Journal of The Korean Association
of Information Education Vol. 17, No. 4,
December 2013, pp. 375-382

http://dx.doi.org/10.14352/jkaie.2013.17.4.375

메모리 기반 협력필터링을 위한 평가 등급 범위를 이용한 유사도 척도

이수정

경인교육대학교 컴퓨터교육과

요 약

협력 필터링은 사용자가 선호했던 항목들의 기록을 토대로 항목을 추천하는 방법으로서 상업 사이트에서 매우 널리 사용되어 왔다. 이 방식의 기본 개념은 유사한 사용자들을 찾아서 그들의 평가등급을 통합하여 새로운 항목 추천에 이용하는 것이다. 따라서 유사도의 정확한 측정은 추천 성능에 매우 중요한 일이다. 본 논문에서는 사용자가 과거에 부여했던 평가등급들을 기준으로 하여 상대적으로 각 평가치를 다루는 새로운 유사도 공식을 제안한다. 광범위한 실험을 통해 제안된 공식이 기존 공식들보다 더 신뢰할 수 있음을 밝혔는데, 이는 극단적인 유사도값의 발생이 현저히 감소하였고, 유사도가 큰 이웃들만을 참조하였을 때 성능이 개선되었기 때문이다. 특히 실험 결과, 제안 공식은 평가 범위가 큰 데이터셋에 대해 기존 공식들보다 우수한 성능을 나타냈다.

키워드 : 추천 시스템, 웹 개인화, 협력 필터링, 유사도

A Similarity Measure Using Rating Ranges for Memory-based Collaborative Filtering

Soojung Lee

Dept. of Computer Education, Gyeongin National University of Education

ABSTRACT

Collaborative filtering has been most widely used in commercial sites to recommend items based on the history of user preferences for items. The basic idea behind this method is to find similar users whose ratings for items are incorporated to make recommendations for new items. Hence, similarity calculation is most critical in recommendation performance. This paper presents a new similarity measure that takes each rating of a user relatively to his own ratings. Extensive experiments revealed that the proposed measure is more reliable than the classic measures in that it significantly decreases generation of extreme similarity values and its performance improves when consulting neighbors with high similarites only. In particular, the results show that the proposed measure is superior to the classic ones for datasets with large rating scales.

Key words : Recommender System, Web Personalization, Collaborative Filtering, Similarity

---

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2012R1A1A3012320)
논문투고 : 2013-07-24
논문심사 : 2013-07-25
심사완료 : 2013-12-04

1. 서론

인터넷 상거래가 활발해짐에 따라 추천시스템(recommender system)은 온라인 고객에게 개인화된 정보를 근거로 유용한 도움을 주는데 큰 역할을 해왔다. 추천시스템은 사용자의 흥미에 부합하는 항목을 추천하거나, 흥미 있는 항목에 대한 평가등급을 예측한다. 협력필터링(Collaborative Filtering, CF)은 가장 널리 사용되는 추천시스템 방식으로서, 실험적 메일 시스템인 Tapestry system[6], Usenet 뉴스 기사를 추천하는 GroupLens[8], 음악을 추천하는 Ringo[13], 서적과 CD를 판매하는 Amazon.com 등이 대표적이다.

본 논문의 초점인 메모리 기반의 CF는 유사한 사용자들이 부여한 과거 평가등급들을 참조하여 현재 사용자를 위해 새로운 항목들의 평가등급을 추정하는 것이다. 따라서 높은 평가 예측치를 가진 항목들을 사용자에게 추천한다. 기본적인 CF 방법의 절차는 다음과 같다.

1. 현 사용자와 다른 사용자들 각각에 대해 유사도를 측정한다. 이 때 두 사용자가 공통으로 평가했던 항목들의 평가등급들을 이용한다.
2. 현 사용자와 가장 유사한 사용자들 (이웃)을 선정한다.
3. 사용자 u와 인접한 사용자들의 집합을 Nu라고 할 때, 사용자 u가 등급을 부여하지 않은 항목 x에 대한 예측 등급 $\hat{r}_{u,x}$는 대개 다음과 같이 산출한다[14][7]. 아래 식에서 $r_{u,x}$는 사용자 u가 부여한 항목 x에 대한 등급이고, $\bar{r}_u$와 $\bar{r}_v$는 각각 사용자 u와 v가 부여한 평균 등급이다.

$$\hat{r}_{u,x} = \bar{r}_u + \frac{\sum_{v \in N_u} sim(u,v) \times (r_{v,x} - \bar{r}_v)}{\sum_{v \in N_u} |sim(u,v)|}$$

