Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 4, pp. 1129~1139

On the Effect of Significance of Correlation Coefficient for Recommender System

Hee Choon Lee[1)]

Abstract

Pearson's correlation coefficient and vector similarity are generally applied to The users' similarity weight of user based recommender system. This study is needed to find that the correlation coefficient of similarity weight is effected by the number of pair response and significance probability. From the classified correlation coefficient by the significance probability test on the correlation coefficient and pair of response, the change of MAE is studied by comparing the predicted precision of the two. The results are experimentally related with the change of MAE from the significant correlation coefficient and the number of pair response.

Keywords : Collaborative filtering, MAE, Pearson's correlation coefficient, Recommender system

1. 서론

온라인 세계는 개인화(personalization)의 르네상스가 진행 중이다. 개인화를 위한 새로운 기술들의 개발은 온라인 서비스가 진정한 일대일(one-to-one) 개인화 서비스를 달성할 수 있도록 도와주고 있다. 고객(사용자)들은 자신이 필요로 하는 수많은 정보를 관리하기 위해 개인화의 필요성을 인식하고 있다. 추천시스템(Recommender Systems)은 전자상거래(e-commerce) 환경에서 제품, 서비스, 정보와 같은 거래 아이템에 대한 무수한 정보 중 사용자가 관심을 가지는 정보를 제공하여 아이템을 제공하는 추천자의 입장에서는 아이템의 과도한 생산을 줄이고 판매를 증진시키는 관점에서 이익을 실현하고 사용자의 입장에서는 아이템에 대한 필요한 정보뿐만 아니라 자신도 인식하지 못했던 정보를 신속하게 취득하는 기회를 제공받게 된다. 전자상거래의 추천시스템은 웹 기반 시스템에 통합되어 사용자의 구매 행위, 아이템에 대한 선호도, 웹 검색 등에서 얻어지는 각종 데이터를 통하여 아이템을 추천하기 위한 핵심적인 정

1) 강원도 원주시 우산동 660번지 상지대학교 컴퓨터데이터정보학과 교수
E-mail : choolee@sangji.ac.kr

보를 수집하게 된다. 사용자와 아이템간의 관계데이터는 사용자의 인구통계학적 자료와 기업이 보유하고 있는 사용자 정보와 아이템 정보를 바탕으로 향상되어진다. 사용자의 선호도 예측 모형과 알고리즘을 이용하여 사용자가 원하는 추천을 생성하기 위한 추천시스템의 추천엔진에 투입되고 분석되어진다. 생성된 추천은 마케팅활동과 구매의사결정을 지원하기 위해 마케팅 전문가와 사용자에게 제공된다.

2. 연구의 필요성

추천시스템은 데이터 분석기법을 적용하여 특정 사용자에게 선호도의 예측치를 생성하여 관심을 가지는 아이템들을 찾는데 도움을 주는 시스템이다. 아이템 추천은 다양한 방법으로 이루어질 수 있다. 추천은 사용자의 인구통계학적 변수, 전체적인 상위 선택 아이템, 혹은 미래 아이템에 대한 예측을 위해 과거 사용자들의 선택 습관 등을 기반으로 이루어진다. 협력적 필터링(collaborative filtering)은 현재 가장 성공적인 추천기법이다(Shardanand and Maes, 1995). 추천시스템의 목표를 요약하면 다음과 같다

가. 추천시스템의 목표
(1) 각 개인들의 취향과 선호도를 바탕으로 연관성이 있고, 정확한 추천을 제공하여야 한다.
(2) 사용자들의 최소한의 관여로 선호도를 결정하여야만 한다.
(3) 사용자들이 즉각적인 행동을 취할 수 있도록 실시간 추천을 할 수 있어야 한다.

추천시스템을 경험한 사용자들은 추천시스템이 추천하는 목록이 자신들의 선호도와 일치하기를 원하고 또한 신속한 추천이 이루어지기를 바란다. 근접 이웃 알고리즘을 이용한 선호도 예측에서 유사도 가중치를 피어슨 상관계수를 이용하였을 경우가 벡터 유사도를 이용하였을 경우보다 우수한 결과를 나타냈기 때문에 피어슨 상관계수의 특정 조건에서 예측력의 변화를 알아보는 것이 필요하며 가장 최소화 되는 조건을 찾아 전체 자료를 이용한 예측 보다 효율적인 예측이 필요하다.

