Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 3, pp. 775 ~ 784

The Effect of Co-rating on the Recommender System of User Base[1)]

Hee Choon Lee[2)] · Seok Jun Lee[3)] · Young Jun Chung[4)]

Abstract

This study is to investigate the effect of the number of co-rated users to the MAE. User based collaborative algorithm generally uses similarity weight to compute the relation of active user and other users. The original estimation algorithm of the GroupLens used the Pearson's correlation coefficient, soon after other researchers used various weighting. The Pearson's correlation coefficient and Vector similarity, which is used in the field of information retrieval, are commonly used to the estimation algorithm. In prediction, we analyze the effect of the number of co-rated users on the user based recommender system.

Keywords : 추천시스템, Collaborative filtering, MAE, Significant weight

1. 머리말

추천시스템(Recommender Systems)은 제품, 서비스, 정보(일반적으로 통칭하여 제품 혹은 아이템)를 잠재 고객에게 추천하기 위해 다양한 응용분야에서 이용되고 있다. 특히 전자상거래 (e-commerce)에서 추천시스템은 웹 기반 시스템에 통합되어 고객과 제품간의 관계(제품의 구매, 제품에 대한 선호도, 제품 카탈로그 검색 등과 같은 활동)에서 얻어지는 거래 데이터와 웹상에서 실시간으로 얻어지는 고객의 행동 데이터를 통하여 제품을 추천하기 위한 핵심적인 정보를 수집하게 된다. 이와 같이 고객과 제품간의 관계데이터는 고객의 인구통계학적 자료와 기업이 보유하고 있는 고객정보와 제품정보를 바탕으로 향상되어진다. 이 모든 데이터들은 고객의 선호도 예측 모형

1) 2004년도 강원대학교 학술연구조성비로 연구하였음.
2) 강원도 원주시 우산동 660번지 상지대학교 컴퓨터데이터정보학과 교수
E-mail : choolee@sangji.ac.kr
3) 강원도 원주시 우산동 660번지 상지대학교 경영학 박사과정
E-mail : crco909@yahoo.co.kr
4) 강원도 춘천시 효자2동 19-1번지 강원대학교 컴퓨터과학과 교수
E-mail : ychung@kangwon.ac.kr

과 알고리즘을 이용하여 고객이 원하는 추천을 생성하기 위한 추천시스템의 추천엔진에 투입되고 분석되어진다. 생성된 추천은 마케팅활동과 구매의사결정을 지원하기 위해 마케팅 전문가와 고객에게 제공된다.

추천시스템은 인구통계학적 자료를 이용한 추천, 상품이나 아이템의 속성을 이용한 내용기반 추천, 사용자와 아이템의 속성은 의도적으로 무시하고 단지 사용자-아이템의 관계 데이터(예: 선호도 평가)만을 이용한 협력적 추천, 내용기반 추천의 장점과 협력적 추천의 장점을 혼합한 혼합 추천 등의 방법이 있다. 추천시스템은 초기평가의 문제(Early rater problem), 데이터 희소성의 문제(Sparsity problem), 모호 집단의 문제(Gray sheep problem) 등이 있으며(Claypool(1999)) 초기평가의 문제는 처음 사용자가 추천을 받기 위해서는 예측의 기반이 되는 다른 사용자들의 평가 혹은 의견이 있어야 하는데 추천시스템 초기에는 이러한 데이터의 부족으로 인하여 예측이 불가능하거나 정확한 예측을 얻을 수 없는 경우가 발생한다. 희소성의 문제는 아이템에 대한 선호도 평가와 같은 예측의 기반이 되는 사용자의 평가 자료가 희박한 경우에 발생하며 사용자와 아이템에 대한 정보의 밀도가 낮아 추천의 질이 떨어지는 문제점이다. 모호 집단의 문제는 중소단위의 사용자 집단에서 특정 사용자가 속한 그룹의 다른 사용자들과 지속적으로 의견이 일치하거나 일치하지 않기 때문에 발생하는 문제로 이러한 사용자의 경우 추천시스템의 이득을 취하기 어려우며 전체 시스템에도 영향을 줄 수 있다. 추천시스템은 초기의 시스템에서는 추천시스템의 문제점에 영향을 적게 받을 수 있는 내용기반 추천을 이용하고 일정 데이터의 확보가 이루어지면 협력적 추천의 접근법을 사용하는 혼합적 방법을 취하는 시스템들이 많이 있으며 전자상거래에 적용되는 많은 시스템들이 이러한 방법을 취하고 있다(Schafer(1999)).