위 절차에서 보듯이, CF 방법의 성능은 유사도 척도에 의해 크게 좌우된다. 현재까지는 cosine 기반 또는 상관도 기반의 유사도 측정 방법이 주로 사용되어 왔다. 그러나, H. Ahn의 연구 결과[1]에 따르면, 평가한 항목들이 희소하거나 cold-starting의 환경에서 전통적 유사도 척도를 이용한 CF 방법들의 성능은 매우 저하되어 추천의 질을 심각하게 떨어뜨린다고 하였다. 이같은 문제점을 해결하기 위한 연구 노력이 있었으나, 대개 전통적 유사도 척도에 다른 여러 요소들을 추가적으로 병합시켜 개선하고자 하였기 때문에, 성능 개선에 있어서 그 한계를 뛰어넘기 어려우며, 따라서 성능의 기본 역할을 하는 유사도 측정 방법의 중요성이 대두되었다.

본 논문에서는 메모리 기반 협력 필터링을 위한 새로운 개념의 유사도 척도를 제시한다. 제안된 척도는 각 사용자가 부여했던 평가등급 범위 내의 상대적인 평가등급을 이용하여 유사도를 계산한다. 광범위한 실험을 통하여 제안한 유사도 공식과 기존 공식들을 이용한 CF 방법의 성능을 비교 분석한 결과, 제안한 방법은 특히 평가등급의 범위가 큰 데이터셋에서 우수한 성능을 보임을 확인하였다.

2. 배경

2.1 협력 필터링 관련 연구

Bobadilla 외 3인[3]은 각 항목과 사용자의 중요도를 고려한 새로운 방식의 유사도 척도를 제안하였다. 각 평가치는 중요도를 반영하여 변환되었으며, 변환된 평가등급에 Pearson, cosine, MSD (Mean Squared Differences)를 적용하여 유사도 계산을 실행하였다. Anand과 Bharadwa[2]은 각 평가등급을 변환하는 또 다른 방법을 제시하였는데, 유전자 프로그래밍(genetic programming)을 이용하여 변환 함수를 구하였으며, cosine 유사도를 적용하여 인접 이웃(nearest neighbor)들을 선정한 후 Resnick's formula[11]를 이용하여 미평가된 항목의 등급을 예측하였다.

Gao외 2인[5]은 기존의 항목 기반의 CF 방식의 문제점을 언급하였는데, 모든 사용자가 같은 중요도를 갖고 있다는 것이다. 따라서 사용자의 가중치를 다르게 계산하여 사용자들 간의 유사도를 Adjusted Cosine 공식을 이용하여 산출하였다. Ren 외 2인[10]은 기존의 항목 기반의 CF 방법을 개선하기 위하여 휴리스틱 중복 요소로써 평가항목들의 중복 정도를 파악하였다. 이를 이용하여 기존의 Pearson과 cosine

<표 1> 두 데이터셋에 대한 피어슨 상관도와 코사인 유사도값의 분포

데이타셋	피어슨 상관도 (%)						코사인 유사도 (%)	
	0	1	-1	0-divide	(0,1)	(-1,0)	1	(0,1)
ML	3.15	3.32	2.51	18.43	48.72	23.71	11.94	88.06
BX	0.47	5.25	3.99	83.83	3.68	2.73	74.29	25.71

척도에 휴리스틱 중복 요소를 반영하는 새로운 유사도 공식을 제안하였다.

한편 기존 유사도 공식에 추가 정보를 접목하여 성능을 개선하려는 시도가 있었는데, Bobadilla 외 2인[4]의 연구에서는 Jaccard metric [9]으로 표현된 공통 평가항목개수 정보를 MSD와 접목하여 실험 대상의 척도들 중에서 가장 뛰어난 성능 결과를 보였음을 밝혔다. 결론적으로, 기존 연구에서는 전통적 유사도 척도를 개선하고자 공통평가항목 개수 또는 사용자의 중요도를 통합하여 새로운 척도를 제안하였다. 따라서, 그러한 척도들의 기본이 되는 전통적인 유사도 척도 자체의 성능이 성공적인 CF 방법에 매우 중요함을 알 수 있으므로, 이에 대한 보다 상세한 성능 분석과 개선의 노력이 필요하다.

