이웃 선정 조건에 따른 협력 필터링의 성능 향상 분석

이수정†

요 약

협력 필터링을 통한 추천 시스템은 정보 검색 편의성을 제공함으로써 다방면에서 성공적으로 활용되어왔다. 유사도 측정은 추천인들의 범위를 결정하는 기준이 되기 때문에 이러한 시스템의 성능을 좌우하는 결정적 요소이다. 본 연구에서는 기존의 유사도 측정 공식에서 산출되는 유사도값의 분포를 분석하고, 유사도값과 공통평가항목수와의 관계를 조사하였다. 이를 통해 발견된 문제점을 보완하기 위하여 유사도값의 제한을 통하여 신뢰할 만한 추천인들을 선정하는 방법을 제시하였다. 실험 결과, 유사도의 상한값과 하한값을 동시에 제한하는 방법이 기존보다 월등한 성능 향상을 가져왔다. 특히 적은 수의 최인접이웃을 참조했을 때 두드러졌는데, 코사인 유사도에 대해서는 최대 0.047, 피어슨에 대해서는 최대 0.03의 추천 성능 향상을 보였다. 이 결과는 피어슨과 코사인 유사도를 이용하는 협력필터링 시스템에서 매우 높거나 낮은 유사도의 이웃의 평가 등급은 참조하지 않는 것이 바람직함을 암시한다.

주제어 : 협력 필터링, 추천 시스템, 유사도 척도

Analysis of Performance Improvement of Collaborative Filtering based on Neighbor Selection Criteria

Soojung Lee†

ABSTRACT

Recommender systems through collaborative filtering has been utilized successfully in various areas by providing with convenience in searching information. Measuring similarity is critical in determining performance of these systems, because it is the criteria for the range of recommenders. This study analyzes distributions of similarity from traditional measures and investigates relations between similarities and the number of co-rated items. With this, this study suggests a method for selecting reliable recommenders by restricting similarities, which compensates for the drawbacks of previous measures. Experimental results showed that restricting similarities of neighbors by upper and lower thresholds yield superior performance than previous methods, especially when consulting fewer nearest neighbors. Maximum improvement of 0.047 for cosine similarity and that of 0.03 for Pearson was achieved. This result tells that a collaborative filtering system using Pearson or cosine similarities should not consult neighbors with very high or low similarities.

Keywords : Collaborative Filtering, Recommender System, Similarity Measure

† 정 회 원: 경인교육대학교 교수(교신저자)
논문접수: 2015년 2월 18일, 심사완료: 2015년 5월 14일, 게재확정: 2015년 6월 2일
* 본 논문은 2012년 교육과학기술부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2012R1A1A3012320)

1. 서론

현재 인터넷은 매우 급격하게 발달하고 확산되어 그 사용자 수가 기하급수적으로 늘어나고 있으며, 특히 스마트폰을 이용하여 더욱 쉽게 인터넷망에 접속할 수 있게 되었다. 이로 인해 인터넷의 가장 큰 장점인 정보 검색이 손쉽게 이루어져 다양한 측면에서 생활의 편리함을 가져왔다. 그러나, 이와 같은 장점에도 불구하고 그 혜택을 충분히 누리지 못할 가능성이 커졌는데, 그 주된 이유는 정보의 홍수 때문이다.

정보 과부하 문제를 해결하기 위한 매우 유용한 방법으로 추천 시스템이 대두되었다[1]. 추천 시스템은 사용자의 흥미에 부합하는 웹 환경을 조성하여 주는 개인화 작업이 그 본질을 이루며 웹 개인화의 가장 유명한 예이다. 이러한 시스템을 통해 고객들은 원하는 상품을 찾을 수 있도록 도움을 받을 수 있는데, 현재 서적, 영화, 음악 등 다양한 분야에서 성공적으로 서비스되고 있다. 예를 들어, Tapestry system, GroupLens, Video Recommender, Ringo, Amazon.com, PHOAKS system 등이 있다[2].

개인화를 실행하는 주된 방법인 협력 필터링(collaborative filtering) 방법의 주요 아이디어는 유사한 흥미를 가진 여러 다른 사용자들의 의견에 따라 항목을 추천하는 것이다. 이 방법의 장점은 추천되는 항목의 내용을 고려하지 않기 때문에, 다른 많은 사람들이 선호한다는 이유만으로, 현 사용자가 선호할만한 새로운 항목들을 발견할 수 있다는 점이다. 그러나 축적된 사용자 선호 정보가 없는 새로운 항목은 추천될 수 없고, 많은 사용자들의 평가를 필요로 한다는 단점이 있다.