2. 연구목적

본 연구에서는 GroupLens에서 제시한 특정 사용자의 이웃 기반의 예측 알고리즘(neighborhood based algorithm)(Resnick(1994))을 이용하여 5개의 dataset에 대해 사용자들의 선호도의 상관관계를 나타내는 유사도 가중치인 피어슨 상관계수(Pearson's correlation coefficient)와 벡터 유사도(Vector similarity)를 이용하여 응답 쌍의 개수를 고려한 유의성 가중치(Significance weighting)에 따른 MAE의 변화량을 비교 연구하였다.

3. 기존연구

3.1 협력적 필터링(collaborative filtering)

협력적 필터링은 아이템의 특성이나 사용자의 프로파일과 같은 속성을 의도적으로 무시하고 사용자들이 아이템에 대해 선호도에 대한 평가, 즉 사용자-아이템 간의 관계 데이터만을 이용하여 예측하는 접근법이다(Hill(1995), Resnick(1994), Shardanand과 Maes(1995)). 협력적 필터링은 추천접근법에서 가장 성공적인 접근법으로 가장 간단한 협력적 필터링의 예로는 가장 잘 알려진 아이템을 모든 사용자들에게 추천하는

것이라 할 수 있다.

협력적 필터링은 이미 많은 상업적 전자상거래에서 실용화되고 있으며 추천 알고리즘 연구의 근간을 이루고 있다. GroupLens는 넷 뉴스의 기사의 선호도 예측을 위하여 예측을 하고자 하는 문서에 대해 다른 사람이 평가한 선호도를 이용하는 사용자 이웃 기반의 예측 알고리즘(neighborhood based algorithm)을 이용하여 사용자가 접하지 않은 새로운 문서에 대한 예측을 하였다. 사용자 이웃 기반의 예측 알고리즘은 이미 문서를 읽은 다른 사람의 견해를 이용하여 예측을 하게 되며 이때 문서를 읽은 사람이 없다면 그 문서에 대한 예측을 할 수 없게 된다(Resnick(1994), Shardanand (1995)). 초기 GroupLens의 예측 알고리즘은 아이템에 대한 사용자들의 선호도 평가치를 바탕으로 사용자간의 평가치의 거리를 계산하고 사용자가 아이템을 얼마나 좋아하는지에 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 척도화(일반적으로 5점,7점 척도를 이용)시킨 평가치를 사용한다(Herlocker(1999), Breese(1998)는 사용자들 간의 유사성을 피어슨 상관계수(Peason's correlation coefficient), 벡터 유사도(Vector similarity), 사용자가 평가하지 않은 아이템에 대해 예측치 계산에서 제외시키지 않고 기본선호도를(default voting) 부여하여 유사도를 구하는 방법, 많이 검색된 아이템을 찾는 것 보다 검색되지 않은 아이템을 찾는 방법을 택하는 역사용자빈도(Inverse user frequency)와 같은 유사도 가중치에 대해 연구하였다(Breese(1998), Resnick(1994), Resnick(1997)).

협력적 필터링은 사용자와의 이웃을 형성하기 위해 사용자들의 관계 데이터를 이용하는 사용자 기반(user-based)의 협력적 필터링(Sarwar(1998), Claypool(1999))과 반대로 아이템의 관계 데이터를 이용하는 아이템 기반(item-based)의 협력적 필터링으로 나누어진다(Sarwar(2001), Deshpande(2004)). 많은 전자상거래 사이트에서는 아이템 기반의 알고리즘을 이용하고 있다(Schafer(2001)).

근접 이웃 형성(nearest-neighborhood formation), 분류화 알고리즘(classification algorithm), 연관규칙 마이닝(association rule minning), 베이지안 네트워크(bayesian network), 클러스터 모형(cluster models) 등의 데이터 분석 알고리즘이 협력적 필터링 문제에 적용되어 연구되고 있다.

