퍼지 및 지능시스템 학회 2003, Vol. 13, No. 5, pp. 526-529

# 사용자 성향의 시간적 변화를 고려한 협업 필터링 알고리즘에 관한 연구

# Study on Collaborative Filtering Algorithm Considering Temporal Variation of User Preference

박영용 · 이학성

Young-Yong Park and Hak-Sung Lee

세종대학교 전자공학과

## 요 약

추천 시스템 또는 협업 필터링은 특정 사용자에게 잠재적으로 흥미가 있거나 가치가 있는 항목을 분류하는 방법이다. 유사한 성향을 같는 사용자는 유사한 형태의 항목을 좋아하리라는 가정 하에, 이 방법은 사용자들의 성향에 관한 데이터베이스를 이용하여 아직 평가되지 못한 항목에 대한 평가값을 예측하는데 사용된다. 보통 추천하고자 하는 사람의 성향은 시간에 따라 변할 수 있고 이러한 시간적인 변화는 사용자 성향에 대한 분류 혹은 예측에 대한 정확성을 떨어뜨릴 수 있다. 본 논문에서는 협업 필터링 알고리즘의 예측 성능을 향상하기 위해서 사용자 성향의 시간적 변화를 이용한 방법을 제안하고자 한다. 이를 위해 기존의 통계적 협업 필터링의 일반적인 형태인 GroupLens 시스템의 상관 가중치가 최근 사용자의 유사성을 반영하기 위해 변형되었다. 제안된 방법은 EachMovie 데이터셋을 이용해 평가하였고 GroupLens 시스템과 비교하여 더 나은 예측 결과를 보였다.

## Abstract

Recommender systems or collaborative filtering are methods to identify potentially interesting or valuable items to a particular user. Under the assumption that people with similar interest tend to like the similar types of items, these methods use a database on the preference of a set of users and predict the rating on the items that the user has not rated. Usually the preference of a particular user is liable to vary with time and this temporal variation may cause an inaccurate identification and prediction. In this paper we propose a method to adapt the temporal variation of the user preference in order to improve the predictive performance of a collaborative filtering algorithm. To be more specific, the correlation weight of the GroupLens system which is a general formulation of statistical collaborative filtering algorithm is modified to reflect only recent similarity between two user. The proposed method is evaluated for EachMovie dataset and shows much better prediction results compared with GroupLens system.

**Key Words** : 지능형 디지털 재설계, Takagi-Sugeno 퍼지 시스템, 선형 행렬 부등식, 퍼지 모델 기반 관측기

## 1. 서 론

최근의 인터넷 보급 확대와 정보기술의 발전에 힘입어 인터넷을 기반으로 하는 전자 상거래(e-commerce)의 보급이 급속히 진행되고 있다. 전자 상거래의 대표적인 예로 B2C (business-to-customer)를 들 수 있는데, 이는 인터넷상에서 공급자와 실소비자간에 행해지는 소매형태의 전자 상거래를 의미한다. 이와 같은 B2C환경에서의 실소비자 혹은 고객은 불특정 다수인 경우가 일반적이므로 공급을 담당하는 기업에서는 잘 정리된 방법으로 고객관계를 관리할 필요가 있다.

이러한 고객 관리의 일환으로 기업은 거래에 관련된 고객 정보를 저장하여 관리하는 것이 일반적이다. 저장된 고객 정보를 활용하여 기업 마켓팅에 응용하는 방법 중 대표적인 것으로 추천 시스템(Recommender System)을 들 수 있다. 추천 시스템 또는 협업 필터링(Collaborative Filtering)은 특정한 사용자에게 잠재적으로 흥미가 있거나 가치가 있는 항목을 분류하는 방법이다[1,2]. 유사한 성향을 갖는 사용자는 유사한 형태의 항목을 좋아하리라는 가정 하에, 이 방법은 사용자들의 성향에 관한 데이터베이스를 이용하여 아직 평가되지 못한 항목에 대한 평가값을 예측하는데 사용된다.

한편 일반적으로 사람의 성향은 시간의 흐름에 따라 변할 수 있다. 따라서 시간의 변화에 대한 고려가 없이 전체 데이터베이스를 이용하여 추출된 사용자의 성향은 특정 시점의 사용자의 성향과 다를 수 있으며 이로 인해 그 시점에서의 추천 시스템의 예측률이 현저하게 떨어질 가능성이 있다.

