협력 필터링 시스템을 위한 순위 기반의 유사도 척도

이수정†

요 약

협력 필터링은 공통된 흥미를 가진 다른 사용자들로부터 정보를 획득하여 그들의 의견에 따라 웹 사이트를 추천하는 방법이다. 과거 수년간, 이 방법은 서적, 식품, 영화 등 다양한 e-commerce 영역에서 사용되었다. 본 논문에서는 협력 필터링 시스템에서 추천 항목을 결정하기 위한 사용자 간의 유사도 측정 방법을 제시하였다. 기존 연구에서는 사용자가 부여했던 전체 평가등급들의 분포를 고려하지 않은 채 각 평가등급을 독립적으로 취급하여 사용자간 유사도를 산출하였으나, 본 연구에서는 사용자의 평가 등급 범위 내에서의 등급의 위치와 순위 정보를 이용하여 유사도를 산출하였다. 실제 데이터집합 상에서 평균 절대 오차의 성능을 측정한 결과, 대부분의 기존 방법들에 비해 제안 방법은 매우 우수하였고, 특히 정해진 등급범위가 클 경우에 그러하였다.

주제어 : 추천 시스템, 웹 개인화, 협력 필터링, 유사도

A Rank-based Similarity Measure for Collaborative Filtering Systems

Soojung Lee†

ABSTRACT

Collaborative filtering is a methodology to recommend websites by obtaining data and opinions from the other users with similar tastes. During the past few years, this method has been used in various fields such as books, food, and movies in e-commerce systems. This study addresses the computation of similarity between users to determine items to be recommended in collaborative filtering systems. Previous studies measured similarity between users by treating each user's ratings independently without considering the distribution of the user's ratings. In contrast, this study measures similarity by utilizing position and rank information of each rating in the range of the user's ratings. The result of the experiments on the real datasets demonstrated that the proposed method improves the mean absolute error significantly, compared to the previous methods, especially when the predetermined range of ratings is large.

Keywords : Recommender System, Web Personalization, Collaborative Filtering, Similarity

† 정 회 원: 경인교육대학교 컴퓨터교육과 교수
논문접수: 2011년 07월 26일, 심사완료: 2011년 09월 23일, 게재확정: 2011년 09월 25일

1. 서 론

대부분의 검색 엔진은 사용자가 누구이건 간에 상관없이 질의어에 대해 동일한 검색 결과를 제공한다. 따라서 검색 작업에 상당한 시간과 노력을 투자하는데도 불구하고 사용자는 자신이 선호하는 흥미 있는 정보를 얻지 못할 수 있다. 이러한 문제를 경감시키기 위해, 사용자의 흥미에 부합하는 웹 환경을 조성하여 주는 개인화 작업이 주목을 받고 있다[1]. 웹 개인화는 특정 사용자의 요구에 맞도록 웹 사이트를 적응시켜 나가는 과정으로서 주로 네 가지 범주가 있는데, 웹 검색경로 예측, 개인화된 정보 보조, 검색 내용의 개인화, 그리고 검색 결과의 개인화이다[2]. 웹 개인화의 가장 유명한 예는 추천 시스템으로서 주로 고객이 원하는 상품을 찾을 수 있도록 도움을 주는데, Amazon.com[3], MovieLens[4], VERSIFI Technologies[5], GroupLens 시스템[6] 등이 있다.

개인화를 실행하는 주된 방법은 협력 필터링(collaborative filtering)과 내용 기반 필터링(content-based filtering)이다[7]. 두 방법은 모두 사용자가 관심을 보일만한 항목들을 식별하여 정보 과부하를 경감시키려 하였다. 협력 필터링은 공통 흥미를 가진 여러 다른 사용자들의 의견에 따라 항목을 추천하며 서적, 식품점, 예술과 엔터테인먼트 등 다양한 영역에서 사용되었다. 이 방법의 주요 장점은, 추천되는 항목의 내용을 고려하지 않기 때문에, 다른 많은 사람들이 선호한다는 이유 하나만으로, 새로운 항목들을 발견할 수 있다는 점이다. 그러나 축적된 사용자 선호 정보가 없는 새로운 문서는 추천할 수 없고, 많은 사용자들의 평가를 필요로 하는 단점이 있다.