3.2 사용자 기반의 협력적 필터링(user based collaborative filtering)

GroupLens에서 제시한 사용자 이웃 기반의 예측 알고리즘을 이용한 협력적 필터링은 특정 고객의 상품에 대한 선호도를 예측하기 위하여 대부분의 경우 식(2)의 피어슨 상관 계수를 이용하여 유사한 선호도를 가지는 이웃들을 정하고 식(1)에 의해 예측 선호도 값을 계산한다.

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

여기서,

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

단, U_x는 아이템 x에 대한 특정사용자 u의 선호도 예측치이고, r_{uj}는 특정사용자 u와 이웃한 사용자 j의 상관관계를 나타내는 피어슨 상관계수이다. J_x는 이웃사용자 j의 아이템 x에 선호도이고 $\overline{J}$는 이웃사용자 j의 선호도 평균이다. Raters는 테스트 상품에 대해 선호도를 표시한 고객들을 의미한다. r_{uj}는 사용자 u와 j의 유사도 가중치로 일반적으로 피어슨 상관계수와 벡터 유사도가 널리 사용된다. 본 논문에서는 피어슨 상관계수와 벡터 유사도 두 가지의 유사도 가중치를 이용하여 사용자 u의 선호도를 예측한다. 식 (3)은 사용자 u와 j의 관계를 나타내는 코사인 벡터 유사도이다.

$$r_{uj} = \frac{U \cdot J}{|U| \cdot |J|} \qquad (3)$$

Herlocker et. al.의 연구에서 사용자간의 상관관계를 나타내는 상관계수는 두 사용자가 공통으로 평가한 응답 쌍에 영향을 받고 있음을 알 수 있다. 상관관계가 큰 사용자들 중에는 일반적으로 공통으로 응답한 쌍의 개수가 매우 작은 사용자들이 있다. 피어슨 상관계수의 경우 공통으로 평가한 응답 쌍이 2개일 경우 가질 수 있는 값은 1, 혹은 -1의 값이다. 그러나 2개의 공통 응답 쌍의 개수로 두 사용자간의 상관관계를 설명하기에는 무리가 있으며 또한 과대하게 연관성을 부여받은 가중치로 인해 예측의 오차가 커질 것으로 보고 응답 쌍의 개수에 따라 유사도 가중치인 상관계수를 보정해 주는 유의성 가중치(Significance weight)를 설정하였다. Herlocker et. al.은 응답쌍의 개수에 따른 영향을 50개로 보고 응답 쌍의 개수가 50보다 적은 상관계수에는 n/50의 가중치를 부여하고 응답쌍이 50보다 큰 경우 상관계수에 1의 가중치를 부여하여 정확도가 향상된 결과를 보였다. 이희춘(2006)은 사용자가 응답한 영화의 응답편수와 상관계수에 따라 MAE의 차이가 있을 것으로 보고 변화량에 대해 연구하였다.

본 논문에서는 사용자의 선호도 예측식인 식 (1)에 사용자 간의 상관관계를 나타내는 유사도 가중치인 r_{uj}를 피어슨 상관계수와 벡터 유사도로 구분하고 이때 각 유사도 가중치에 유의성 가중치인 sw를 각각 곱하여 응답 쌍의 개수에 따른 가중치를 각각 부여하여 예측을 하였다.

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%으로 분할한 5개의 dataset을 이용하여 평가하였다.

(1) 응답 쌍(pair of response)의 빈도분포

응답 쌍(pair of response)의 빈도분포표는 다음 <표 1>과 같다.

<표 1> 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 응답 쌍의 빈도분포표에서 알 수 있듯이 응답쌍이 없는 경우가 3.4%이며 응답쌍이 5개 이하의 경우 31.6%를 차지하고 있어 응답 쌍의 개수에 따라 예측치는 영향을 받을 수 있다.

4.2 성능평가

본 논문에서 제시한 알고리즘과 기존의 알고리즘의 정확도를 평가하기 위해 MAE(Mean Absolute Error)를 사용하여 실제 선호도와 예측 선호도간의 정확도를 평가하였다. MAE는 실제 사용자의 선호도 평가치와 예측치의 차이에 대한 절대값의 평균으로 나타내며 알고리즘의 예측에 대한 정확성을 알 수 있으며 식 (4)에 의해 정의된다.