2.2 기존 유사도 공식의 문제점

유사도 측정은 최인접 이웃을 알아내는 중요한 역할을 하므로 협력 필터링의 성능을 좌우한다. 현재까지 가장 널리 사용되는 측정방법으로 피어슨(Pearson) 상관계수와 코사인(cosine) 척도가 있다[7]. 피어슨 계수는 두 사용자의 공통항목 평가치들을 벡터로 간주하여 선형적 의존관계를 알아내는 것이며, -1부터 +1 사이의 값을 갖는다. 코사인 척도는 두 벡터 간 각도의 코사인 값을 측정한다.

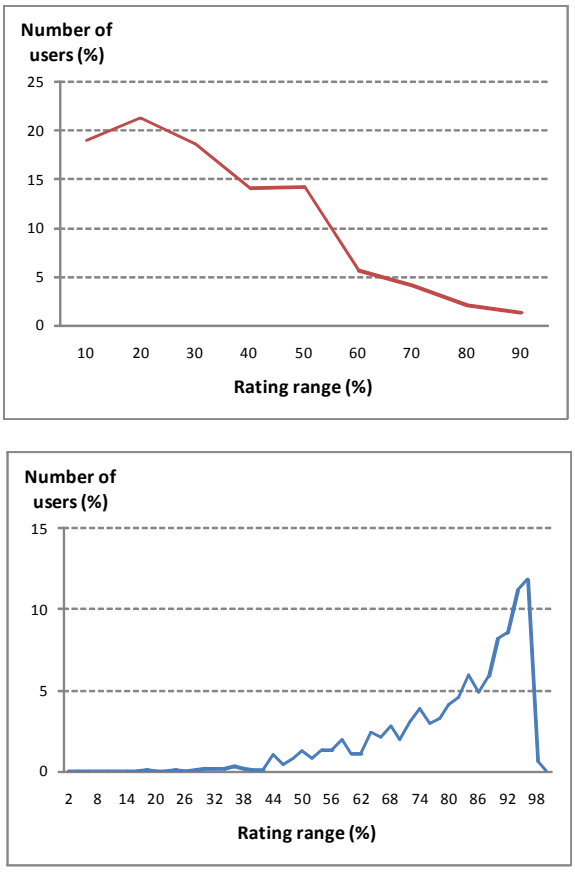
두 대표적 유사도 척도의 문제점은 첫째, 두 사용자의 공통평가항목개수가 단 하나일 때 극단적인 값, 즉, 1, -1, 또는 0-divide의 값을 산출한다는 것이다. 또한, 평가치가 모두 동일할 경우에도 같은 결과가 발생한다. 둘째, 피어슨 상관도는 평가치의 분산을 고려하지 않으므로, 만약 두 개의 공통평가항목인 i와 j에 대해, $r_{u,i}$=5, $r_{u,j}$=4이고, $r_{v,i}$=5, $r_{v,j}$=1이라면, 두 사용자 u와 v간의 피어슨 상관도는 1이 된다. 그러나, 허락된 평가범위가 1~5라고 가정할 때, 이 경우에 두 사용자의 평가의도는 매우 다르다고 할 수 있으므로 높은 유사도가 발생하는 것은 문제가 있다.

피어슨 상관도와 코사인 유사도에 대한 좀 더 깊이 있는 분석을 위해 기존 연구에서 널리 사용되고 있는 MovieLens(ML)와 Book-Crossing(BX)의 두 데이터셋에 대해 공통평가항목이 존재하는 모든 가능한 두 사용자 간의 유사도 값 분포를 알아보았다. <표 1>에서 보듯이, 피어슨 상관의 극단적 값들은 두 데이터셋 모두에서 0-divide가 가장 많이 차지하였는데, ML 데이터셋에선 약 18%, BX 데이터셋에선 약 84%나 되었다. 이와 반대로 비극단적 범위, 즉, (0,1) 또는 (-1,0), 내의 유사도 값은 매우 작은 비중을 차지함을 알 수 있다.