3. 연구목적

본 연구에서는 GroupLens에서 제시한 특정 사용자의 이웃 기반의 예측 알고리즘인 근접 이웃 알고리즘(nearest neighborhood algorithm)(Resnick(1994))을 이용하여 MovieLens dataset에 대해 예측하였다. 사용자들의 선호도의 상관관계를 나타내는 유사도 가중치는 피어슨 상관계수(Pearson's correlation coefficient)를 이용하였으며 예측에 영향을 주는 응답 쌍의 개수를 고려하여 분석하였으며 피어슨 상관계수를 유의수준에 따라 나누어 예측의 정확도를 비교 연구하였다.

4. 선행연구

추천시스템은 정보필터링 기법을 이용하여 사용자가 원하는 아이템을 추천하기 위한 자동화된 기법을 말한다. 추천시스템은 처음으로 Goldberg(1992)에 의해 개념이 도입되어 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).

4.1 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), Shardanand and 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. GroupLens의 Resnick(1994)은 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. Herlocker(1999)은 사용자의 선호도는 사용자간의 성향을 구분하기 위해 척도화한 평가치를 사용하였다. Breese(1998)는 사용자들 간의 유사성을 피어슨 상관계수(Pearson'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.Sarwar(1998), Claypool(1999)는 특정 사용자의 이웃을 형성하기 위해 이웃 사용자들의 관계 데이터를 이용하는 사용자 기반(user-based)의 협력적 필터링에 대해 연구하였고, 반대로 Sarwar(2001), Deshpande(2004)는 아이템의 관계 데이터를 이용하여 아이템 기반(item-based)의 협력적 필터링을 연구하였다. Schafer(2001)는 전자상거래 사이트에서 활용하고 있는 추천시스템에 대해 연구하였으며 아이템 기반의 알고리즘이 성공적으로 이용되고 있음

을 연구하였다. 이희춘(2006a)은 MovieLens dataset을 이용하여 예측의 정확도를 높이기 위해 인구통계변수와 사용자의 응답 쌍의 영향력에 대해 연구하였으며 이희춘(2006b)에서는 개선 알고리즘을 제시하여 예측의 정확도를 높이는 연구를 하였다. 이희춘 등(2006)은 유사도 가중치인 피어슨 상관계수와 벡터 유사도에 응답 쌍의 영향을 세분화시킨 유의성 가중치를 적용하여 예측의 정확도가 향상됨을 연구하였다.

4.2 사용자 기반의 협력적 필터링(user based collaborative filtering)

GroupLens에서 제시한 사용자 이웃 기반의 예측 알고리즘을 이용한 협력적 필터링은 특정 사용자의 아이템에 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.

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

여기서,

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

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 아이템에 대해 선호도를 표시한 사용자들을 의미한다. r_{uj}는 사용자 u와 j의 유사도 가중치로 일반적으로 피어슨 상관계수와 벡터 유사도가 널리 사용된다. 본 논문에서는 벡터 유사도 보다 예측력이 우수한 피어슨 상관계수를 이용하여 예측하였다.

본 연구에서는 통계적으로 유의한 상관계수가 예측의 정확도를 높일 것이라는 가정하에 유의확률에 따른 예측결과의 MAE 변화량을 살펴보고 유의확률에 따른 예측 정확도의 관계를 분석하였다. 또한 응답 쌍의 개수와 유의확률을 고려하여 MAE에 영향이 있는지를 분석하였다.

