Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 3, pp. 717 ~ 726

Improved Algorithm for User Based Recommender System[1)]

Hee Choon Lee[2)]

Abstract

This study is to investigate the MAE of prediction value by collaborative filtering algorithm originated by GroupLens and improved algorithm. To decrease the MAE on the collaborative recommender system on user based, this research proposes the improved algorithm, which reduces the possibility of over estimation of active user's preference mean collaboratively using other user's preference mean. The result shows the MAE of prediction by improved algorithm is better than original algorithm, so the active user's preference mean used in prediction formula is possibly over estimated.

Keywords : 추천시스템, Collaborative filtering, MAE

1. 머리말

초고속 인터넷 인프라의 확산과 인터넷의 대중화로 다양한 형태의 전자상거래(e-commerce)가 활발하게 진행되고 있으며 주요한 거래수단으로 대두되고 있다. 전자상거래에서 제공되는 거래는 물품의 거래뿐만 아니라 영화나 음악 등과 같은 서비스와 개인의 취향에 맞춘 여행상품과 같은 다양한 형태의 서비스가 급성장하고 있다. 우리나라에서는 인터넷을 통한 매출과 가입자의 증가율과 같은 지표가 과거의 지표에 비해 줄어들고 있는 시장 성숙기에 해당하는 특성을 보여주고 있다. 이로 인해 전자상거래 영역에서의 기업들은 더욱 더 치열해진 경쟁에서 생존하기 위해 다양한 마케팅 전략을 구사하여야 하고 고객들도 기존의 서비스에 비해 차별화된 서비스를 원하고 있다. 개인화 추천시스템은 자동화된 정보필터링 기술을 적용하여 사용자의 취향에 맞는 상품을 추천해 주는 시스템이다(정경용(2005)). 이미 추천시스템은 Amazon.com, CD Now.com, e-Bay등과 같은 전자상거래 사이트에 적용되어 성공적

1) 이 논문은 2004년도 상지대학교 교내 연구비 지원에 의한 것임.
2) 강원도 원주시 우산동 660번지 상지대학교 컴퓨터데이터정보학과 교수
E-mail : choolee@sangji.ac.kr

으로 운영되고 있다(Schafer (1999) 참조).

본 논문은 GroupLens의 100K movielens data를 이용하여 GroupLens에서 제시한 dataset을 training datset 과 test dataset을 각각 80%, 20%로 분할한 5개 dataset을 구성하여 GroupLens에서 제시한 특정 사용자 이웃 기반의 예측알고리즘(neighborhood based algorithm)과 제안하는 예측알고리즘의 예측치와 평가치간의 MAE를 구하여 평가하였다.

2. 추천시스템의 분류

추천시스템은 정보필터링(Information retrieval) 기법을 이용하여 고객이 원하는 제품을 추천하기 위한 자동화된 기법을 말한다. 추천시스템은 처음으로 Goldberg(1992)에 의해 개념이 도입되어 수 십여년간 다양한 추천접근법과 알고리즘들이 개발되었으며 다양한 기법들이 등장하였다. 정보검색과 추천시스템은 동일한 개념으로 이해되기도 하지만 정보검색은 명시적인 사용자의 선호도에 대응되는 관련정보를 검색하는 것이고 추천시스템은 사용자의 특성, 사용자에 의해 구매된 아이템, 사용자의 행위 등과 같은 관찰된 정보가 잠재적 사용자의 선호도에 반영된다고 가정하고 사용자의 특성과 아이템간의 관계를 이용하여 잠재적 사용자의 선호도를 예측하고 특정 사용자에게 적합한 아이템에 대한 예측을 제공한다는 점에서 구분된다. 추천시스템은 추천접근법에 따라 속성기반(content-based), 협력적 필터링(collaborative filtering), 혼합적 필터링(hybrid), 지식공학적(Knowledge engineering) 접근법으로 나누어 볼 수 있다(Huang(2004)).