한편 Ahn [1]은 기존 유사도 측정 방식의 문제점을 극복하고자 새로운 개념의 방법을 제안했는데, 각 공통평가항목에 대한 두 평가치의 크기 차이를 고려한 것이다. 즉, 두 평가치가 평균 또는 중앙값으로부터 떨어진 정도와 방향을 측정하였다. PIP라고 불리는 이 방법의 문제점은 첫째, 해당 공식의 구성요소들이 상호 독립적이지 않고 특정 개념을 중복적으로 측정한다는 것이다. 두 번째 문제점은 유사도 계산에 있어서 각 공통평가항목들이 서로 독립적으로 취급된다는 것이다. 예로써, (1) $r_{u,i}$=5, $r_{u,j}$=5, $r_{v,i}$=1, $r_{v,j}$=3이고 (2) $r_{u,i}$=1, $r_{u,j}$=5, $r_{v,i}$=5, $r_{v,j}$=3일 때, 두 경우에 PIP 유사도 값은 동일하다. 그러나, 두 사용자의 각 경우에 있어서의 평가 의도는 동일하다고 보기 어려우므로, 유사도 값은 적어도 같지 않아야 한다. 예를 들어, 코사인 유사도 값은 (1)의 경우 0.8944, (2)의 경우 0.6727이다. 이러한 문제점의 원인은 PIP 유사도가 사용자의 전반적인 평가 패턴을 고려하지 않고 각 공통항목에 대한 평가치를 별개로 취급하기 때문이다.

평가등급의 허락된 범위가 클 경우, 각 사용자의 평가치 범주가 전체 범위의 일부분만으로 구성될 수

있고, 또한 각 사용자의 평가 범주가 다를 수 있다. 이를 확인하기 위해, BX와 Jester 데이터셋에 대해 실험하였다. 전자는 1~10의 정수값, 후자는 -10~+10의 실수값을 평가 범주로 정의한다. (그림 1)에서 BX 데이터셋에 대해서 87% 가량의 사용자들이 전체 평가범위의 60% 이하만을 사용하였고, Jester 데이터셋에 대해서는 약 40%의 사용자들이 전체 범위의 80% 이하를 사용하였다. 이는 물론 각 사용자의 평가개수가 다르기 때문이지만, 사용자들 평가치의 각기 다른 범주는 유사도 계산에 있어서 고려할 필요가 있다.



(그림 1) Book-Crossing(위)과 Jester(아래) 데이터셋에서 사용된 평가치의 각 범위 비율에 해당하는 사용자수의 상대적 비율

3. 제안 유사도 공식

제안 유사도 공식은 사용자가 새로운 항목을 평가할 때 자신이 과거에 부여한 평가등급들을 염두에 둔다는 매우 간단한 아이디어로부터 출발한다. 즉, 시스템에서 허락된 평가범주와 자신의 과거 평가범주가 혼합된 형태가 사용자가 인식하는 새로운 범주가 된다. 물론, 이는 사용자가 충분히 많은 평가를 행하여 자신의 평가패턴이 드러난 경우를 가정한다. 이러한 아이디어를 구현하기 위하여, 우선 아래 변수를 정의하여 사용자의 각 평가치를 자신의 과거 평가 범위에 대한 상대적 값으로 변환하기로 한다.

$$pos_u(r_{u,i}) = \begin{cases} \dfrac{r_{u,i} - r_{u,\min}}{r_{u,\max} - r_{u,\min}}, & \text{if } r_{u,\max} > r_{u,\min} \\ 1, & otherwise \end{cases}$$

위 식에서 i는 항목을, $r_{u,i}$는 사용자 u가 i 항목에 부여한 평가치, $r_{u,\max}$는 사용자 u의 과거 평가범위 중 최대치, $r_{u,\min}$는 사용자 u의 과거 평가범위 중 최소치를 나타낸다.

물론, 각 공통평가항목에 대해 동일한 pos 값을 갖는 두 사용자간의 유사도는 매우 높아야할 것이다. 따라서, 제안된 유사도 공식은 공통항목의 pos 값의 차이를 반영하여 정의하기로 한다. 단, 차이의 반영 정도를 극대화 (적은 차이일 때는 높은 유사도를, 그렇지 않은 경우엔 낮은 유사도 값을 발생)하기 위하여, pos 차이에 대한 지수함수를 선택하였다. I를 사용자 u와 v의 공통평가항목들의 집합이라고 할 때, 두 사용자간 유사도는 다음과 같이 정의한다.

$$sim(u,v) = \frac{2}{1 + e^{\alpha \cdot D_{pos}(u,v)}}, \quad \alpha > 1$$

$$D_{pos}(u,v) = \sum_{i \in I} \frac{1}{|I|} |pos_u(r_{u,i}) - pos_v(r_{v,i})|$$

위와 같이 정의한 유사도 측정 방법에 따라 유사도값의 분포를 ML과 BX 데이터셋에 대해 실험하여, <표 1>의 결과와 비교하였다. ML 데이터셋에 대해

서, (0, 1) 구간 내의 유사도값은 제안 방식에 따르면 97.3%의 분포를 나타냈으며, 이는 <표 1>의 코사인과 피어슨 방식에 비하여 크게 월등한 수치이다. 또한 BX 데이터셋에 대해서는 그 차이가 더욱 컸는데, 피어슨 유사도가 3.68%, 코사인 유사도가 25.71%이었는데 비해, 제안 공식은 89.02%의 분포를 보였다.