본 연구에서는 협력 필터링 시스템의 예측 성능 및 추천 성능을 향상시키는 방법을 제시한다. 주된 아이디어는 추천 작업을 위하여 평가치를 참조할 다른 사용자들과의 유사도의 가장 적합한 상한선과 하한선 기준을 실험을 통해 알아낸다. 논문의 구성은 다음과 같다. 2절에서는 관련 연구를 기술하고, 3절에서 제안 방법 및 관련 실험 결과를 설명하며 4절에서 논문의 결론을 맺는다.

2. 배경

2.1 사용자 기반 협력 필터링

협력 필터링을 이용한 추천 시스템은 크게 메모리 기반과 모델 기반으로 분류되며, 전자는 다시 사용자 기반과 항목 기반으로 구분한다[2]. 본 연구에서는 사용자 기반 협력 필터링 시스템에 초점을 둔다. 이 시스템의 핵심은 두 명의 사용자가 유사한 기호를 갖고 있다면, 다른 새로운 항목들에 대해서도 유사한 반응을 보일 것이라는 가정에 기초한다. 이를 바탕으로 하여, 새로운 항목들에 대해 임의의 사용자가 부여할 등급을 예측할 수 있으며, 가장 높은 예측등급을 부여받은 항목들을 추천하는 것이다. 정확한 예측을 위하여, 사용자들이 과거에 부여하였던 모든 등급 이력을 보관해야 하며, 공통 평가 항목에 대해 부여하였던 등급들을 토대로 사용자간 유사도를 계산한다.

그러므로, 시스템에서 예측한 결과, 높은 등급을 받은 항목들을 실제로 사용자가 선호할 것인지가 시스템의 질을 판단하는 주요 요소이다. 이를 위하여 현 사용자와 가장 비슷한 성향을 가진 다른 사용자들의 평가 기록을 참조해야 한다. 따라서, 정확한 사용자 간의 유사도 산출은 매우 중요한 연구주제 중 하나이다[3][4]. 현 사용자 u에게 새로운 항목 x에 대해, 그 평가 예측치를 산출하는 절차는 다음과 같다.

1. u와 가장 유사한 사용자들을 구하고, 이들을 u의 최인접 이웃이라고 정의한다.
2. u의 최인접 이웃 중에서 x를 평가했던 사용자들을 N_u라고 표기한다.
3. u가 x에 대해 부여할 평가치, $r_{u,x}$는 아래와 같이 Resnick's formula에 의해 예측한다[3]. 이 때, $\overline{r_u}$는 u의 평균 평가치, sim(u, v)는 사용자 u와 v 간의 유사도값이다.

$$r_{u,x} = \overline{r_u} + \frac{\sum_{v \in N_u} sim(u,v) \times (r_{v,x} - \overline{r_v})}{\sum_{v \in N_u} |sim(u,v)|} \quad \text{.....(식 1)}$$

2.2 관련 연구

두 사용자 간의 유사도 측정은 최인접 이웃을

결정하는 기준이 되므로 매우 중요하다. 왜냐하면 협력 필터링은 (식 1)에서와 같이 이웃의 평가치와 유사도값을 바탕으로 하여 미평가된 항목의 평가치를 추정하고, 높게 추정 평가된 항목들을 추천하는 방식이기 때문이다. 결과적으로, 유사도값을 제대로 측정하는 일은 추천된 항목들이 사용자가 선호할 것인지 아닌지와 결부되므로 시스템의 성능에 큰 영향을 끼친다.

새로운 유사도 측정 방식을 개발하려는 연구는 지속적으로 진행되어 왔다. Kwon과 Hong은 기조의 MSD(Mean Squared Difference)[5]의 향상된 방식을 제안하였는데, 이는 moment를 근간으로 하여 평가척도의 범위 및 각 평가등급 차이의 k 제곱승을 고려한 것이다[6]. Ren 외 2인은 Pearson이나 cosine에 weight를 부여함으로써 이들 두 방법을 발전시켰는데, weight는 두 항목을 모두 평가한 상대적 사용자 수를 표현한 것으로써. 즉, 두 개 항목을 모두 평가한 사용자들이 상대적으로 많을수록 유사도는 높아지도록 하였다[7]. MovieLens 데이터셋에서 실험한 결과, 기존의 피어슨이나 코사인보다 좋은 성능을 보인 것으로 발표하였다. Gao 외 2인은 유사도 계산에 있어서 모든 사용자에게 동일한 가중치를 부여하는 기존 방식에 문제점을 제기하여, 중요한 가중치를 갖는 사용자들에게 우선순위를 부여한 채로 기존공식을 개선하는 방법을 제안하였다[8].