이와는 반대로 내용 기반 필터링은 내용 분석을 토대로, 사용자 요구나 선호도를 반영한 프로필을 구축한다. 프로필의 구축은 사용자가 직접 입력하거나 또는 그의 행위로부터 간접적으로 학습한다. 예로서, Persona 시스템[8]은 사용자의 검색 경로로부터 관심과 비관심 영역을 분류하여 학습하고, [9]의 시스템은 사용자의 검색 이력으로부터 선호 범주를 학습한다. 이 같은 내용기반 필터링 방법은 아직 미평가된 문서에 대해 그 내용을 살펴보아 사용자의 흥미 여부를 예측할 수 있지만, 프로필에 축적된 흥미도 외에 새로운 흥미로운 정보를 발견할 수 없다는 단점이 있다.

협력 필터링을 이용한 추천 시스템은 크게 메모리 기반과 모델 기반으로 분류된다[10]. 후자는 과거 등급들을 이용하여 모델을 학습하고 이를 토대로 새로운 항목의 등급을 예측한다[11]. 모델의 학습은 주로 확률, 통계나 기계 학습 기법을 활용하지만 좀 더 복잡한 기법으로서 Bayesian, 선형 회귀분석, 최대 엔트로피 모델, 마코프 결정 프로세스 등이 있다. 메모리 기반 시스템은 두 명의 사용자가 유사한 기호를 갖고 있다면, 다른 새로운 항목들에 대해서도 유사한 반응을 보일 것이라는 가정에 기초한다. 이 추천 시스템은 사용자에게 새로운 항목들에 대해 그가 부여할 등급을 예측하여, 가장 높은 예측등급의 항목을 추천한다. 따라서 가능한 한 정확히 예측하는 것이 관건이다. 이를 위해 우선 과거에 사용자들이 부여하였던 모든 등급 이력을 보관하고, 공통 평가 항목에 대해 부여하였던 등급들을 토대로 사용자간 또는 항목 간 유사도를 계산한다.

메모리 기반 협력 필터링 시스템은 사용자 기반과 항목 기반으로 분류하는데[12], 전자는 N명의 최인접 이웃이 부여한 최고 등급의 항목을 추천하고, 항목 기반 시스템인 경우는 사용자가 선호했던 항목들과 가장 유사한 항목을 추천한다. 메모리 기반 시스템에서 새로운 항목의 등급을 예측하기 위해서 과거에 부여했던 등급들의 평균을 주로 이용하였으나, [13]에서는 과거등급들의 범위를 기반으로 예측하는 방법을 제시하고 Book-crossing 데이터셋[14]에서 제안 방법의 우수성을 입증하였다. 사용자가 실제로 선호할 항목을 추천하는 것은 시스템 성능을 결정하는 주요 요소이며, 이에 기초로서 정확한 유사도 산출은 매우 중요한 연구주제 중 하나이다.

본 논문에서는 메모리 기반 협력 필터링을 통한 추천 시스템을 위해서 사용자 간 유사도를 측정하기 위한 방법을 제안한다. 논문의 구성은 다음과 같다. 2절에서는 관련 연구를 기술하고, 3절에서 제안 방법을 설명하며 4절에서 실험을 통한 성능을 입증하고 5절에서 논문의 결론을 맺는다.

<표 1> 피어슨 상관도와 코사인 유사도의 한계점 예시

		피어슨 상관도	코사인 유사도
• 공통 평가항목이 단 한 개일 때		+1, −1, 또는 계산불가	1
• 공통 평가항목이 둘 이상: $r_{u,i}$, $r_{v,i}$가 각각 모두 동일한 값이며 $r_{u,i} \neq r_{v,i}$	$\overline{r_u} = r_{u,i}$, $\overline{r_v} = r_{v,i}$	계산불가	1
	$\overline{r_u} \neq r_{u,i}$, $\overline{r_v} \neq r_{v,i}$	+1 또는 −1	1

2. 유사도 관련 기존 연구

2.1. 피어슨 상관도와 코사인 유사도

유사도 측정은 최인접 이웃을 알아내는 중요한 역할을 하므로 협력 필터링의 성능을 좌우한다. 현재까지 주로 사용되는 측정방법으로 Pearson 상관계수와 cosine similarity가 있다[15]. 사용자 u와 v가 모두 등급을 매긴 항목들의 집합을 I, $r_{u,i}$와 $r_{v,i}$를 각각 사용자 u와 v가 부여한 항목 i에 대한 등급, $\overline{r_u}$와 $\overline{r_v}$를 각각 사용자 u와 v가 부여한 모든 등급들의 평균이라고 할 때, 두 사용자 간의 Pearson 상관계수는 다음과 같이 산출한다.