위 유사도 공식으로 가장 인접한 이웃들이 선정된 후에, 미평가된 항목 x에 대한 예측치는 평가범위 $R_u = r_{u,\max} - r_{u,\min}$를 반영하여 다음과 같이 산출한다.

$$\widehat{r_{u,x}} = \min_u + \frac{\sum_{v \in N_u} sim(u,v)(r_{v,x} - \min_v)R_u / R_v}{\sum_{v \in N_u} |sim(u,v)|}$$

4. 실험

4.1 실험 배경

제안한 예측 방법의 성능을 평가하기 위하여, Book-Crossing(BX)과 Jester 데이터셋을 사용하였다. 이들 데이터셋은 기존 관련 연구에서 성능 평가를 위해 매우 널리 사용되고 있다. 두 데이터셋에서 정의한 평가범위가 다르므로, 용이한 성능 비교를 위하여 각 실험결과는 같은 범위로 정규화하였다. 각 데이터셋에 대해 <표 2>에 상세 기술하였는데, 희소성 수준이란 행렬 내 데이터가 없는 요소, 즉, 평가가 매겨지지 않은 요소의 비율을 의미하며, (값이 0인 요소 개수)/(행렬의 크기)로 산출한다.

시스템에서 예측한 평가등급의 정확도는 여러 척도로서 측정할 수 있는데, 본 논문에서는 관련 연구에서 주로 사용하는 MAE(Mean Absolute Error)[12]를 도입하였고, N개의 실제 평가치 r_i에 대한 예측치 $\hat{r_i}$에 대해, $MAE = \frac{1}{N}\sum_i |r_i - \hat{r_i}|$로 정의한다. 이 밖에 성능 평가 척도로서, 추천항목들의 질을 precision과 recall 척도[5]로써 평가할 수 있는데, 이들의 조화평균값인 F1을 사용하기로 한다.

비교 대상의 유사도는 기존의 CF 시스템에서 주로 사용되었던 피어슨 상관도(PRS), 코사인 유사도(COS), 스피어만 순위상관계수(SPR), PIP 유사도(PIP)를 선정하였고, 본 연구에서 제안한 유사도는 RANGEB로 표기하였다. 성능에 영향을 미치는 다양한 요소들과 그 영향 정도를 분석하기 위해, 최인접 사용자 수(Number of NNs)와 제안한 유사도 공식에 사용된 α값을 변화시켜 실험하였다.

실험 결과의 신뢰도를 높이기 위하여 각 실험 결과는 5회 크로스 확인(5-fold cross validation)의 평균값으로 산출하였는데, 각 회마다 서로 다른 훈련 데이타 집합과 시험 데이타 집합을 8:2으로 구성하였다. 모든 실험은 1.96GM RAM과 3.16GHz Intel Core 2 Duo CPU의 PC 상에서 C 프로그램을 작성하여 진행하였다.