3. 관련연구

3.1 내용기반 필터링(content-based)

내용기반의 접근법은 아이템의 속성을 분석하기 위하여 정보검색(Information retrieval) 분야의 기법을 이용한다. 추천은 사용자가 이전에 경험한 아이템의 특성을 바탕으로 사용자의 프로파일을 구성하고 아이템에서 추출된 속성과의 일치성 정도를 이용하여 이루어진다. 내용기반의 접근법은 이전에 사용자가 경험한 아이템의 특성을 이용하여 유사한 특성을 지닌 아이템에게만 추천이 이루어진다는 제약을 가지고 있다. 내용기반 접근법의 근본적인 문제점은 단지 추천을 하고자 하는 목표 사용자 자신의 경험이나 피드백만이 아이템의 추천에 이용된다는 것이다. 다시 말해, 내용기반 추천의 문제점은 다른 사용자의 취향이 반영되거나 영향을 받음에도 불구하고 추천을 받고자 하는 개별 사용자들 자신의 우선적인 경험이 이용되는 문제점을 안고 있다(Balabanovic와 Shoham(1997)).

3.2 협력적 필터링(collaborative filtering)

협력적 필터링은 사용자와 아이템 간의 관계 데이터만을 이용하는 접근법(Hill(1995), Resnick(1994) 참조)이다. 협력적 필터링은 사용자와 아이템의 관계를 User×Item의 행렬로 표기하고 사용자간의 관계를 이용한 분석인 사용자 기반(user-based)과 사용자를 제외한 아이템간의 관계를 이용한(item-based)의 두 가지로 나눌 수 있다. 협력적 필터링은 가장 널리 이용되고 성공적인 추천 접근법으로 추천 알고리즘 연구의 근간을 이루고 있다(Breese(1998), Resnick(1994), Resnick(1997)).

3.3 선행연구

GroupLens는 넷뉴스의 뉴스 기사의 선호도 예측을 위하여 예측을 하고자 하는 문서에 대해 다른 사람이 평가한 선호도를 이용하는 사용자 이웃 기반의 예측 알고리즘(neighborhood based algorithm)을 이용하여 새로운 문서에 대한 선호도 예측을 하였다. 내용기반의 필터링은 문서에서 단어의 출현 혹은 비 출현의 빈도를 이용하여 사용자의 선호도를 예측하는 반면 사용자 이웃 기반의 예측 알고리즘은 이미 문서를 읽은 다른 사람의 견해를 이용하여 예측을 하게 된다. 이때 각 사용자 간의 유사도를 타나내기 위한 유사도 가중치(similarity weight)를 피어슨 상관계수(Pearson's correlation coefficient)를 이용하였으며 문서를 읽은 사람이 없다면 그 문서에 대한 예측을 할 수 없게 된다. 이는 추천시스템에서 문제가 되는 초기예측의 문제가 된다. 초기 추천시스템에서 사용된 알고리즘은(Resnick(1994), Shardanand(1995)) 아이템에 대한 사용자들의 선호도 평가치를 바탕으로 사용자간의 평가치의 거리를 계산하고 사용자가 아이템을 얼마나 좋아 하는지에 대한 예측은 그 아이템에 대해 근접한 이웃들의 집합에서 선호도의 가중 평균을 계산하여 이루어진다. 여기서 아이템에 대한 선호도를 표시하지 않은 사용자들은 무시된다. 사용자의 선호도는 사용자간의 성향을 구분하기 위해 척도화시킨 평가치를 사용한다(Herlocker(1999)). Breese(1998)는 협력적 필터링의 유사도 가중치에 대하여 피어슨 상관계수, 벡터 유사도(vector similarity), 기본선호도(default voting), 역사용자 빈도수(inverse user frequency), 사례확대(case amplification)의 방법을 이용하여 정확도와 유효범위를 향상시키는 것에 관한 연구와 기존의 확률적 방법인 베이지안(Bayesian) 방식의 모델 기반(model-based) 방법을 협력적 방법에 응용하였다. Herlocker(1999)는 협력적 필터링에 대한 다양한 평가방법을 제안하였으며 박지선(2001)은 협력적 필터링 기법에서 적용하고 있는 피어슨 상관계수를 이용하는 방법에서 비슷한 선호도 패턴을 가지는 고객들을 적절히 군집화하여 이 군집에 속하는 고객들의 평가를 기반으로 하여 협력적 필터링 기법을 수행하는 방법을 제안하였다. 김택헌(2004)은 ordered clustering 방법을 이용하여 each movie dataset에 대해 예측의 정확도를 실험하였다. 이희춘(2006)은 movielens dataset을 이용하여 예측의 정확도를 높이기 위해 인구통계변수와 사용자의 응답쌍의 영향력에 대해 연구했다.