5. 연구방법

5.1 평가자료의 분석

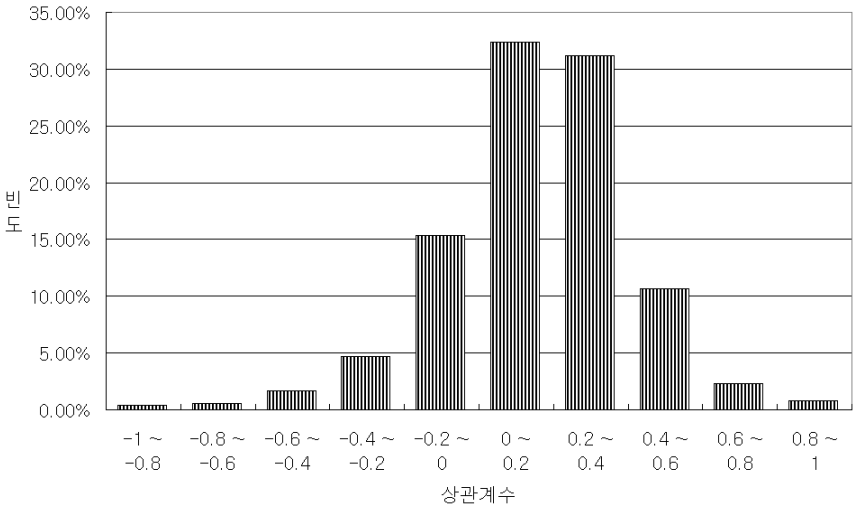
본 논문에서 사용된 dataset은 GroupLens의 MovieLens 100K dataset을 이용하여 실험을 하였다. MovieLens dataset은 943명의 평가자들이 1682편의 영화에 대해 최소 20편에서 최대 737편의 영화를 1-5점으로 평가를 하였다. 1682편의 영화에 943명이 평가한 평가치의 수는 100,000개이다. 본 논문에서는 GroupLens에서 제공되는 MovieLens 100K dataset에 대해 기술통계량은 MovieLens 100K 전체 dataset에 대하여 분석하였으며 상관계수와 응답 쌍의 개수가 MAE의 변화에 어떤 영향을 가지고 있는지의 분석은 MovieLens 100K dataset을 80%:20%의 training dataset과 test dataset으로 나누어 분석하였다. 유의성의 영향에 대한 분석에서 응답쌍이 2개 이하인 경우는 분석에서 제외하였다. 응답쌍이 2개 일 때 피어슨 상관계수를 구하더라도 피어슨 상관계수가 1또는 -1로 나타나 검정식에 의한 유의확률을 구할 수 없다. 이희춘(2006)은 상관계수가 높다는 것이 예측에 좋은 영향만을 미치는 것은 아니라는 결과를 얻었다. 따라서 본 연구에서는 응답쌍을 최소 3개 이상으로 제한 하였다.

다음은 MovieLens 100K dataset의 전체 자료에 대한 상관계수, 유의확률, 응답 쌍에 대한 분석이다. MovieLens 100K dataset 전체 자료의 선호도 예측에 필요한 피어슨 상관계수의 분포는 다음 <표 1>, <그림 1>과 같다.

<표 1> 피어슨 상관계수 빈도 분포표

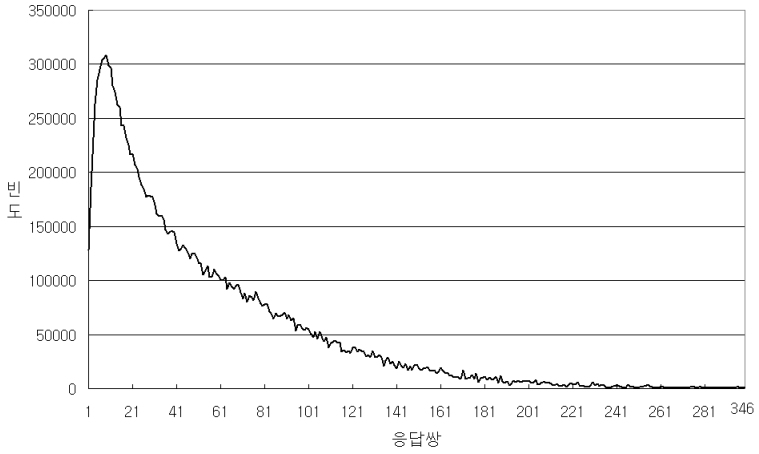
상관계수	빈도	퍼센트	누적 퍼센트
-1 ~ -0.8	59253	0.36%	0.36%
-0.8 ~ -0.6	91718	0.56%	0.92%
-0.6 ~ -0.4	279536	1.70%	2.61%
-0.4 ~ -0.2	768422	4.66%	7.28%
-0.2 ~ 0	2536000	15.39%	22.67%
0 ~ 0.2	5330016	32.35%	55.01%
0.2 ~ 0.4	5136603	31.17%	86.19%
0.4 ~ 0.6	1757065	10.66%	96.85%
0.6 ~ 0.8	386638	2.35%	99.19%
0.8 ~ 1	132659	0.81%	100.00%
합계	16477910	100.00%	100.00%