4.2 실험 결과

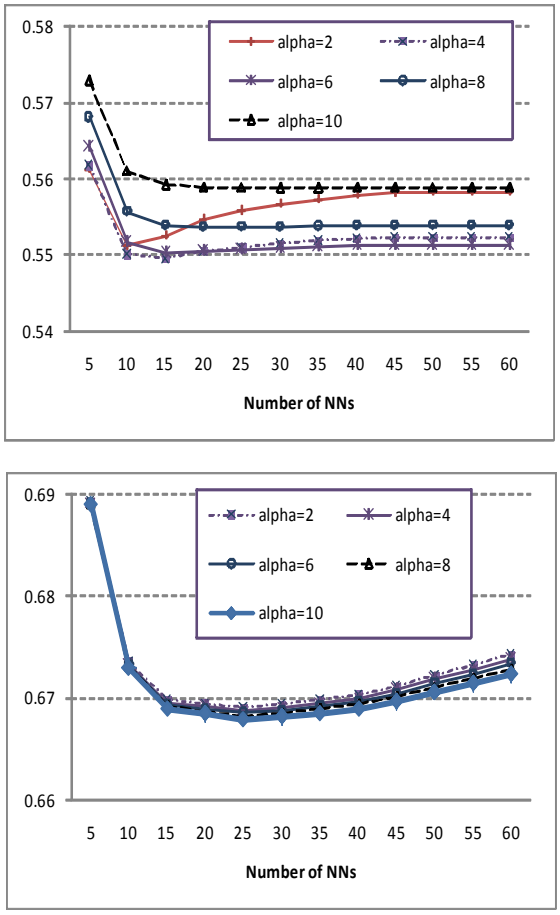
4.2.1 α값의 영향 분석

(그림 2)는 본 연구에서 제안한 유사도 공식의 α값의 영향을 최인접 이웃수를 변화시켜 실험한 결과이다. BX와 Jester 데이터셋에 대한 MAE 측정 결과에서 후자의 경우 α값이 큰 영향을 주지 않은 반면에, 전자의 경우엔 α값의 변화에 따라 MAE 결과에 다소 차이를 보였는데, 그럼에도 불구하고, 그 차이는 최대 0.01를 넘지 못함을 볼 수 있다. 이러한 결과에 대한

<표 2> 실험 데이터 집합

	BX	Jester
평가개수	각 사용자 당 10개, 각 서적 당 20개 초과	각 사용자 당 24개 초과
행렬크기(사용자수×항목수)	1014 × 883	1498 × 480
평가범위	1~10의 정수	-10~+10의 실수
희소성 수준	0.9775	0.2936

이유는 <표 2>에 제시하였듯이, BX 데이터셋의 경우에 Jester보다 희소성 수준이 매우 높으므로, 작은 유사도 값 차이가 상대적으로 큰 성능 차이를 보인 것으로 판단된다. 따라서, BX에 대한 실험 결과를 바탕으로 하여, 가장 좋은 성능을 보인 α값인 6을 선택하여 실험을 계속 진행하였다.



(그림 2) Book-Crossing(위)과 Jester(아래) 데이터셋에 대한 α값 변화에 따른 MAE 성능

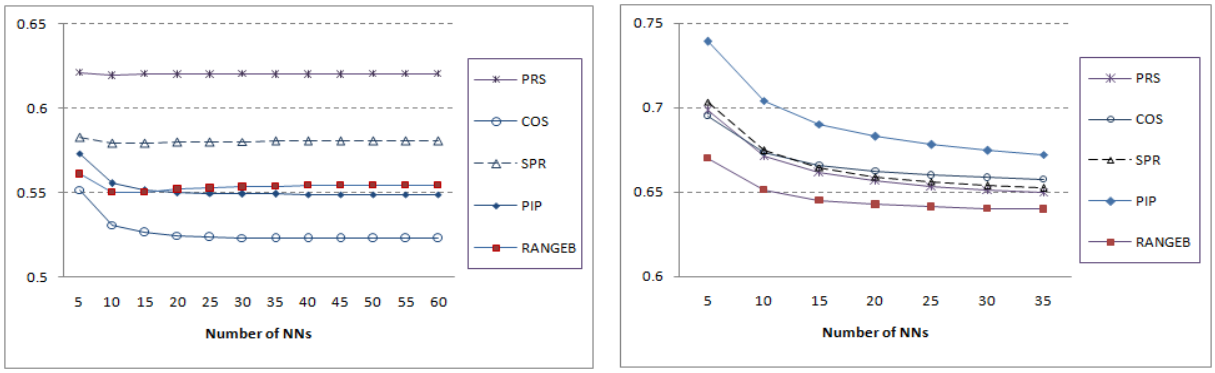
4.2.2 최인접 이웃수에 따른 성능 분석

최인접 사용자수(NN)를 5부터 35까지 변화시켜 성능에 대한 영향 정도를 알아보았다. (그림 3)에 MAE 결과를 제시하였는데, 각 유사도 척도를 이용한 CF 성능은 NN이 커짐에 따라 점차적으로 안정화되는 것을 확인할 수 있다. 이는 NN을 30명 이상으로 증가시키는 것은 평가치 예측의 정확도를 개선하는데 있어서 큰 도움을 주지 않음을 말한다.