이와 같이 기존 공식에 근간을 두지 않고, 전혀 새로운 방법의 연구결과도 발표되었는데, 대표적으로 Ahn이 제시한 PIP라는 유사도 측정 방식은 각각의 평가치 쌍의 크기 차이를 고려하여, 두 평가치가 평균 또는 중앙값으로부터 떨어진 정도와 방향을 측정하였다[9]. 그러나, 각 공통평가항목들이 서로 독립적으로 취급되어 사용자의 전반적인 평가 패턴을 고려하지 않는다는 문제점이 있다. 이 밖에 보다 정교한 측정 방법을 개발하기 위하여 데이터 마이닝 기법을 이용한 연구 결과들이 있는데, [10]이나 [11]에서 제시한 유전자 기법을 이용한 측정 방식들이 이에 속한다.

2.3 유사도 산출

현재까지 주로 사용된 유사도 측정방법으로는 피어슨(Pearson) 상관계수와 코사인(cosine) similarity가 있다[5]. 사용자 u와 v 사이의 피어슨 상관계수는 다음과 같이 산출하며, [-1, 1] 범위 내의 값을 갖는다. 이 때 I는 두 사용자가 공통으로 평가한 항목들의 집합이다.

$$sim(u,v) = \frac{\sum_{i \in I}(r_{u,i} - \overline{r_u})(r_{v,i} - \overline{r_v})}{\sqrt{\sum_{i \in I}(r_{u,i} - \overline{r_u})^2 \sum_{i \in I}(r_{v,i} - \overline{r_v})^2}}$$

코사인 유사도는 각 사용자를 벡터로 간주하여 두 벡터 간의 cosine 값을 유사도 값으로 하며, 아래 식으로 측정하며, [0, 1] 사이의 값을 갖는다[9].

$$sim(u,v) = \frac{\vec{u} \bullet \vec{v}}{\| \vec{u} \| \times \| \vec{v} \|} = \frac{\sum_{i \in I} r_{u,i} r_{v,i}}{\sqrt{\sum_{i \in I} r_{u,i}^2} \sqrt{\sum_{i \in I} r_{v,i}^2}}$$

그러나, 이들 공식은 특정 조건에서 그 한계를 보이는데, 예를 들어, 공통 평가항목수가 단 하나뿐일 때 유사도의 최대값인 1이 산출되거나, 공통 평가항목들에게 일정한 평가치를 부여했을 때 유사도를 산출할 수 없다든지 하는 경우이다. 이에 대한 상세한 내용은 [12]에 기술되어 있다. 따라서, 이러한 단점들을 극복하고 성능 향상을 이루기 위해 다수의 연구자들이 유사도 공식을 개발하여 왔다[3][4][9]. 본 연구에서는 전통적으로 가장 널리 사용되어 왔던 피어슨과 코사인 공식에 초점을 두어, 이들의 산출값에 따른 협력 필터링의 성능을 측정하고, 가장 적합한 상한선과 하한선의 유사도 값을 실험적으로 도출한다.

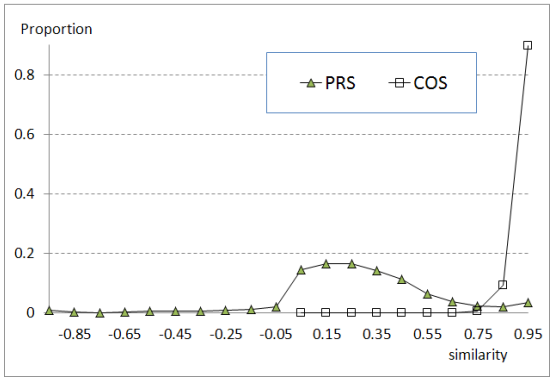
3. 실험 연구

3.1 유사도값의 분포

기존의 두가지 대표적 유사도 산출 공식인 피어슨과 코사인 공식의 산출값의 분포를 알아보았다. <그림 1>은 MovieLens 데이터셋[13]에 대하여 실험 측정한 결과로서, x값은 각 구간의 중위값을 표현하였다. 코사인(COS)과 피어슨(PRS) 유사도값의 분포 차이가 매우 크다는 것을 주목할 수 있다. 즉, 코사인은 대개 0.8 이상의 큰 값들을 가지며, 0.9 이상의 값들이 약 90%를 차지하였다.