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

Cosine similarity는 각 사용자를 |I| 차원의 벡터로 간주하여 두 벡터 간 각도의 cosine 값으로써 아래 식으로 측정한다.

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

두 사용자의 공통 평가 항목이 단 하나일 때, 피어슨 상관도와 코사인 유사도는 한계를 갖는데, 코사인 유사도는 1, 피어슨 상관도는 −1, 1, 또는 제수가 0이어서 계산불가이기 때문이다. 공통 항목들이 다수일지라도, 사용자가 이들에 대해 동일한 등급을 부여한다면 유사한 결과를 초래한다. <표 1>에 이들 방식의 단점들을 정리하였다.

2.2. PIP 유사도

[16]에서는 시스템의 신규 사용자들에게 항목을 추천하는 어려움을 해결하기 위하여 새로운 유사도 방법을 고안하였다. 이 방법은 공통 항목에 대한 두 평가등급의 거리차를 기본으로 하되, 평가등급의 중앙값 또는 평균값을 기준으로 두 평가등급이 각기 이들 값에서 멀어져 있는 정도와 방향을 수치화하였다. PIP라고 명명된 유사도는 세 구성 요소의 곱으로 이루어지는데, 즉, Proximity, Impact, Popularity이다. Proximity는 두 등급 차가 작을수록 커지나, 이들이 중앙값에서 서로 반대편에 있다면 원래의 두 배가 차이나는 것으로 간주한다. Impact는 두 등급이 중앙값에서 멀어질수록 커지나, 중앙값의 서로 반대편으로 멀어질수록 오히려 더 작아지는 값을 가진다. 마지막 요소인 Popularity는 평균을 고려하여, 두 등급이 해당 항목의 평균에서 같은 방향으로 멀어질수록 큰 값을 갖도록 고안하였다. <표 2>는 PIP 공식을 구체적으로 제시하였다. $r_{\max}$와 $r_{\min}$은 각각 최대

<표 2> PIP 유사도의 구성 요소

	$Agreement(r_{u,i}, r_{v,i}) = true$	$Agreement(r_{u,i}, r_{v,i}) = false$
$\mathrm{Pro}ximity(r_{u,i}, r_{v,i})$	$(2(r_{\max} - r_{\min}) + 1 - \lvert r_{u,i} - r_{v,i} \rvert)^2$	$(2(r_{\max} - r_{\min}) + 1 - 2\lvert r_{u,i} - r_{v,i} \rvert)^2$
$Impact(r_{u,i}, r_{v,i})$	$(\lvert r_{u,i} - r_{med} \rvert + 1)(\lvert r_{v,i} - r_{med} \rvert + 1)$	$\frac{1}{(\lvert r_{u,i} - r_{med} \rvert + 1)(\lvert r_{v,i} - r_{med} \rvert + 1)}$

μ_i: 모든 사용자의 항목 i에 대한 평균 등급

$$Popularity(r_{u,i}, r_{v,i}) = \begin{cases} 1 + (\frac{r_{u,i} + r_{v,i}}{2} - \mu_i)^2, & \text{if}(r_{u,i} > \mu_i \ \ r_{v,i} > \mu_i) \text{ or } (r_{u,i} < \mu_i \ \ r_{v,i} < \mu_i) \\ 1, & otherwise \end{cases}$$

$$sim(u,v) = \sum_{i \in I} \mathrm{Pro}ximity(r_{u,i}, r_{v,i}) \bullet Impact(r_{u,i}, r_{v,i}) \bullet Popularity(r_{u,i}, r_{v,i})$$

와 최소 평가등급, r_{med}는 평가등급의 중앙값, 즉 $(r_{\max}+r_{\min})/2$이며 다음의 부울변수를 이용한다.

$$Agreement(r_{u,i}, r_{v,i}) = \begin{cases} true, \text{ if } r_{u,i}, r_{v,i} > r_{med} \text{ or } r_{u,i}, r_{v,i} < r_{med} \\ false, \ otherwise \end{cases}$$