<표 1>의 상관계수의 분포표를 보면 양의 상관계수 분포의 비율이 음의 상관계수 분포의 비율보다 더 많다는 것으로 알 수 있고 상관계수가 -0.2~0.4 에 78.9%가 분포되어있는 것으로 나타났다.



<그림 1> 상관계수 빈도 분포도

<그림 2>는 MovieLens의 100K 전체 자료에서 사용자 간 상관계수의 분포이다.



<그림 2> 응답 쌍의 분포도

<그림 2>에서 사용자간의 상관계수를 구하기 위한 응답 쌍은 최대 346개의 응답 쌍이 있으며 최소 3개의 응답 쌍이 있는 것으로 나타났다. 사용자 간에 유사도 가중치인 상관계수의 유의성 검정을 위해 최소 3개의 쌍으로 구성되어 있어야 하기 때문에 0~2개의 응답 쌍은 분석에서 제외시켰다.

5.2 성능평가

5.2.1 MAE(Mean Absolute Error)

MAE(Mean Absolute Error)는 협력적 필터링에 의한 예측치의 성능을 평가하기 위해 가장 일반적으로 적용되는 평가 척도이다. MAE는 계산된 선호도 예측치와 이에 대응하는 실제 선호도 평가치의 절대 편차의 평균으로 계산된다.

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

여기서, N은 모든 사용자들에 대한 예측의 총 개수를 나타낸다. MAE에 의한 성능평가의 결과는 MAE가 낮을수록 전체 예측 알고리즘의 정확도가 높다. MAE와 유사한 평가 척도로 MSE(Mean Squared Error), RMSE(Root Mean Squared Error), 그리고 MAE를 표준화 시킨 NMAE(Normalized Mean Absolute Error)등이 있으며 일반적으로 전체 시스템의 정확도는 MAE를 이용하여 성능을 평가한다.

6. 분석 및 결과

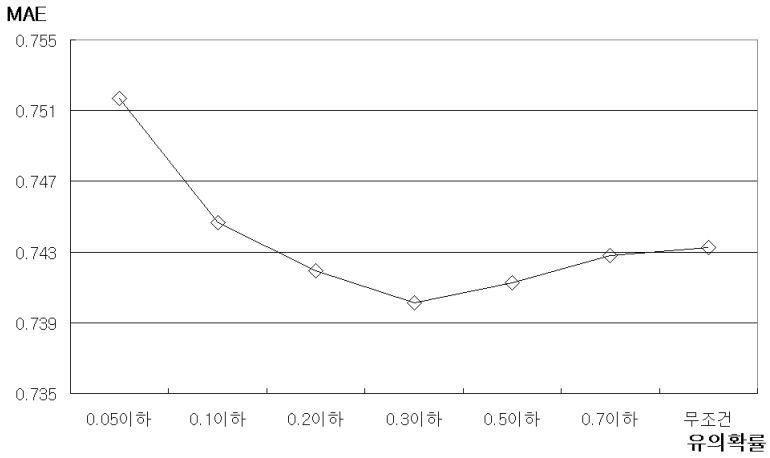
본 연구에서는 상관계수의 크기뿐만 아니라 통계적으로 유의한 상관계수가 예측식의 정확도를 높이는 중요한 요인으로 보고 특정 사용자 와 이웃한 사용자 의 유사도 가중치인 피어슨 상관계수를 유의수준에 따라 통계적으로 유의한 상관계수로 구분하여 예측을 하였다. 상관계수 을 이용하여 두 사용자 와 의 상관계수를 검정하여 유의확률을 각각 {0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 무조건}으로 구분하고 dataset의 예측치에 대해 MAE의 변화량을 살펴보았다. 또한 응답 쌍의 개수를 {5, 7, 10, 20, 30, 무조건}으로 나누고 유의확률을 고려하여 MAE의 변화량을 살펴보았다.