반면에 피어슨은 음의 값은 거의 없으며, 0~0.5 사이의 값이 약 73%로서 대부분을 차지하였고, 0.9 이상은 약 3.5%에 불과하였다.

이와 같이 유사도 공식의 종류에 따라서 산출값이 매우 다르다. 특히 코사인은 1에 가까운 큰 값들이 대부분인데 그 원인은 주로 데이터의 희박성(data sparsity)이 클 때 및 공통평가항목수가 적을 때, 코사인 공식이 이를 고려하지 않음으로 최대값을 산출하는 경우가 많기 때문이다.



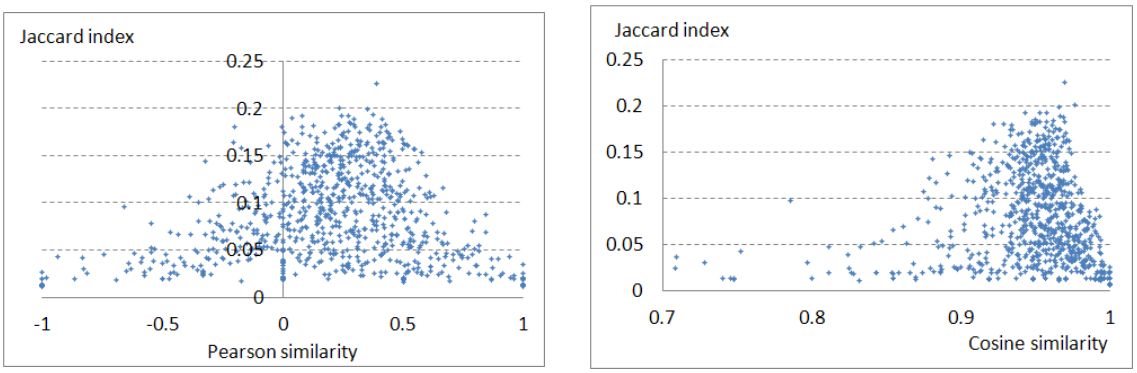
<그림 1> 피어슨과 코사인 유사도값의 분포

3.2 유사도값과 공통평가항목수의 분포

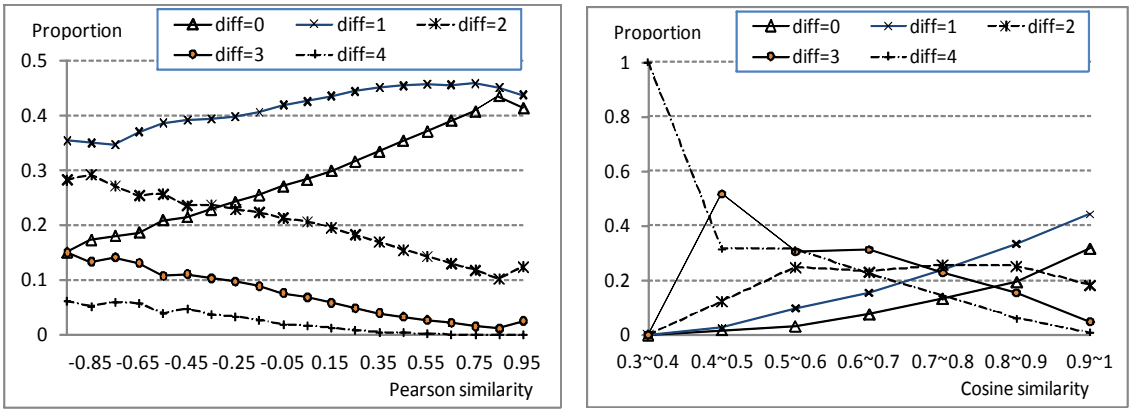
유사도 공식에 대한 신뢰 여부를 살펴보기 위하여 유사도값에 따른 공통 평가 항목수의 분포를 조사하였다. 높은 유사도값이 산출된 경우 공통 평가 항목수가 많을 것으로 기대할 수 있다. 공통 평가 항목수는 Jaccard Index를 통하여 측정하였는데, I_u를 사용자 u가 평가한 항목 집합이라고 할 때, 사용자 u와 v간의 Jaccard index, Jaccard(u, v)는 아래와 같이 산출한다[14].

$$Jaccard(u,v) = \frac{|I_u \cap I_v|}{|I_u| + |I_v| - |I_u \cap I_v|}$$

<그림 2>는 MovieLens 데이터셋[13]을 이용하여 실험한 결과이다. 그래프 상의 각 점은 두 사용자 간의 유사도값에 대한 Jaccard index값을 나타낸다. 전반적으로 유사도가 높다고 공통항목수의 비율도 높다고 볼 수 없으며, 특히 코사인의 경우 공통항목수의 비율이 낮아도 1에 가까운 유사도값을 갖는 경우가 많아, 신빙성이 의심된다.