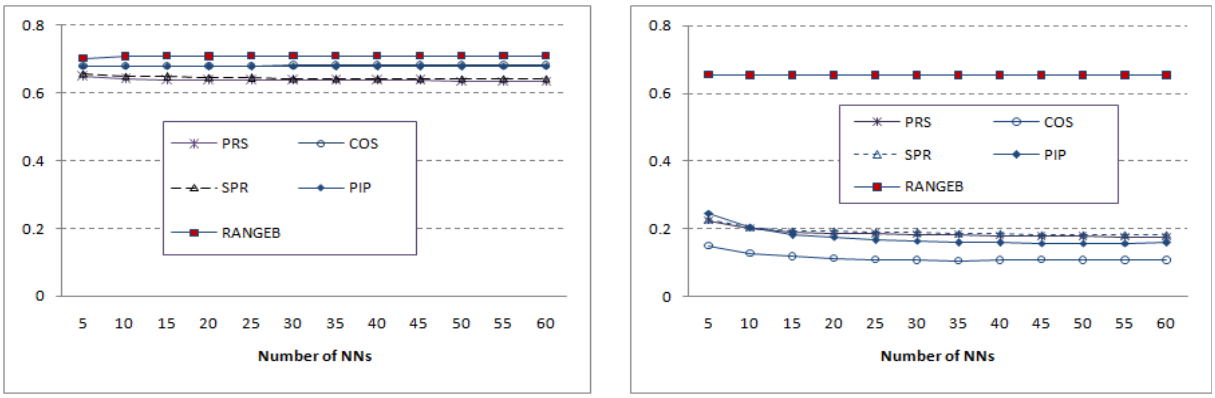
두 데이터셋 상의 실험에 있어서 모두 가장 우세한 성능의 유사도 척도는 없음을 확인할 수 있는데, 특히, BX 데이터셋에 있어서는 Jester 데이터셋보다 유사도 척도 간에 상대적으로 더 큰 차이를 보였다. 이는 BX 데이터셋에서 희소성 수준이 훨씬 크기 때문에 유사도가 상대적으로 낮은 이웃이 포함될 수 있고, 그러한 이웃들의 평가치가 일부 유사도 척도의 성능 저하에 큰 영향을 주는 것으로 판단된다.

특히, PRS는 <표 1>에 제시하였듯이, BX 데이터셋에서 매우 큰 비율의 극단적 유사도 값을 발생시키는데, 이러한 원인으로 (그림 3)에서 BX 데이터셋에 대한 MAE 성능이 현저히 떨어지는 것으로 파악된다. 또한 높은 희소성 수준은 SPR의 성능에도 악영향을 주는 것으로 나타났는데, 그 이유는 공통평가항목이 단 하나일 경우 SPR값은 1이 되고, 둘일 경우엔 ±1이 되어 부정확한 값이 산출되기 때문인 것으로 판단된다. 또한, [1]에서 언급한 바와 같이, PIP는 주로 cold-starting 조건을 대비하여 만들어진 유사도이기 때문에, BX보다 상대적으로 밀집된 Jester 데이터셋에서 그 성능은 다른 유사도들보다 더욱 저하됨을 확인할 수 있다. 다만, COS은 기존 유사도들에 비해 가장 우수한 성능 결과를 보였고, 특히 BX 데이터셋에서 RANGEB보다 우수하였다. 반면에 제안한 유사도인 RANGEB는 BX 데이터셋에선 COS을 제외한 모든 유사도 척도를 능가하였으며, 그보다 더 희소성 수준이 낮고 큰 범위의 평가등급을 정의한 Jester 데이터셋 하에서는 월등한 성능 차이를 보이므로, 전반적으로 그 우수성이 입증되었다. 이는 평가범위를 크게 정의한 데이터셋을 타깃으로 한 제안 방법의 의도가 그 효과를 발휘하였음을 입증했다고 볼 수 있다.

(그림 4)는 F1을 통한 새로운 항목 추천의 성능 결과를 나타낸다. 대부분의 척도들이 매우 대등한 성능을 보이나, RANGEB는 특히 Jester 데이터셋에서 매우 확연히 우수한 성능을 보였다. F1을 구성하는 두 가지 요소인 precision과 recall 결과값을 살펴본 바에 따르면, precision 값은 두 데이터셋 모두에서 실험한



(그림 3) Book-Crossing(좌)과 Jester(우) 데이터셋에 대한 최인접 이웃수 변화에 따른 MAE 성능



(그림 4) Book-Crossing(좌)과 Jester(우) 데이터셋에 대한 최인접 이웃수 변화에 따른 F1 성능

유사도 척도 간에 거의 유사한 결과를 보였으며, 반면에 recall 값에 대해서는 RANGEB가 Jester 데이터셋에서 상대적으로 월등한 성능을 보여 종합적으로 RANGEB의 F1 성능이 매우 우세한 결과를 보였다. 이는 RANGEB를 적용한 추천시스템이 높은 평가치를 부여 받은 항목들을 보다 빠짐없이 추천함을 말하는 것이므로, 상대적으로 신뢰성이 높다고 할 수 있다.