(1) 사용자 이웃 기반의 예측 알고리즘

최초의 자동화된 협력적 필터링 알고리즘은 사용자 이웃 기반의 예측 알고리즘으로

GroupLens에서 특정 사용자에 대하여 다른 사용자들의 기사에 대한 선호도를 바탕으로 특정 사용자 자신의 선호도 평균과 이웃의 선호도 평균에 대한 가중치를 이용하여 특정 사용자가 접하지 않은 기사에 대한 선호도를 예측하였으며 이때 이웃 사용자와의 관계를 상관계수를 이용하여 가중하였다. 일반적으로 이웃과의 가중치는 피어슨 상관계수, 코사인 벡터, 스피어맨 순위상관계수 등이 사용되며 GroupLens에서 제시한 알고리즘에서는 피어슨 상관계수를 사용하였다(Herlocker(1999)).

다음은 GroupLens에서 제시한 문서에 대한 선호도 예측의 예제와 선호도 예측 알고리즘을 이용한 계산결과이다.

<표 1> 문서에 대한 각 사용자의 선호도

Item \ User	Ken	Lee	Meg	Nan
1	1	4	2	2
2	5	2	4	4
3			3	
4	2	5		5
5	4	1		1
6	?	2	5	

$$K_{6_{Pred}} = \overline{K} + \frac{\sum_{J \in raters} (J_6 - \overline{J}) r_{kj}}{\sum_j | r_{kj} |}$$

$$= 3 + \frac{2r_{km} - r_{kl}}{| r_{km} | + | r_{kl} |} = 3 + \frac{2 - (-0.8)}{| 1 | + | -0.8 |} = 4.56$$

<표 1>에서 Ken, Lee, Meg, Nan은 각각 문서를 접한 사용자가 되고 1열은 각 사용자들이 접한 문서, 즉 아이템의 종류가 된다. 예측은 사용자 Ken의 6번째 문서에 대한 선호도를 구하는 것으로 먼저 각 문서에 대한 사용자들의 선호도 평가들의 관계를 각 사용자들의 상관관계를 이용하여 관련성을 구하게 된다. 이때 사용자 간의 상관관계는 피어슨 상관계수, 코사인 벡터 등이 이용될 수 있으며, GroupLens가 제시한 알고리즘에서는 피어슨 상관계수가 사용되었다.

여기서 $\overline{K}$는 Ken의 선호도 평균이 되고, $\overline{J}$는 Ken에 이웃한 Lee, Meg, Nan 각각의 선호도 평균이 된다. r_{kj}는 Ken과 이웃한 사용자 Lee, Meg, Nan과의 피어슨 상관계수이다. 즉, r_{km}은 Ken과 Meg, r_{kl}은 Ken과 Lee의 피어슨 상관계수이다. Ken의 선호도 평균과 이웃의 선호도 평균과 상관계수를 이용하여 Ken의 6번째 아이템에 대한 선호도를 예측하면 4.56으로 구해진다.

(2) 제안 알고리즘을 이용한 협력적 필터링

GroupLens에서 제시한 사용자 이웃 기반의 예측 알고리즘을 이용한 협력적 필터링은 어떤 상품에 대한 특정 고객의 선호도를 예측하기 위하여 식 (2)의 피어슨 상관계수를 이용하여 유사한 선호도를 가지는 이웃들을 정하고 식 (1)에 의해 특정 사용자의 선호도 예측을 계산한다.