<그림 2> 피어슨과 코사인 유사도 값에 따른 Jaccard index 분포



<그림 3> 피어슨과 코사인 유사도값에 따른 예측치의 정확도 분포

3.3 유사도값에 따른 예측 정확성

본 절에서는 유사도값에 따라서 얼마나 정확히 평가치를 예측하는지를 살펴보았다. 관련 연구에서 매우 널리 사용되어왔던 MovieLens 데이터셋[13]을 이용하여 실험하였고, 각 유사도값 구간별로 예측치와 실제치의 차이(diff)가 어떠한 분포를 나타내는지 알아 보았다. MovieLens는 1~5까지의 평가점을 가지므로 diff는 0~4의 값을 가진다.

<그림 3>에서 피어슨 상관도가 증가할수록 diff=0의 비율이 가장 크게 증가하였다. 즉, 예상한 바대로, 피어슨 유사도 값이 증가하면 예측치는 상대적으로 더욱 정확해짐을 말한다. 마찬가지로 diff=1의 비율 또한 증가 추세를 보였으나, 그보다 큰 diff값이 차지하는 비율은 점차적으로 감소함을 알 수 있다. 이는 유사도가 커질수록 예측의 정확도가 높아짐을 의미한다. 코사인 유사도에 대한 실험 결과 또한 피어슨 결과와 본질적으로 같았다. 다만, 코사인 유사도는 대개 0.8 이상이므로, 그보다 낮은 값에 대한 분포 그래프는 유연하게 나타나지 않았다.

각 유사도 구간별로 diff의 평균(Mdiff)을 비교하여 <표 1>에 나타냈다. 유사도 값의 증가에 따라 Mdiff는 당연히 감소하는 추세를 보였으나, 피어슨의 경우, 가장 큰 유사도 구간에서 Mdiff가 오히려 증가한 것은 주목할 필요가 있다. 또한 <그림 2>에서 코사인 유사도가 1에 가까울 때 Jaccard index값이 작은 경우가 많으므로, 본 연구에서는 유사도 값이 낮을 때뿐만 아니라, 높을 때의 이웃을 배제하고, 평가 예측치의 정확도가 얼마나 향상되는지 알아보았다.

3.4 유사도의 상한선과 하한선 실험

3.4.1 실험 배경

본 연구의 실험은 MovieLens 데이터셋 상에서 실시되었는데, 이 데이타셋은 1682개의 영화에 대해 943명의 사용자가 1부터 5 사이의 등급으로 평가한 기록을 갖고 있으며, 사용자당 100개가 넘는 평가개수를 포함한다. 실험은 인접 사용자(nearest neighbors)와의 유사도에 따른 영향을 조사하기 위한 것이므로, 최인접 사용자(topNN) 수를 5부터 60까지 변화시켜 실험하였다. 실험 결과의 신뢰도를 높이기 위하여 각 실험의 결과는 10회 크로스 확인(10-fold cross validation)의 평균값으로 산출하였는데, 각 회마다 서로 다른 훈련 데이타 집합과 시험 데이타 집합을 80:20으로 구성하였다. 훈련 데이터를 이용하여 큰 유사도 순서의 최인접 사용자들을 구하고, 이들의 평가등급을 이용하여 (식 1)을 적용한 예측 평가등급을 산출하였다.

성능 분석을 위하여 기존 문헌에서 가장 흔히 사용되는 두가지 평가 척도를 이용하였는데, 첫째 예측 성능을 알아볼 수 있는 MAE(Mean Absolute Error)는 평가 예측치와 실제치의 차이의 절대값 평균을 말하며, 낮을수록 좋은 성능을 나타낸다[5]. 두 번째로, 협력 필터링 시스템의 추천 성능을 가늠하는 척도로서 대개 precision(P)과 recall(R)을 사용하나, 이 둘의 조화평균값인 F1(2PR/(P+R))으로써 한번에 성능을 평가할 수 있으므로 본 연구에서는 F1 결과를 제시하였다[8]. Precision이란 추천시스템이 추천한 항목들 중에서 실제 평가치가 기준치 이상으로 높은 항목들의 비율을 말한다. Recall은 기준치 이상의 높은 실제 평가치의 항목들 중에서 추천된 항목들의 비율을 의미한다.