PIP의 첫 번째 문제점은 구성 요소들이 독립적이지 않고 다소 중복된 개념을 반영한 점이다. 예로써, 중앙값에서 서로 반대편에 위치한 두 등급 간에 차이가 클수록 Proximity나 Impact는 모두 점차 작아진다. 다른 예로, 중앙값과 평균이 거의 동일할 경우, 이들을 기준으로 같은 방향에 놓인 등급 간 Impact와 Popularity는 두 등급이 클수록 마찬가지로 커진다. PIP의 두 번째 한계점은 더욱 치명적인 것으로 판단되는데, 각 공통평가항목에 대한 두 등급의 유사도를 다른 공통평가항목들의 등급과 상관없이 독립적으로 산출한다는 점이다. 예로써, 사용자 u와 v 사이에 공통평가항목 i와 j가 존재할 때, $r_{u,i}$=5, $r_{u,j}$=5, $r_{v,i}$=2, $r_{v,j}$=4인 경우와 $r_{u,i}$=2, $r_{u,j}$=5, $r_{v,i}$=5, $r_{v,j}$=4인 경우의 PIP 유사도는 같다. 그러나 이들 두 경우에 사용자 u와 v의 평가 의도가 확연히 다르므로 유사도는 다른 값이어야 한다. 이러한 문제점의 원인은 사용자의 전체적인 평가의도를 고려하지 않고, 각 항목의 평가등급을 별도로 취급하기 때문이다.

3. 제안 방법

앞 절에서 언급한 기존 유사도 산출 방식의 가장 큰 문제점의 원인을 한마디로 요약하면, 각 공통 평가항목을 독립적으로 취급한다는데 있다. 그러나, 각 사용자의 평가등급 결정은 주관적이어서, 어떤 사용자에게 등급 r은 낮은 등급이나, 다른 이에게는 높은 등급일 수 있다. 이러한 주관적 판단의 차이를 수치화하여 유사도 산출에 반영하기 위해, 본 연구에서는 평가 범위 전체를 고려한 방법을 제시한다. 즉, 사용자가 부여한 임의의 평가등급이 그가 부여한 과거 평가등급들 중 어떤 순위를 차지하는지를 측정한 후, 각 사용자의 평가등급의 순위 정보끼리 비교하여 유사도를 산출한다. 두 사용자 u와 v가 등급을 매긴 공통 항목들의 집합을 I라 하고, $rank_u(r_{u,i})$를 $r_{u,i}$의 순위라고 할 때, 유사도는 다음과 같다.

$$rank_u(r_{u,i}) = \frac{|\{j | r_{u,i} < r_{u,j}\}|}{\text{사용자 } u \text{가 평가한 항목들의 총개수}}$$

$$sim(u,v) = 1 - \frac{1}{|I|}\sum_{i \in I} |rank_u(r_{u,i}) - rank_v(r_{v,i})|$$

순위 정보는 임의의 평가등급을 사용자의 전체 평가범위를 기준으로 부여한 수치이긴 하나, 사용자가 부여한 평가등급들이 정규 분포를 이룬다고 볼 수 없으므로, 만약 대개의 평가등급들이 높고, 한두 개의 평가등급이 매우 낮다면, 높은 등급일지라도 그 한두 개의 평가등급 순위와 별로 차이없는 낮은 순위를 가질 수 있어 불합리하다. 따라서, 위 식에서처럼 순위뿐만 아니라, 임의의 평가등급이 사용자가 과거에 부여한 평가범위 내에서 어느 위치를 차지하는지의 정보도 고려하도록 한다. 사용자 u가 부여한 평가등급의 최대치와 최소치를 $r_{u,\max}$와 $r_{u,\min}$이라 하고 $pos_u(r_{u,i})$가 $r_{u,i}$의 위치값일 때, 최종 유사도는 다음과 같다.

