부가정보와의 관계를 파악하였다. 부가정보와 선호도 예측치와의 관계를 파악하기 위하여 부가정보로 얻어진 이웃 수의 4분위수로 4집단으로 구분하였다. 1million dataset을 이용한 실험 dataset의 결과 이웃 수는 최소 1명에서 최대 2699명이 선정되었다. 다음 <그림6>은 구분 집단에서 예측 오차의 평균을 나타내고 있다.

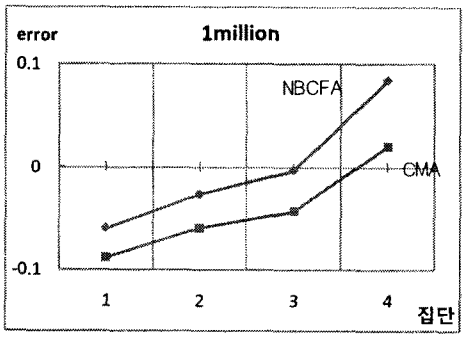

그림 6. 구분 집단에서 예측 오차의 평균

<그림6>의 결과를 통하여 이웃의 수에 따라 선호도 예측치가 과소 또는 과대 예측되고 있음을 알 수 있다.

다음 <표1>은 구분 집단 간 예측 오차의 평균에 대한 차이의 검정결과이다. 결과를 통하여 이웃 수로 구분된 집단 간 오차의 평균에는 차이가 있음을 알 수 있다.

표 1. 이웃 수에 따른 집단 간 오차평균 검정

오차	구분	제곱합	자유도	평균제곱	F	유의확률
NBCFA	집단-간	568.3	3	189.44	227.30	0.000**
	집단-내	166652.2	199959	0.83	-	
	합계	167220.5	199962	-		
CMA	집단-간	313.4	3	104.46	132.19	0.000**
	집단-내	158008.8	199959	0.79	-	
	합계	158322.2	199962	-		

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

<그림6>과 <표1>의 결과를 통하여 오차를 보정할 수 있는 수식의 제안이 가능하다.

3-3. 실험 결과

실험을 통하여 수식(1)과(2)에서 제시된 NBCFA와 CMA에 의해 선호도 예측치의 오차를 개선하기 위하여 다음과 같이 이웃의 수를 이용한 선형 보정함수를 제안할 수 있다.

$$New\ \hat{U}_x = \hat{U}_x + f(N) \qquad (3)$$

수식(3)에서 NBCFA와 CMA에서 생성된 선호도 예측치에 선호도 예측 과정에서 획득된 추가정보인 이웃 수를 이용하여 예측 오차를 최소화 시켜 줄 수 있는 선형 보정함수인 $f(N)$을 추가하여 다음 <그림7>과 같은 선호도 예측 오차의 개선을 가져올 수 있었다.

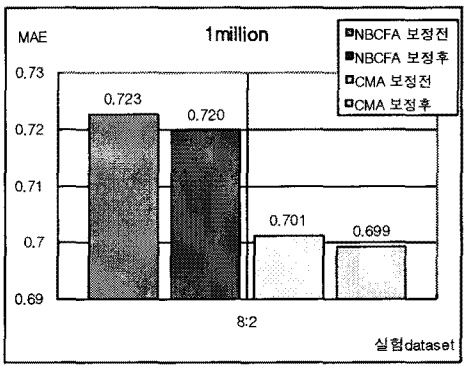

그림 7. 보정함수 적용 전/후의 예측 정확도

또한 다음 <표2>는 각 알고리즘에 의해 얻어진 기존의 예측 정확도와 보정함수를 적용한 예측치의 차이에 대한 검정 결과이다.

표 2. 보정 전/후의 예측치의 t검정 결과

알고리즘	대응차평균	t값	유의확률
NBCFA	0.0028	20.513	0.000**
CMA	0.0020	14.962	0.000**

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

실험 결과를 통하여 선호도 예측 과정에서 얻을 수 있는 부가정보를 이용하여 선호도 예측 오차를 줄일 수 있는 방법이 있음을 알 수 있다. 그러나 실험에서 얻어진 부가정보가 사전평가와의 관련성은 아직 연구되지 않았다. 만약 사전평가 방법과의 관계가 밝혀지면 적정 수준의 이웃 수를 이용하여 선호도 예측 결과를 보정할 수 있는 보정함수를 개발 할 수 있을 것으로 기대된다.

Ⅳ. 연구의 의의

본 연구는 협력적 필터링 알고리즘을 이용한 고객의 선호도 예측 과정에서 부가적으로 생성된 정보를 이용하여 선호도 예측 정확도를 개선할 수 있는 방법에 대하여 제안하고 있다. 아직 그 예측 정확도의 개선이 미약하지만 추가적 연구를 통하여 개선효과가 큰 보정함수를 제안 할 수 있는 기초를 제시하고 있다. 또한 기존의 사전평가와 같은 연구결과와 연계할 수 있다면 협력적 필터링 알고리즘에서 이웃에 따른 시스템 부하를 줄이는 동시에 선호도 예측 정확도를 개선할 수 있는 방법이 개발될 것으로 기대한다. 본 연구의 목적은 차기 정교한 선호도 예측 정확도 보정 함수의 개발에 기초를 마련하고 부가정보의 활용에 대한 가능성을 제시하는데 그 목적이 있다.

참 고 문 헌

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

[2] 김용수 (2005), "전자상거래에서 고객의 탐색 및 행동 패턴을 고려한 추천시스템의 개발", 한국과학기술원, 박사학위 논문.

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

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

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

[6] 이희춘, 이석준 (2006). "사용자 기반 추천시스템에서 근접이웃 알고리즘과 수정알고리즘의 예측 정확도에 관한 연구", *Journal of the Korean Data Analysis Society*, 8, No.5, 1893-1904.

[7] 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.

[8] 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.

[9] 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.

[10] 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.