<표 1> 피어슨과 코사인 유사도값 구간별 예측치의 정확도 평균(Mdiff)

유사도 / Mdiff	0~0.1	0.1~0.2	0.2~0.3	0.3~0.4	0.4~0.5	0.5~0.6	0.6~0.7	0.7~0.8	0.8~0.9	0.9~1.0
피어슨	1.1076	1.0472	0.9874	0.8271	0.8727	0.8297	0.7833	0.7412	0.6904	0.764
코사인	N/A	N/A	N/A	4	3.0876	2.7718	2.4542	2.0083	1.5516	0.9924

<표 2> 각 방식별 최인접 이웃(topNN)의 유사도값 허용 범위

	PRS-L	PRS-U	PRS-LU	COS-L	COS-U	COS-LU
topNN의 유사도 범위	[0.09, 1]	[-1, 0.8]	[0.09, 0.8]	[0.8, 1]	[0, 0.97]	[0.8, 0.97]

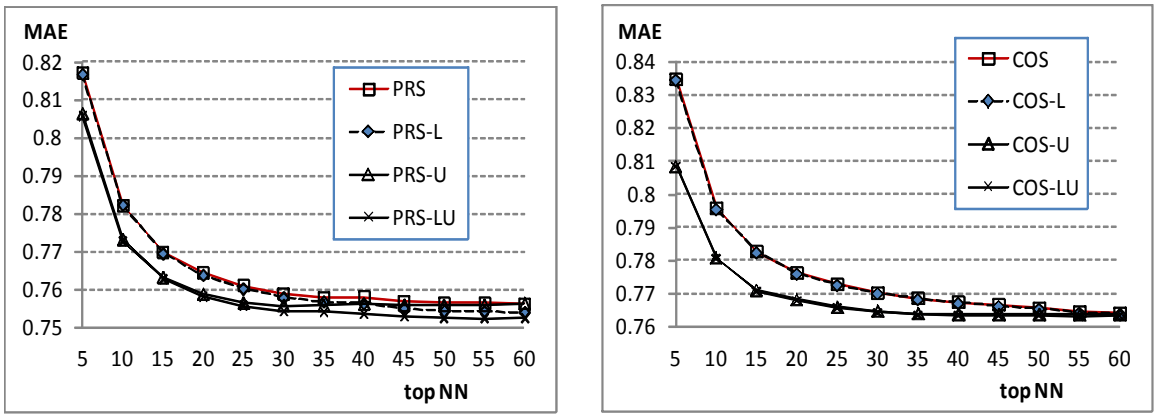
3.4.2 성능 측정 결과

최적의 성능을 위해 최인접 이웃(topNN)이 가져야 할 유사도값 범위를 알아보기 위하여, MovieLens 데이터셋을 이용하여 실험하였다. 최인접 이웃을 결정하는데 있어서, 피어슨(PRS)과 코사인(COS) 유사도의 하한값을 다양하게 하여 실험하였다. 각각의 실험에서 MAE(Mean Absolute Error)를 기준으로 하여 가장 좋은 성능 결과를 가져온 하한값을 구하였다.