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

$$sim(u,v) = 1 - \sum_{i \in I} \frac{1}{2|I|}(|rank_u(r_{u,i}) - rank_v(r_{v,i})| + |pos_u(r_{u,i}) - pos_v(r_{v,i})|)$$

이같이 산출한 유사도를 이용하여, 사용자 기반의 협력 필터링은 인접사용자들이 부여한 등급 정보를 취합하여, 높은 등급을 얻은 항목을 추천한다. 따라서 추천 항목의 등급이 사용자가 부여할 실제 등급에 얼마나 부합할지의 예측 정확도가 매우 중요하다. 사용자 u의 최인접 사용자들의 집합을 Nu라 할 때, 사용자 u가 등급을 미부여한 항목 x에 대한 등급 $r_{u,x}$는 대개 다음과 같이 예측한다[15].

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

4. 실험 연구

4.1. 실험 배경

제안한 유사도 방법의 성능을 평가하기 위하여, MovieLens[17]와 Book-Crossing 데이터[14]를 사용하였다. Book-Crossing 데이타는 271,379권의

<표 3> 실험 데이터 집합

Data Set	평가개수	행렬크기 (사용자수×서적수)	희소성수준	사용자 당 평가개수	항목당 피평가개수	평가등급범위
BX5-10	48436	3072 × 2094	0.992470	> 5	>10	1~10
BX5-20	17861	1498 × 480	0.975160	> 5	>20	1~10
MovieLens	100000	943 × 1682	0.936953	> 100	-	1~5

<표 4> 각 데이터집합의 Pearson과 Cosine 유사도 분포

Data Set	공통평가항목이 존재하지 않는 경우(%)	공통 평가항목이 존재하는 경우 각 유사도값의 분포 (%)								
		비율 (%)	Pearson 유사도 (PRS)						코사인 유사도(COS)	
			0	1	-1	zero divide	(0, 1)	(-1, 0)	1	(0, 1)
BX5-10	95.47	4.53	0.15	36.10	29.0	5.22	15.94	12.40	80.84	19.16
BX5-20	91.49	8.51	0.19	33.22	27.5	5.98	17.95	14.09	76.01	23.99
MovieLens	7.62	92.38	0.18	4.93	4.35	0.36	58.83	31.15	12.13	87.87

서적에 대하여 278,858명의 사용자들이 평가한 1-10 사이의 등급을 1,149,790개 포함하고 있다. 그러나 많은 사용자 및 서적의 평가수가 매우 적어 성능 측정의 정확성이 떨어지므로, 원데이터 행렬의 희소수준을 향상시키는 사전 처리를 하여 [15]에서처럼 희소수준이 다른 두 집합을 산출하고 각각 BX5-10과 BX5-20으로 명명하였는데, 이는 희소수준이 성능에 미치는 영향을 알아보고자 함이다. 희소수준이란 사용자수×항목수의 행렬 내 평가가 매겨지지 않은 요소의 비율이며, (값이 0인 요소 개수)/(행렬의 크기)로 산출한다. 각 집합에 대해 <표 3>에 상세 기술하였다.

기존에 많이 활용되었던 유사도 계산 방식인 Pearson correlation(PRS), Cosine similarity (COS), PIP, 그리고 제안 방식(REL)의 성능을 비교하기 위해, 이들 각각을 이용한 결과 등급이 사용자가 부여한 실제 평가등급과 얼마나 부합되는지를 측정하였다. 이러한 예측 정확도는 주로 MAE(Mean Absolute Error)[10]로 나타내며 다음과 같이 정의된다.