$$MAE = \frac{\sum_{i=0} | \varepsilon_i |}{N} \tag{4}$$

4.3 분석

본 연구에서 GroupLens의 movielens dataset를 분석한 결과 응답 쌍의 개수 분포가 1-5사이의 경우 전체의 28.2%를 차지하고 50까지의 비율이 90.9%를 차지하는 것으로 나타나 공통 응답 쌍의 개수가 적어 상관계수가 과대하게 평가되어 있음을 알 수 있었다. 본 연구에서는 사용자 이웃 기반의 알고리즘에 유사도 가중치를 피어슨 상관계수와 벡터 유사도로 나누어 적용하여 예측을 하고 유사도 가중치에 응답 쌍의 개수를 고려한 유의성 가중치를 세분화하여 이에 따른 MAE의 변화를 살펴보았다.

$$U_x = \overline{U} + \frac{\sum_{J \in Raters} (J_x - \overline{J}) r'_{uj}}{\sum_{J \in Raters} | r'_{uj} |} \tag{5}$$

여기서,

$$r'_{uj} = r_{uj} \cdot sw$$

sw는 응답쌍의 개수에 따라 다음과 같은 가중치를 부여하였다

$$sw = \begin{cases} \frac{n}{C}, & C = 3,5,7,10,15,\cdots,50,60,\cdots,100,120,150,180,200,300 \\ 1, & n \geq C \end{cases}$$

여기서 n은 응답 쌍의 개수이다.

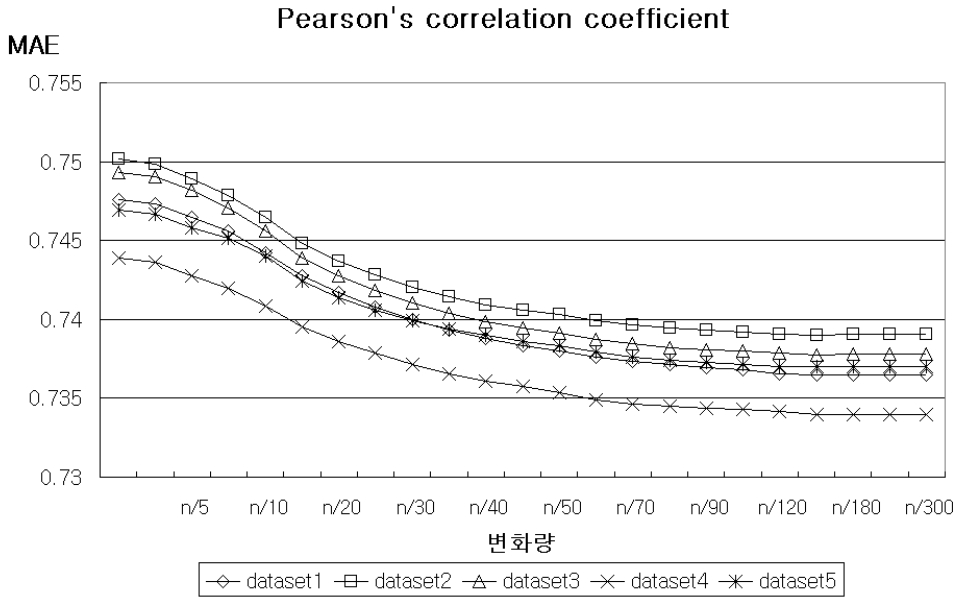
4.4 분석 결과

<표 3> 유의성 가중치에 따른 분석결과

Dataset	Max / Min	상관계수	조건	벡터 유사도	조건
1	max	0.747598	무조건	0.755477	무조건
	min	0.736453	n/150	0.750469	n/180
2	max	0.750199	무조건	0.754684	무조건
	min	0.739025	n/150	0.74932	n/150
3	max	0.749318	무조건	0.757654	무조건
	min	0.737768	n/150	0.751885	n/150
4	max	0.74392	무조건	0.751836	무조건
	min	0.733957	n/200	0.746467	n/200
5	max	0.746925	무조건	0.75585	무조건
	min	0.736987	n/150	0.751195	n/180