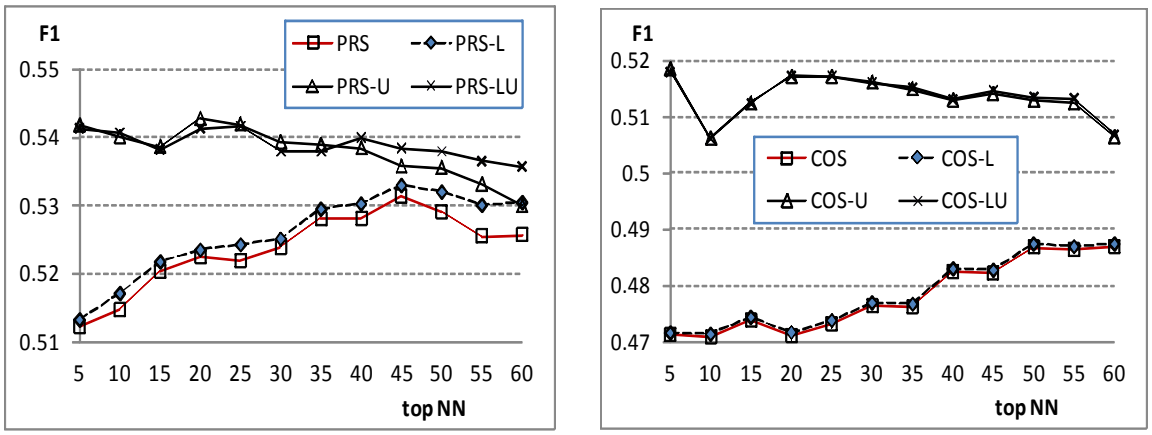
이와 같이 정한 최적의 하한값은 <표 2>에 제시한대로 PRS-L, COS-L로 각각 표기하였는데, PRS-L은 평가치를 참조할 topNN과 현사용자와의 PRS 유사도의 최저치는 0.09임을 의미한다. 마찬가지로, COS-L에서는 그 최저치가 0.8이다. 상한값의 결정에 있어서도 같은 방식을 적용하였으며, <표 2>에서와 같이 PRS-U, COS-U로 표기하였다. 즉, topNN의 PRS 유사도는 0.8까지로 제한하였을 때, 다른 제한값을 적용하였을 때보다 더욱 좋은 성능을 보였다. 또한, COS 유사도값이 0.97 이하인 topNN의 평가치만을 참조할 때가 상대적으로 가장 좋은 성능을 가져왔다. 더 나아가, 이들을 통합한 방식, 즉, 하한값과 상한값을 둘다 적용한 방식도 실험하였는데, 이를 PRS-LU, COS-LU라고 표기하였다. 따라서, PRS-LU 방식에서는 유사도가 [0.09, 0.8] 범위내의 이웃들만을 최인접이웃으로 선정하여 그들의 평가점수를 참조한 것이다. 또한 COS-LU에서는 최인접이웃과의 유사도 범위를 [0.8, 0.97]로 제한하였다. 이로써 3.2와 3.3절에서 기술한대로, 유사도 측정방식의 문제점으로 말미암아, 상대적으로 매우 작은 유사도뿐만 아니라 큰 유사도값의 이웃을 평가참조에 있어서 배제하는 것이 오히려 성능 향상을 가져옴을 증명하는 것이다.

각 방식의 예측 성능은 MAE로서 추출하였는데, <그림 4>에 제시하였다. PRS와 COS 모두 상한값과 하한값을 통합하여 적용한 방식, 즉, PRS-LU와 COS-LU가 가장 좋은 성능을 보였다. PRS의 경우, 방식들 간에 큰 성능 차를 보이진 않았으나, COS-LU는 COS보다 최대 0.026의 성능 향상을 가져왔다. 특히, 두 유사도 공식 모두에서 topNN이 작을수록 효과가 컸으며, 상한선을 적용한 결과가 하한선을 적용한 결과보다 좋은 결과를 보였다. 상한선을 적용한 결과, 즉, 매우 큰 유사도값의 이웃을 topNN에 불포함시키는 것이 더 좋은 성능을 가져오는 이유는 3.2절과 [12]에서 언급한대로, 두 사용자의 평가패턴이 다름에도 불구하고 COS 유사도값이 1인 경우가 발생하기 때문이다.

각 방식의 추천 성능을 알아 보기 위하여 <그림 5>와 같이 F1 성능을 비교하였다. MAE 성능 결과와는 다르게 각 방식의 결과가 차이가 큼을 알 수 있는데, 특히 MAE에서와 마찬가지로 topNN이 작을 때 더욱 차이가 컸다. topNN 값에 상관없이 전반적으로 PRS-LU와 PRS-U의 성능이 다른 두 방식보다 훨씬 좋은데, PRS-LU는 topNN이 25 이하일 때 PRS보다 0.02~0.03의 성능 향상을 가져왔다. 단, PRS-L의 성능은 topNN이 커질수록 점차적으로 가장 크게 향상됨을 주목할 수 있다. 이는 평가치를 참조할 최인접 이웃을 선정할 때 낮은 유사도값을 가진 이웃을 포함시키는 것이 오히려 F1 성능에 있어서 해가 된다는 것을 의미한다.