$$MAE = \frac{\sum_{x} |r_{u,x} - p_{u,x}|}{N}$$

$r_{u,x}$: 항목 x에 대해 사용자 u가 부여한 등급

$p_{u,x}$: 항목 x에 대한 사용자 u의 등급 예측값

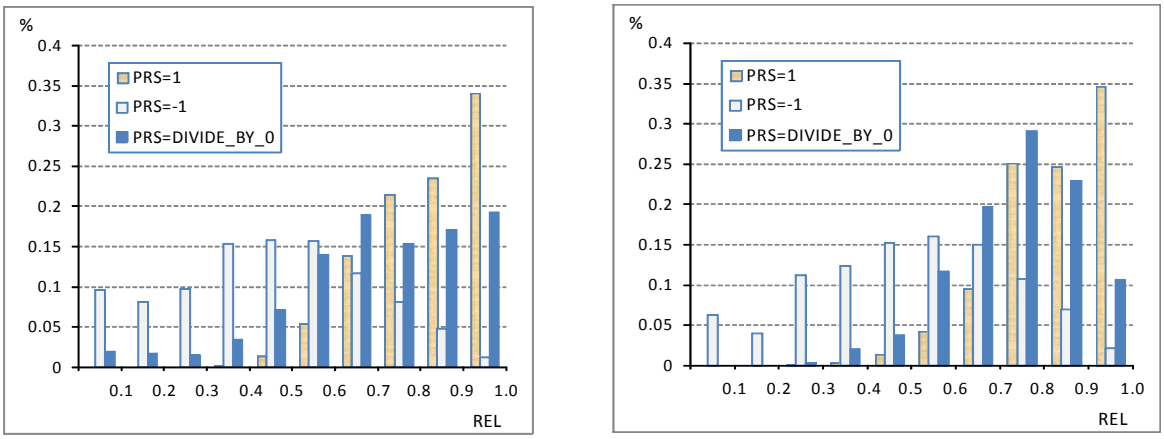
N : 성능 평가 대상 항목의 총개수

실험 결과의 신뢰도를 높이기 위해 MAE는 5회 크로스 확인(5-fold cross validation)의 평균으로 산출하였고, 각 회마다 서로 다른 훈련 데이터와 시험 데이타 집합을 80:20으로 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. 유사도 값 분포

2.1절에서 언급한 대로 Pearson 상관도(PRS)는 극단적인 값, 즉, 0, 1, 또는 -1의 값을 갖기 쉬운데, 이는 특히 두 사용자간에 공통된 평가항목수가 극히 적은 경우에 그러하다. 각 데이터집합별로 이들 극단 값의 산출 분포를 알아본 결과 <표 4>와 같았다. 공통 평가 항목이 존재하여 유사도 산출이 가능한 비율은 희소수준이 가장 낮은 MovieLens가 92.38%로 가장 컸으며, 그다음 BX5-20, BX5-10 순이었다. 유사도 산출이 가능한 경우에 Book-Crossing 데이터집합에서는 PRS가 1인 경우가 36.1%과 33.22%로서 가장 컸고 그 다음으로 큰 비중을 차지하는 유사도 값은 -1임을 알 수 있다. 따라서 새로운 사용자가 많거나 사용자가 항목에 대한 평가를 거의 하지 않는 경우, PRS는 신뢰성이 크게 떨어짐을 알 수 있다. 이러한 현상은 COS에서도 마찬가지인데, 값이 1인 유사도를 가진 사용자 쌍이 80.84%, 76.01%인 것으로써 증명된다. 한 사용자가 100개 이상을 평



<그림 1> Pearson 유사도의 비정상 값에 대한 해당 사용자의 REL 유사도 값 분포. (a) BX5-20. (b) MovieLens.

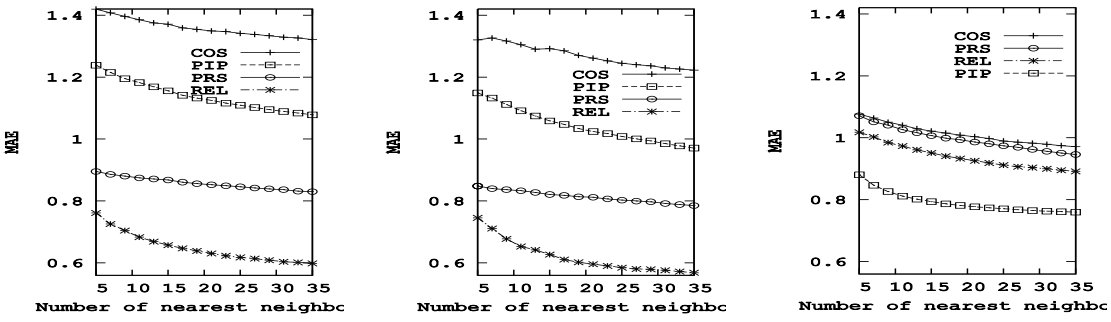
가하는 MovieLens 집합에 대해서조차, PRS는 9.82%, COS는 12.13%의 극단 값을 갖는 것을 볼 때, 희소한 평가수를 가진 대부분의 상업 사이트에 대비하여 보다 정확한 유사도의 산출이 매우 중요하다는 것을 알 수 있다.