6.1 분석 결과

다음의 <표 2>는 유의확률에 따른 MAE 결과이며 <그림 3>은 유의확률에 따른 MAE의 변화도이다.

<표 2> 유의확률에 따른 MAE 결과표

유의확률 / dataset	0.05이하	0.1이하	0.2이하	0.3이하	0.5이하	0.7이하	무조건
dataset	0.751678	0.744669	0.741942	0.740121	0.741254	0.742783	0.743276



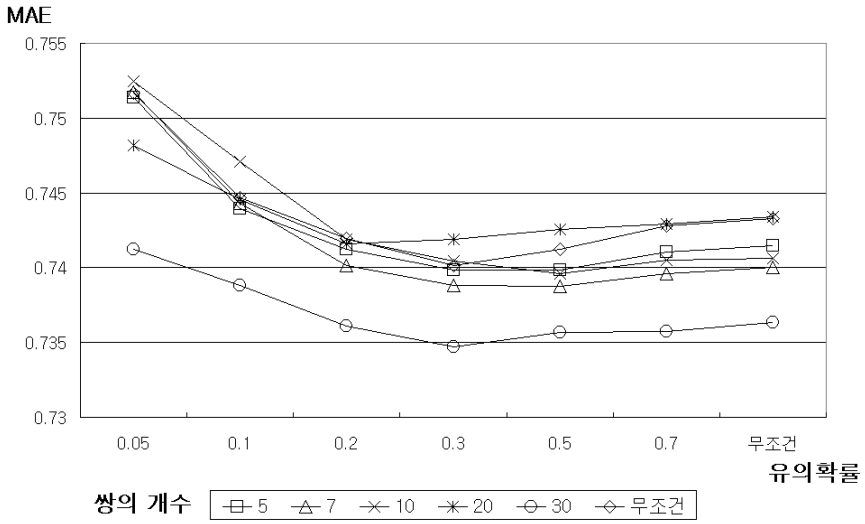
<그림 3> 유의확률에 따른 MAE의 변화도

유의확률이 0.3이하 일 때 MAE가 가장 작게 나타났으며 MAE의 변화는 유의확률 0.3을 기점으로 감소와 증가를 보이는 오목형의 변화량을 보이고 있다.

다음의 <표 3>은 dataset의 사용자간의 응답 쌍과 유의확률을 동시에 고려한 MAE 결과이며 <그림 4>는 dataset의 응답 쌍과 유의확률을 동시에 고려한 MAE의 변화도이다.

<표 3> 응답 쌍과 유의확률에 따른 dataset의 MAE 결과

유의확률 / 응답 쌍	0.05이하	0.1이하	0.2이하	0.3이하	0.5이하	0.7이하	무조건
3이상	0.751678	0.744669	0.741942	0.740121	0.741254	0.742783	0.743276
5이상	0.751398	0.743936	0.741258	0.739857	0.73982	0.74106	0.741486
7이상	0.75171	0.744293	0.740144	0.738818	0.738778	0.739629	0.740048
10이상	0.752457	0.747094	0.741901	0.740437	0.739585	0.740502	0.740604
20이상	0.748175	0.744538	0.741593	0.741867	0.742586	0.74295	0.743402
30이상	0.741207	0.738845	0.736127	0.734726	0.735696	0.735766	0.736329



<그림 4> 응답 쌍과 유의확률에 따른 dataset의 MAE의 변화도

응답 쌍과 유의확률을 동시에 고려한 dataset의 분석에서 응답 쌍의 개수가 30개 이상, 유의확률을 0.3이하로 제한한 상관계수를 이용한 예측치의 MAE가 가장 작음을 보여주고 있다. Herlocker 등(2004)이 연구한 예측 정확도의 연구에서 나타난 'magic barrier'인 0.73에 가까운 MAE를 구할 수 있었다. 이는 예측 알고리즘에 사용되는 상관계수가 응답 쌍과 통계적으로 유의한 상관계수에 제한을 두고 예측하였을 경우 정확도에 영향을 미침을 알 수 있다.