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

여기서

$$r_{uj} = \frac{\sum (U - \overline{U})(J - \bar{J})}{\sqrt{\sum (U - \overline{U})^2 \cdot \sum (J - \bar{J})^2}} \quad , \qquad -1 \le r_{uj} \le 1 \tag{2}$$

U_x는 아이템 x에 대한 특정 사용자 u의 선호도 예측치이고, r_{uj}는 특정 사용자 u와 이웃한 사용자 j의 상관관계를 나타내는 피어슨 상관계수이다. J_x는 이웃사용자 j의 아이템 x에 대한 선호도이고 $\bar{J}$는 이웃사용자 j의 선호도 전체의 평균이다. r_{uj}가 1에 가까울수록 두 고객의 선호도 경향이 매우 유사함을 나타내고 -1에 가까울수록 반대의 선호 경향을 나타낸다. Raters는 테스트 상품에 대해 선호도를 표시한 고객들을 의미한다. 협력적 필터링의 많은 연구들이 GroupLens에서 제시한 알고리즘을 이용하여 선호도를 예측하였다.

본 논문에서는 GroupLens에서 제시한 알고리즘과 제안 알고리즘 1, 2를 비교하여 예측의 정확도를 MAE를 구하여 비교하였다. GroupLens에서 제시한 알고리즘에서는 $\overline{U}$는 특정사용자가 나타낸 선호도 전체의 평균을 나타내고 있다. 그러나 특정 사용자 u와 j의 상관관계를 나타내는 가중치 r_{uj}는 u와 j가 공통으로 평가한 항목으로만 구하게 된다. 여기서 $\overline{U}$는 사용자 u의 선호도 전체의 평균을 이용하게 되면 사용자 u의 선호도가 과대평가되어 사용자 j의 선호도를 충분히 반영하지 못하게 될 가능성이 있다. 그래서 제안 알고리즘 1은 $\overline{U}$를 사용자 u의 선호도 전체의 평균과 사용자 u와 이웃 사용자 j가 공통으로 표기한 선호도의 평균들의 평균인 $\overline{U}_{match}$로 나누어 두 예측식의 정확도를 MAE를 구하여 비교하였다. 또한 사용자 u와 이웃 사용자 j가 공통으로 표기한 선호도를 이용하여 사용자 u의 선호도를 예측하기 때문에 이웃 사용자 j

의 선호도도 사용자 u와 공통으로 표기한 선호도만을 이용하는 것이 필요할 것으로 판단되어 $\bar{J}$를 사용자 u와 공통으로 표기한 아이템들에 대한 선호도인 $\bar{J}_{match}$로 나눈 제안 알고리즘 2를 이용한 예측식의 정확도를 MAE를 구하여 비교하였다.

제안 알고리즘 1과 제안 알고리즘 2는 다음 수식 (3),(4)와 같다.

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

$$U_x = \bar{U}_{match} + \frac{\sum_{J \in Raters} (J_x - \bar{J}_{match}) r_{uj}}{\sum_{J \in Raters} | r_{uj} |} \tag{4}$$

여기서, $\bar{U}_{match}$는 사용자 u와 이웃 사용자 j가 공통으로 표기한 아이템에 대한 사용자 u의 선호도 평균들의 평균이다. $\bar{J}_{match}$는 사용자 u와 이웃 사용자 j가 공통으로 표기한 아이템들에 대한 사용자 j의 선호도 평균이다.

4. 분석 결과

4.1 평가자료의 분석

본 논문에서 사용된 dataset은 GroupLens의 movielens dataset을 이용하여 실험을 하였다. movielens dataset은 943명의 평가자들은 1682편의 영화에 대해 최소 20편을 평가하였으며 최대 737편의 영화에 평가를 하였다. 평가점수는 1-5점으로 평가하였다. 1682편의 영화에 943명이 평가한 평가의 수는 100,000개이다. 본 논문에서는 GroupLens에서 제공되는 movielens dataset을 training dataset과 test dataset으로 각각 80%, 20%로 분할한 5개의 dataset을 이용하여 평가하였다.

(1) 응답 평가자의 응답편수의 빈도 분포

movielens dataset의 응답 평가자의 응답편수의 빈도 분포는 다음 <표 2>와 같다.