PRS 값이 극단일 경우, 해당 사용자 쌍이 갖는 REL 유사도 값을 알아보았다. <그림 1>에 제시한 대로, BX5-20과 MovieLens 집합은 비슷한 양상을 보이는데, PRS가 1인 경우 REL은 대개 0.5 이상이며 1에 근접할수록 더욱 큰 비중을 나타내었다. PRS 산출이 불가한 경우에도 유사한 양상을 보였으며, PRS가 -1일 때 REL값은 대략 0.3~0.7 사이의 중앙에 위치하는 경우가 많았고 그보다 클수록 빈도수도 줄었다. 즉, 제안 공식은 공통 평가항목수가 극히 적어도 극단 값을 드물게 산출하고 대개 0과 1 사이에 고르게 분포한다.

4.2.2. 성능 결과

<그림 2>는 각 유사도 공식의 MAE 성능을 측정한 결과이다. 예측 성능에 대한 인접사용자수(Number of nearest neighbors)의 영향을 조사하기 위하여, 그 수를 5부터 35까지 변화시켜 실험하였다. 각 공식을 이용하여 최인접 사용자들을 결정한 후, 이들이 부여한 등급을 기초로 예측하였을 때 그 정확도를 데이터집합별로 제시하였다. 그림에서 각 방법의 성능은 모든 집합에서 유사한 행태를 보인다. 즉, 인접사용자수가 증가함에 따라 성능이 점차 향상된다. 이는 인접사용자수가 늘어나면 해당 항목에 대해 참조할 평가등급 개수가 증가하므로 예측 정확도가 커지기 때문이다.

<그림 2>(a)와 (b)에서 두 Book-Crossing 데이터집합에서 성능은 매우 유사한데, 다만 희소수준



<그림 2> 각 유사도 공식을 사용한 평가등급 예측 정확도. (a) BX5-10, (b) BX5-20, (c) MovieLens

이 향상된 BX5-20에서 성능이 다소 개선되었다. 유사도 간의 성능 차이는 두 집합 모두에서 매우 큰 것을 볼 수 있다. BX5-10 집합에서, REL은 COS, PIP, PRS 보다 각각 최대 0.72, 0.48, 0.23 향상된 결과를 보이며, 이러한 향상 정도는 인접 사용자수가 증가하면 대체로 함께 증가한다. 또한 BX5-20 집합에서 REL은 COS, PIP, PRS 보다 각각 최대 0.67, 0.44, 0.22 향상된 결과를 가져왔다. 특히, <그림 2>(a)와 (b)에서 PIP와 REL은 나머지 두 방법보다 인접 사용자수가 증가함에 따라 성능이 좀 더 가파르게 개선되었다.

한편, [16]에서 세 데이터집합에 대해 서로 다른 유사도 공식의 성능을 비교하였는데, 매우 근소한 차이를 보였으나, COS 성능이 가장 낮았고, PIP와 PRS는 서로 대등한 성능을 보였다. 단, 이는 인접사용자수를 제한하지 않고 모든 사용자에 대해 실험한 결과이므로, 본 연구 결과와 비교하기엔 무리가 있으나, COS 성능이 가장 낮은 점은 일치한다. 또한 [16]에서는 신규사용자들의 비율이 커짐에 따라 PIP 성능의 우수성을 입증하였으나, 본 연구에서는 신규사용자 비율에 따른 실험은 진행하지 않았으므로, <그림 2>(a)와 (b)에서처럼 REL이 Book-Crossing 데이터집합에서 모든 조건에서 우수하다고 결론 내리기엔 무리가 있다.

<그림 2>(c)의 MovieLens에 대한 실험 결과에서 COS와 PIP의 성능은 현저히 개선되었으나, REL와 PRS의 성능은 Book-Crossing 집합에 비해 저하된 것을 주목할 수 있다. COS와 PIP의 성능 개선 이유는 희소수준이 Book-Crossing 집합보다 매우 향상되었기 때문인 것으로 판단되나, PRS의 경우 <표 2>에 제시한대로 극단 값의 비율이 대폭 줄고 정상 범위 내의 값이 58.83%와 31.15%로 늘어났음에도 불구하고 성능이 저하된 것을 볼 때 PRS 유사도의 성능과 희소수준과의 관련도가 COS에 비해 상대적으로 낮다는 것을 알 수 있다. 상대적 위치와 순위를 기반으로 하는 REL에 대해서는, MovieLens의 평가범위는 1-5이므로, 1-10일 때보다 변별력이 떨어지게 되어 성능이 저하된 것으로 판단된다. MovieLens에서 PIP의 성능은 COS, PRS, REL보다 각각 최대 0.23, 0.21, 0.16 향상되었다. 그러나, REL도 가장 흔히 사용되는 COS와 PRS 보다 좋은 결과를 나타내어, 최대 0.08이 개선되었다. 결론적으로 REL 방법은 평가범위가 큰 경우 그 효과가 극대화되는 장점이 있고, 반면 상대적으로 작은 경우, 기존의 COS나 PRS 방식보다는 월등하나, PIP에 비해 저조한 성능을 보였다.