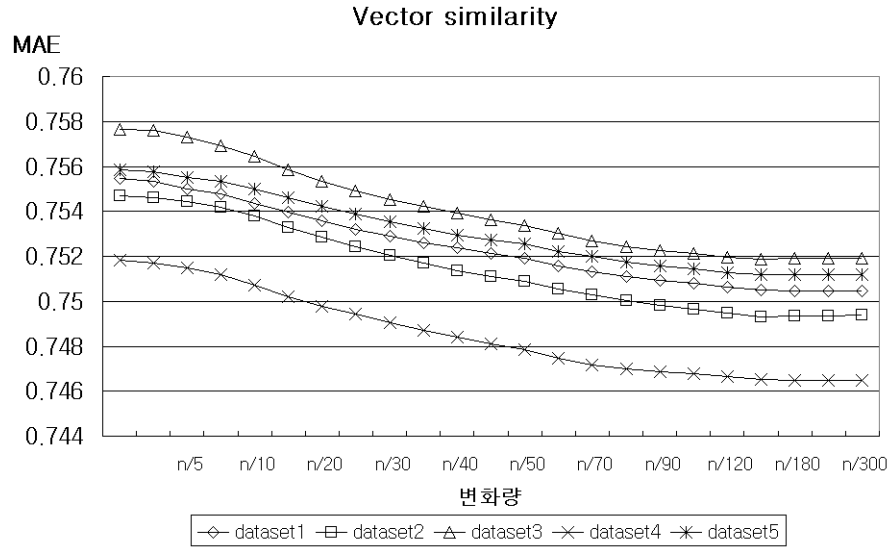
<표 3> 유의성 가중치 sw에 따른 각 dataset별 분석결과이다. dataset1에서 피어슨 상관계수의 유의성 가중치 n/150일 경우 MAE가 가장 작게 나타났으며 유의성 가중치가 무조건일 경우 가장 크게 나타났다. dataset2의 경우 유의성 가중치 n/150일 경우 가장 작았으며 무조건일 경우 가장 크게 나타났다. dataset3에서는 각각 n/150, 무조건, dataset4에서는 각각 n/200, 무조건, dataset5에서는 각각 n/150, 무조건일 경우 가장 큰 결과를 보여 응답 쌍의 개수를 고려한 유의성 가중치의 영향을 받음을 알 수 있었다. 벡터 유사도도 유의성 가중치의 영향을 받아 응답쌍의 개수에 따라 MAE의 변화가 관찰되었다. dataset1에서는 유의성 가중치가 n/180일 경우 MAE가 가장 작게 나타났으며 유의성 가중치가 무조건일 경우 MAE가 가장 크게 나타났다. dataset2에서는 각각 n/150일 경우, 무조건일 경우, dataset3에서는 각각 n/150일 경우, 무조건일 경우, dataset4에서는 n/200일 경우, 무조건일 경우, dataset5에서는 n/180일 경우, 무조건일 경우 MAE가 가장 크게 나타났다. 이것은 응답 쌍의 개수를 고려한 유의성 가중치의 영향을 받음을 알 수 있다.

다음의 <그림 1>은 각 dataset의 유사도 가중치인 피어슨 상관계수에 대해 유의성 가중치의 변화량에 따른 MAE의 결과이다.



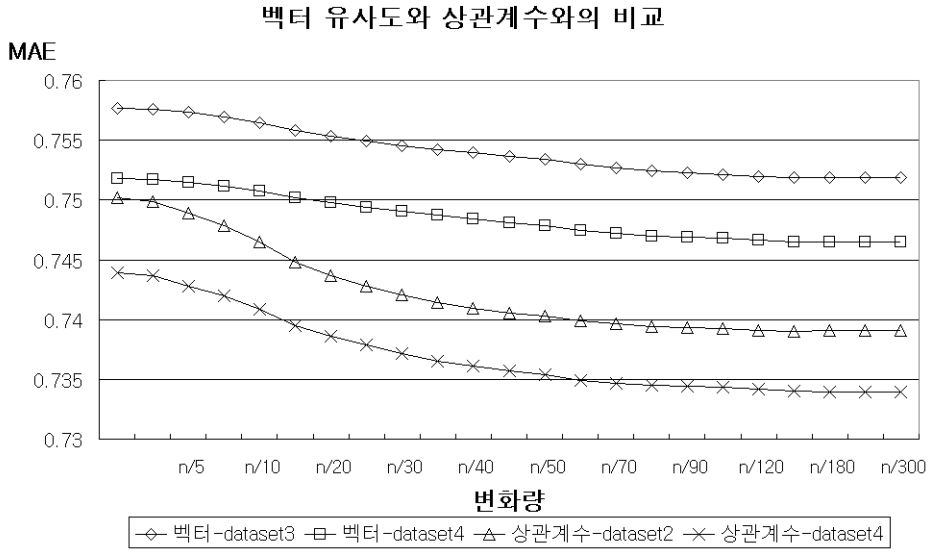
<그림 1> 상관계수 유사도 가중치에 적용된 유의성 가중치 변화에 따른 MAE 변화

다음의 <그림 2>는 각 dataset의 벡터 유사도 가중치에 대해 유의성 가중치의 변화량에 따른 MAE의 결과이다.