7. 결론

본 논문에서는 사용자 기반의 예측 알고리즘에서 사용자간의 관계를 나타내는 유사도 가중치인 피어슨 상관계수에 대하여 응답 쌍과 상관계수의 유의성에 따른 영향에 대해 연구하였다. 예측에서 유의한 상관계수의 사용이 예측의 정확도에 영향을 미치는 것으로 나타났다. 응답 쌍과 유의적인 상관계수에 따른 예측에서도 MAE가 감소하는 것으로 나타났으며 유의적인 상관계수만을 고려하는 것보다 상관계수의 유의성과 응답 쌍을 고려한 예측이 MAE를 감소시키는 것으로 나타났다. 차후의 연구에서는 응답자(사용자)의 응답의 noise를 고려한 응답자(사용자)의 필터링(filtering) 연구가 필요할 것으로 기대된다.

참고문헌

1. 이희춘(2006). An Exploratory Study for Decreasing Error of Prediction Value of Recommender System on User Based, *Journal of the Korean Data & Information Science Society,* Vol. 17, No. 1, 77-86.
2. 이희춘 (2006b). Improved Algorithm for User Based Recommender System, *Journal of the Korean Data & Information Science Society*, Vol 17. No. 3, 717-726.
3. 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, 775-784.
4. 정경용, 이정현(2005). 개인화 추천 시스템의 예측 정확도 향상을 위한 사용자 유사도 가중치에 대한 비교 평가, *대한전자공학회, 전자공학회논문지*, 제42권 6호, 63-74.
5. Breese, J. S., Heckerman, D., Kadie. C.(1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering, *In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, 43-52.
6. Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M.(1999). Combining Content-Based and Collaborative Filters in an Online Newspaper, *In Proceedings of ACM SIGIR Workshop on Recommender Systems.*
7. Deshpande, M., Karypis, G.(2004). Item-based top-N recommendation algorithms, *ACM Transactions on Information Systems*, 22-1, 143-177.
8. Goldberg, D., Nichols, D., Oki, B. M., Terry, D.(1992). Using collaborative filtering to weave an information tapestry, *Communications of the ACM* , Volume 35, Issue 12, 61-70
9. Herlocker, J. L., Konstan, J. A., Terveen, L. G., Riedl, J.(2004) Evaluating collaborative filtering recommender systems, *ACM Transactions on Information Systems*, Volume 22, Issue 1, 5-53.
10. Hill, W., Stead, L., Rosenstein, M., Furnas, G.(1995). Recommending and Evaluating Choices in A Virtual Community of use, *Proceedings of the SIGCHI conference on Human factors in computing systems*.
11. Konstan, H. J., Borchers, J. A., Riedl, J.(1999). An Algorithmic Framework for Performing Collaborative Filtering, *In Proceedings of the 1999 Conference on Research and Development in Information Retrieval.*
12. Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P., Riedl, J.(1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews, *In Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, 175-186.
13. Sarwar, B. M., Konstan, J. A., Borchers, A., Herlocker, J. L., Miller, B.

N., Riedl, J.(1998). Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System, *Computer Supported Cooperative Work*, 345-354.
14. Sarwar, B. M., Karypis, G., Konstan, J. A., Reidl, J.(2001). Item-based collaborative filtering recommendation algorithms, *In Proceedings of Tenth International World Wide Web Conference*, 285-295.
15. Schafer, J. B., Konstan J. A., Riedle, J.(1999). Recommender systems in e-commerce, *ACM Conference on Electronic Commerce*, 158-166.
16. Schafer, J. B., Konstan, J. A., Riedl, J.(2001). E-Commerce Recommendation Applications, *Data Mining and Knowledge Discovery*, 5-1, 115-153.
17. Shardanand, U., Maes, P.(1995). Social Information Filtering: Algorithms for Automating "Word of Mouth", *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, volume 1, 210-217.
18. 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 9월 접수, 2006년 11월 채택]