<표 2> 응답편수의 빈도 분포표

응답편수	빈도	퍼센트
20-25	137	14.5
26-30	76	8.1
31-40	92	9.7
41-60	144	15.3
61-80	87	9.2
81-100	46	4.2
101-150	435	13.9
151-200	78	8.3
200이상	148	15.7

movielens 자료는 943명의 응답자가 1682편의 영화 중최소 20편의 영화에 응답하였다. 응답자는 평균 106.0개의 영화에 대해 평가를 하였다.

(2) 응답쌍(pair of response)의 빈도분포

응답쌍(pair of response)의 빈도분포표는 다음 <표 3>와 같다.

<표 3> 응답쌍(pair of response)의 빈도 분포표

응답쌍의 수	빈도	퍼센트
0	15043	3.4
1-5	125478	28.2
6-10	93886	21.2
11-15	57135	12.8
16-20	34846	7.9
21-25	23389	5.3
26-30	16585	3.7
31-50	37329	8.4
51이상	39235	9.1

movielens dataset의 응답 평가자의 응답편수의 응답쌍의 빈도분포표에서 알 수 있듯이 응답쌍이 없는 경우가 3.4%이며 응답쌍이 5개 이하의 경우 31.6%를 차지하고 있어 이상점(outlier)가 있을 경우 상관계수는 이상점의 영향을 받을 수 있다.

4.2 성능평가

추천시스템의 정확성을 평가하기 위하여 일반적으로 MAE(Mean Absolute Error)가 사용된다(Breese et al.(1998), Herlocker et al.(1999), Shardanand and Maes(1995)). 본 논문에서 제시한 알고리즘과 기존의 알고리즘의 정확도의 평가도 MAE를 사용하

여 실제 선호도와 예측 선호도간의 정확도를 평가하였다. MAE는 선호도 예측치와 사용자의 실제 선호도 평가치 사이의 편차의 절대값의 평균으로 측정한다.

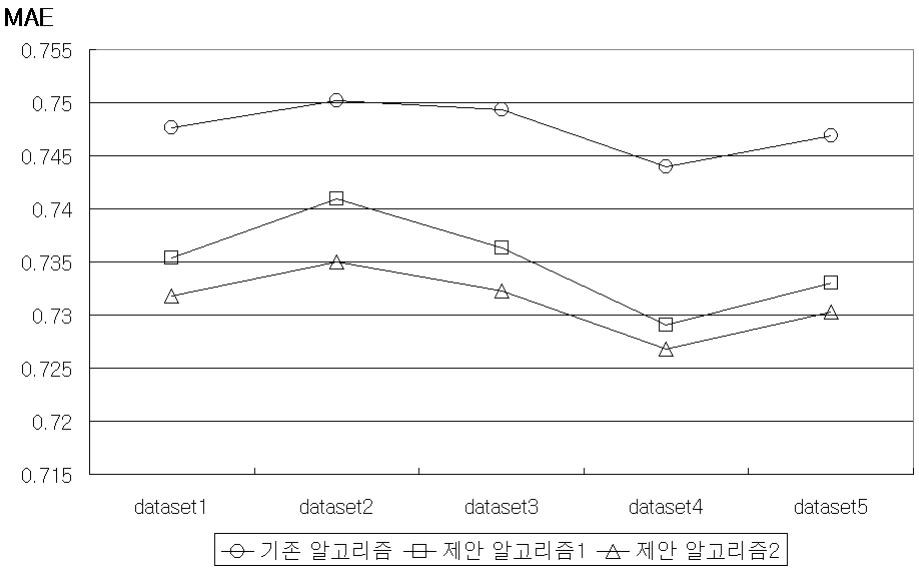
$$MAE = \frac{\sum_{i=0} |\varepsilon_i|}{N} \tag{6}$$

다음은 본 논문에서 제시한 알고리즘과 기존 알고리즘을 이용하여 5개 dataset에 대하여 예측력의 정확도에 대한 MAE분석 결과이다.