<그림 4> 각 유사도값 범위에 따른 MAE 성능 비교



<그림 5> 각 유사도값 범위에 따른 F1 성능 비교

F1 성능 측면에서 COS 결과는 크게 두 그룹으로 나뉘는데, COS과 COS-L, COS-U와 COS-LU이다. 이들 두 그룹의 성능 차이가 PRS에서보다 더욱 크게 나타났다. 구체적으로, topNN이 25 이하일 때 0.044~0.047의 성능 차이를 보였다. 결과적으로 <표 2>에서 기술하였듯이, 코사인 유사도값이 0.97보다 큰 인접이웃을 참조하는 것은 오히려 성능을 저하시키는 결과를 가져온 것이다. 다만, <그림 5>에서 두 그룹 간의 성능차이가 topNN이 증가함에 따라 점차 좁혀지는 것을 볼 때, 이같은 상한값 제한의 효과가 topNN이 증가할수록 적어지는 것을 알 수 있다. 이는 평가등급을 참조할 최인접이웃수가 적은 경우에는, 유사도 측정 공식의 오류로 인해 부당하게 큰 유사도를 갖게 된 이웃을 최인접이웃에 포함시키지 않는 것이 큰 효과를 발휘하여 성능을 향상시키지만, 최인접이웃수가 많은 경우엔, 그러한 이웃을 배제하는 것이 적은 비중을 차지하기 때문에 상대적으로 적은 성능 향상 효과를 가져오기 때문이다.

결과적으로, 참조할 인접이웃의 선정 시에, 유사도의 허용 범위를 제한하는 방식은 PRS 보다 COS에서 더욱 효과가 컸으며, 상한값을 제한했을 때 성능 향상이 더 컸으며, 예측 성능보다 추천 성능에 더욱 큰 영향을 미쳤다.

4. 결론

협력 필터링을 통한 추천 시스템은 정보 과부하 시대에 사용자가 선호할 만한 자료들을 선정하여 추천하므로 서적, 뉴스, 영화 등 다양한 분야에서 유용하게 활용되어 사용자의 편의를 크게 향상시키는 역할을 한다. 이러한 시스템에서는 사용자와 유사한 인접 이웃들이 선호했던 자료들을 추천하기 때문에, 유사도의 정확한 측정은 시스템 성능에 있어서 가장 중요한 요소라고 할 수 있다.

본 연구에서는 기존의 코사인과 피어슨 유사도 공식의 단점들을 분석하고, 이를 해결하기 위한 간단한 아이디어를 제시하여, 이를 적용한 실험을 광범위하게 실시하였다. 성능 분석 결과, 이웃과의 유사도의 상한값과 하한값을 제한하는 방법이 예측 성능과 추천 성능 측면에서 모두 기존 방법보다 월등한 향상을 가져왔다. 즉, 피어슨과 코사인 유사도를 이용하는 협력 필터링 추천 시스템에서 매우 높거나 낮은 유사도의 이웃의 평가 등급은 참조하지 않는 것이 바람직하다는 것을 암시한다. 향후 과제로서 제안 방법을 Jester, Netflix 등 특성이 다른 데이터집합에 대해 실험하고, 또 다른 기존 유사도 측정 방법 등에도 적용 실험하여 그 성능을 조사 연구함이 바람직할 것이다.

참 고 문 헌

[1] Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2011). *Recommender Systems - An Introduction*. Cambridge University Press.

[2] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering, 17*(6), 734 - 749.

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

[4] Liu, H., Hu, Z., Mian, A., Tian, H., & Zhu, X. (2014). A new user similarity model to improve the accuracy of collaborative filtering. *Knowledge-Based Systems, 56*, 156 - 166.

[5] Su, X. & Khoshgoftaar, T.M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*.

[6] Kwon, H.-J. & Hong, K.-S. (2011). Personalized smart TV program recommender based on collaborative filtering and a novel similarity method. *IEEE Transactions on Consumer Electronics, 57*(3), 1416-1423.

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

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

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

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

[11] Bobadilla, J., Ortega, F., Hernando, A., & Alcal J. (2011). Improving collaborative filtering recommender system results and performance using genetic algorithms. *Knowledge-Based Systems, 24*(8), 1310-1316.

[12] Lee, S. (2011). A rank-based similarity measure for collaborative filtering systems. *Journal of the Korean Association of Computer Education, 14*(5), 97-104.

[13] MovieLens, http://www.grouplens.org/

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

이 수 정

1985 이화여자대학교 수학교육과
1990 Texas A&M 대학교 컴퓨터과학과(석사)
1994 Texas A&M 대학교 컴퓨터과학과(박사)
1994~1998 삼성전자 통신개발실 선임연구원
1998~현재 경인교육대학교 컴퓨터교육과 교수
관심분야: 컴퓨터교육, 추천시스템, 정보필터링
E-Mail: sjlee@gin.ac.kr