5. 결 론

협력 필터링을 통한 추천 시스템은 광범위한 자료들 중에서 사용자에게 필요할 만한 자료들만을 골라 제시하므로 서적, 뉴스, 영화 등 다양한 분야에서 매우 유용하게 활용되어 정보화시대에 중요한 역할을 한다. 본 연구는 이러한 시스템에서 추천 항목을 결정하기 위한 사용자 간의 유사도 측정 방법을 제시하였다. 기존 연구에서는 전체 평가등급들의 분포를 고려하지 않은 채 사용자들의 공통 평가항목들의 등급을 독립적으로 취급하였으나, 본 연구에서는 평가 등급 범위 내에서 각 등급의 위치와 순위 정보를 이용하여 유사도를 산출하였다. 제안 방법은 본 연구에서 선택한 두 데이터집합에 대해 대부분의 기존 방법보다 우수한 성능을 보였고 특히 주어진 평가범위가 클 때 매우 우수하였다. 향후 과제로서 제안 방법을 Jester, Netflix 등의 다른 데이터집합에 대해 실험하고, 또 다른 기존 유사도와 비교 실험하여 그 성능을 조사 연구할 것이다.

참 고 문 헌

[1] Arotaritei, D. & Mitra, S. (2004). Web mining: a survey in the fuzzy framework. *Fuzzy Sets and Systems, 148*(1), 5-19.

[2] Zhu, D. & Dreher, H. (2008). Improving web search by categorization, clustering, and personalization. *The 4th International Conf. Advanced data mining and applications,* 659-666.

[3] Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations : item-to-item collaborative filtering. *IEEE Internet Comput.*, *7*(1), 76-80.

[4] Miller, B., Albert, I., Lam, S., Konstan, J., & Riedl, J. (2003). MovieLens unplugged: experiences with an occasionally connected recommender system. *Proc. International Conf. on Intelligent User Interfaces*, 263-266.

[5] Kim, H.-R., & Chan, P. K. (2005). Personalized search results with user interest hierarchies learnt from bookmarks. *7th International Workshop on Knowledge Discovery on the Web*, 158-176.

[6] Good, N., Schafer, J., Konstan, J., Borchers, J., Sarwar, B., Herlocker, J., & Riedl, J. (1999). Combining collaborative filtering with personal agents for better recommendations. *Conference of the American Association of Artificial Intelligence*, 439-446.

[7] Forsati, R., & Meybodi, M.R. (2010). Effective page recommendation algorithms based on distributed learning automata and weighted association rules. *Expert systems with applications, 37*(2), 1316-1330.

[8] Tanudjaja, F., & Mui, L. (2002). Persona: A contextualized and personalized web search. *The 35th Annual Hawaii International Conference on System Sciences,* 67.

[9] Liu, F., Yu, C., & Meng, W. (2004). Personalized web search for improving retrieval effectiveness. *IEEE Trans. Knowl. Data Eng., 16*(1), 28-40.

[10] 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.

[11] Hofmann, T. (2003). Collaborative filtering via gaussian probabilistic latent semantic analysis. *Proc. 26th Ann. Int'l ACM SIGIR Conf,* 259-266.

[12] Wang, T. & Ren, Y. (2009). Research on personalized recommendation based on web usage mining using collaborative filtering technique. *WSEAS Transactions on Information Science and Applications, 6*(1), 62-72.

[13] 이수정 (2011). 협력필터링 시스템을 위한 평가등급 범위 기반의 예측방법. 컴퓨터교육학회논문지, 14(4), 19-27.

[14] http://www.informatik.uni-freiburg.de/~cziegler/BX/

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

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

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

이 수 정

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