<표 4> dataset에 따른 MAE 결과표

	MAE		
	기존 알고리즘	제안 알고리즘 1	제안 알고리즘 2
dataset1	0.7475983	0.7353512	0.7317636
dataset2	0.7501989	0.7409647	0.7349743
dataset3	0.7493183	0.7363431	0.7322902
dataset4	0.7439199	0.729018	0.7268045
dataset5	0.7469254	0.7330533	0.7303055

<표 4>는 GroupLens에서 제시한 기존 알고리즘과 본 논문에서 제안한 알고리즘 1과 2의 MAE 결과표이다. 기존 알고리즘의 MAE에 비해 제안한 알고리즘 1, 2의 MAE가 작게 나타나 예측력이 우수함을 알 수 있다.



<그림 1>기존 알고리즘과 제안 알고리즘과의 dataset별 MAE비교

5. 결 론

본 논문에서 제안한 알고리즘과 기존의 알고리즘을 비교하였을 때 제안한 알고리즘 1, 2의 MAE가 작게 나타나 제안 알고리즘 1, 2가 예측력이 더 우수하다고 할 수 있다. 제안 알고리즘 1, 2의 비교에서는 제안 알고리즘 1에 비해 제안 알고리즘 2의 MAE가 작게 나타나 제안 알고리즘 2가 예측력이 더 좋다고 할 수 있다. 앞으로의 연구에서는 예측 평가순위를 고려하는 연구가 필요할 것으로 기대된다.

참고문헌

1. 김택헌, 양성봉 (2004). 협동적 필터링 기반 추천 시스템을 위한 향상된 이웃 선정 방법, 제 21회 *한국정보처리학회* 춘계학술발표대회. 11권, 1호, 453-456.
2. 박지선, 김택헌, 유영석, 양성봉 (2001). 추천 시스템을 위한 고객 클러스터링 방법을 적용한 예측 알고리즘. *한국정보과학회* 2001년 춘계학술대회, 제28권, 1호, 268-270.
3. 이희춘 (2006). An Exploratory Study for Decreasing Error of Prediction Value of Recommender System on User Based, *한국데이터정보과학회*, 제17권, 1호, 77-86.
4. 정경용, 이정현 (2005). 개인화 추천 시스템의 예측 정확도 향상을 위한 사용자 유사도 가중치에 대한 비교 평가, *전자공학회논문지-CI* 제42권, 6호, 63-74.
5. David Goldberg, David Nichols, Brian M. Oki, Douglas Terry (1992). Using collaborative filtering to weave an information tapestry, *Communications of the ACM*, 35-12, 61-70.
6. Herlocker, J., Konstan, J., Borchers, A., Riedl, J. (1999). An Algorithmic Framework for Performing Collaborative Filtering, *In Proceedings of the 1999 Conference on Research and Development in Information Retrieval,* 230-237.
7. J. Ben Schafer and Joseph A. Konstan and John Riedl. (1999). Recommender systems in E-commerce, *In Proceedings of ACM Conference on Electronic Commerce*, 158-166.
8. John S. Breese and David Heckerman and Carl Kadie. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering, *In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, 43-52.
9. Marko Balabanovic, Yoav Shoham. (1997). Fab: content-based, collaborative recommendation, *Communications of the ACM* , 40-3, 66-72.
10. Paul Resnick and N. Iacovou and M. Suchak and P. Bergstorm and J. Riedl. (1994). GroupLens: An Open Architecture for Collaborative

Filtering of Netnews, *In Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, 175-186.
11. Paul Resnick, Hal R. Varian. (1997). Recommender systems, *Communications of the ACM*, 40-3, 56-58.
12. Upendra Shardanand and Patti Maes. (1995). Filtering: Algorithms for Automating "Word of Mouth", *In Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, 1-1, 210-217.
13. Will Hill, Larry Stead, Mark Rosenstein, George Furnas. (1995). Recommending and Evaluating Choices in A Virtual Community of use, *Proceedings of the SIGCHI conference on Human factors in computing systems,* 194-201.
14. Zan Huang; Wingyan Chung; Hsinchun Chen. (2004). A graph model for E-commerce recommender systems, *Journal of the American Society for Information Science and Technology,* 55-3, 259-274.

[2006년 6월 접수, 2006년 8월 채택]