5. 결론

협력 필터링을 통한 추천 시스템은 광범위한 자료들 중에서 사용자에게 필요할 만한 자료들만을 골라 제시하므로 서적, 뉴스, 영화 등 다양한 분야에서 매우 유용하게 실제로 활용된다. 본 연구에서는 이러한 시스템에서 가장 중요한 역할을 하는 유사도 측정의 새로운 방법을 제시하였다. 제안 방법은 기존의 전통적 방법과는 달리, 각 사용자의 과거 평가치를 기준으로 상대적 값의 평가치를 반영하는 것이다. 이러한 특성 때문에 기존 유사도 공식에서 공통평가항목개수가 적을 경우 대두되었던 극단적인 유사도 값의 발생 문제를 극복할 수 있었다.

다양한 조건하에서의 실험을 통하여 제안 방법의 우수성을 확인하였는데, 특히 평가등급의 범위가 큰 데이터셋에 대해 전통적 유사도 측정 방법들을 크게 능가할 수 있었다. 그러나, 데이타셋의 희소성 수준이 매우 높거나, 정의된 평가등급 범위가 크지 않을 때는 제안 방법의 성능에 상대적인 한계가 있으므로, 이를 개선하기 위한 향후 노력이 요구된다.

참 고 문 헌

[1] H. Ahn (2008). A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem, *Information Sciences, 178-1*, 37 - 51.

[2] D. Anand and K. Bharadwaj (2010). *Adaptive user similarity measures for recommender systems: A genetic programming approach*. The 3rd IEEE International Conference on Computer Science and Information Technology.

[3] J. Bobadilla, A. Hernando, F. Ortega, and A. Gutierrez (2012). Collaborative filtering based on significances. *Information Sciences, 185-1*, 1-17.

[4] J. Bobadilla, F. Serradilla, and J. Bernal (2010). A new collaborative filtering metric that improves the behavior of recommender systems. *Knowledge-Based Systems, 23-6*, 520-528.

[5] M. Gao, Z. Wu, and F. Jiang (2011). Userrank for item-based collaborative filtering recommendation. *Information Processing Letters, 111-9*, 440-446.

[6] D. Goldberg, D. Nichols, B.M. Oki, and D. Terry (1992). Using collaborative filtering to weave an information Tapestry. *Communications of the ACM, 35-12*, 61-70.

[7] B. Jeong, J. Lee, and H. Cho (2010). Improving memory-based collaborative filtering via similarity updating and prediction modulation, *Information Sciences, 180-5*, 602 - 612.

[8] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, and J. Riedl (!997). GroupLens: applying collaborative filtering to usenet news. *Communications of the ACM, 40-3*, 77-87.

[9] G. Koutrica, B. Bercovitz, and H. Garcia (2009). FlexRecs: expresing and combining flexible recommendations. In *SIGMOD*, 745-757.

[10] L. Ren, J. Gu, and W. Xia (2011). A weighted similarity-boosted collaborative filtering approach. *Energy Procedia, 13*, 9060-9067.

[11] P. Resnick, N. Lakovou, M. Sushak, P. Bergstrom, and J. Riedl (1994). Grouplens: An open architecture for collaborative filtering of netnews. In Proc. *the ACM conference on Computer supported cooperative work*, ACM Press, 175-186.

[12] B.M. Sarwar, J.A. Konstan, A. Borchers, J. Herlocker, B. Miller, and J. Riedl (1998). Using filtering agents to improve prediction quality in the GroupLens research collaborative filtering system. *Proc. the 1998 ACM Conference on Computer Supported Cooperative Work*, 345-354.

[13] U. Shardanand and P. Maes (1995). Social information filtering: algorithms for automating ''word of mouth''. *In Proc. SIGCHI Conf. Human Factors in Computing Systems*, Denver, Colorado, United States, 210-217.

[14] X. Su and T.M. Khoshgoftaar (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence 2009*, Article ID 421425, 19 pages

저 자 소 개

이 수 정

1985 이화여자대학교 과학교육과
1990 Texas A&M 대학교 컴퓨터 공학과(석사)
1994 Texas A&M 대학교 컴퓨터 공학과(박사)
1994～1998 삼성전자 통신개발실 선임연구원
1998～현재 경인교육대학교 컴퓨터교육과 교수
관심분야 : 컴퓨터교육, 추천시스템, 웹정보필터링
e-mail : sjlee@gin.ac.kr