<그림 2> 벡터 유사도 가중치에 적용된 유의성 가중치 변화에 따른 MAE 변화

다음의 <그림 3>은 각 dataset 중에서 유사도 가중치인 피어슨 상관계수를 사용했을 때 MAE가 가장 큰 dataset2와 MAE가 가장 작은 dataset4, 벡터 유사도 가중치를 사용했을 때 MAE가 가장 큰 dataset3, MAE가 가장 작은 dataset4에 대한 유의성 가중치의 변화량에 따른 MAE의 결과이다.



<그림 3> 상관계수와 벡터 유사도에 적용된 유의성 가중치 변화에 따른 Max MAE dataset과 Min MAE dataset의 MAE 변화

위의 결과에서 유사도 가중치인 피어슨 상관계수와 벡터 유사도는 응답 쌍의 개수에 따른 유의성 가중치에 영향을 받음을 알 수 있고 피어슨 상관계수의 경우 n/150에서 n/200 사이의 유의성 가중치에서 MAE가 최소가 되고 벡터 유사도의 경우 n/150에서 n/200 사이의 유의성 가중치에서 MAE가 최소가 되었다. 유의성 가중치는 n./300 이상에서는 벡터 유사도와 상관계수 모두 더 이상 영향을 미치지 않는 것으로 실험결과 나타났다. 결국 사용자 이웃 기반의 예측 알고리즘에서 사용되는 유사도 가중치는 사용자들이 공통으로 선호도를 평가한 응답 쌍의 개수에 영향을 받음을 알 수 있어 응답 쌍의 개수를 고려한 가중치를 사용하는 것이 예측치의 MAE가 작다는 것을 알 수 있다.

5. 결 론

본 논문에서는 응답쌍은 예측에 어떤한 영향을 미치는 알아보기위해 응답쌍의 변화에 따른 유사도 가중치를 이용하여 피어슨 상관계수와 벡터 유사도를 변형한 가중치를 예측에 사용한 결과 응답 쌍의 개수에 따른 유의성 가중치에 영향을 받음을 알 수 있다. 응답 쌍의 개수를 고려한 가중치를 사용하는 것이 예측치의 MAE를 줄일 수 있다고 할 수 있다.

참고문헌

1. 이희춘 (2006). An Exploratory Study for Decreasing Error of Prediction Value of Recommender System on User Based, 한국데이터정보과학회, 제17권, 1호, 77-86
2. Badrul M. Sarwar and Joseph A. Konstan and Al Borchers and Jonathan L. Herlocker and Bradley N. Miller and John Riedl. (1998). Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System, Computer Supported Cooperative Work, 345-354.
3. Badrul M. Sarwar and George Karypis and Joseph A. Konstan and John Reidl. (2001). Item-based collaborative filtering recommendation algorithms, In Proceedings of Tenth International World Wide Web Conference, 285-295.
4. 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.
5. 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.

6. J. Ben Schafer and Joseph A. Konstan and John Riedl. (2001). E-Commerce Recommendation Applications, Data Mining and Knowledge Discovery, 5-1, 115-153.
7. 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.
8. M. Claypool and A. Gokhale and T. Miranda and P. Murnikov and D. Netes and M.Sartin. (1999). Combining Content-Based and Collaborative Filters in an Online Newspaper, In Proceedings of ACM SIGIR Workshop on Recommender Systems.
9. Mukund Deshpande, George Karypis. (2004). Item-based top-N recommendation algorithms, ACM Transactions on Information Systems, 22-1, 143-177.
10. Paul Resnick, N. Iacovou, M. Suchak, P. Bergstorm, 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, In Proceedings of the SIGCHI conference on Human factors in computing systems, 194-201.

[2006년 6월 접수, 2006년 8월 채택]