본 논문에서는 사용자 성향의 시간적 변화를 고려한 추천 시스템 알고리즘을 제안하고자 한다. 기존의 알고리즘의 경우 시간적 고려 없이 전체 데이터베이스를 이용하여 사용자 성향을 추출하나, 제시된 방법은 예측하고자 하는 시점의 데이터베이스 정보를 활용하여 사용자 성향을 추출한다. 제시된 방식은 기존의 GroupLens 알고리즘[4]에 적용되었으며,

접수일자 : 2003년 5월 19일
완료일자 : 2003년 10월 9일

기존 방식과 비교하여 높은 예측력을 보여주고 있다.

본 논문의 구성은 다음과 같다. 2장에서 협업 필터링과 GroupLens 알고리즘에 대해서 설명하고, 3장에서는 시간적 성향을 고려한 개선된 알고리즘에 대해서 제안하며, 4장에서는 제안된 방법의 타당성 검증을 위한 모의 실험의 결과를 분석하고 마지막으로 5장에서 결론을 제시한다.

## 2. 협업 필터링(Collaborative Filtering)

추천 시스템에서 많이 사용되는 협업 필터링은 사용자들의 성향에 관한 데이터베이스를 이용하여 아직 평가되지 못한 항목에 대한 특정 사용자의 평가값을 예측하는데 사용된다. 이 방법의 기본 개념은 '유사한 성향을 갖는 사용자는 유사한 형태의 항목을 좋아한다'는 가정을 바탕으로 하고 있다. 즉 어떤 항목에 대해 특정 사용자의 평가값을 예측하고자 할 때, 그 사용자와 유사한 성향을 가지고 있는 다른 사용자들이 그 항목에 대해 어떤 평가값를 나타내었는지를 조사하여 이를 예측하고자 하는 사용자의 평가값로 사용하는 방식이다. 따라서 데이터베이스에는 여러 항목들에 대해 각 사용자별로 작성된 평가값을 포함하고 있어야 한다.

i라는 항목에 대한 사용자 $a$의 평가값을 $v_{a,i}$ 라 하고, 사용자 $a$가 평가값을 평가한 항목들의 집합을 $I_a$ 라 하면, 사용자 $a$에 대한 평균 평가값을 다음과 같이 정의 할 수 있다.

$$\overline{v_a} = \frac{1}{|I_a|}\sum_{i \in I_a} v_{a,i} \tag{1}$$

이제 사용자 $a$가 아직 평가값을 표시하지 않은 항목 $x$ 에 대해 사용자 $a$가 평가내릴 평가값의 추정치를 $P_{a,x}$ 라하자. 협업 필터링에서는 $P_{a,x}$ 를 계산하기 위해 항목 $x$ 에 대해 평가값을 표시한 다른 사용자들의 평가값을 이용한다. 항목 $x$ 에 대해 평가값을 표시한 다른 사용자들의 집합을 B라 하면, 협업 필터링에서는 $P_{a,x}$ 는 다음과 같이 계산된다[3].

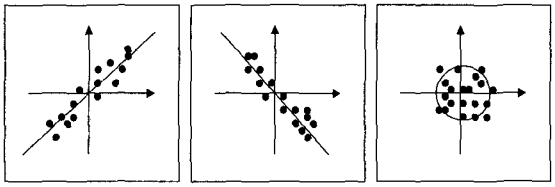
$$p_{a,x} = \overline{v_a} + \frac{\sum_{b \in B} w(a,b)(v_{b,x} - \overline{v_b})}{\sum_{b \in B} w(a,b)} \tag{2}$$

여기서 $w(a,b)$는 사용자 $a$와 $b$간의 유사도를 나타낸다. 이 유사도 $w(a,b)$를 계산하기 위한 방법에 대한 여러 연구 결과들이 있으나 본 논문에서는 상관 계수를 이용한 GroupLens방법에 대해서 다루고자 한다[3][4].

사용자 $a$, b가 공통으로 평가를 내린 항목의 집합을 $I_{a,b}$ 라 하자. GroupLens방식에서는 유사도 $w(a,b)$를 계산하기 위해 통계학에서의 상관계수(Correlation Coefficient)를 이용한다. 즉,

$$w(a,b) = \frac{\sum_{i \in I_{a,b}} (v_{a,i} - \overline{v_a})(v_{b,i} - \overline{v_b})}{\sqrt{\sum_{i \in I_{a,b}} (v_{a,i} - \overline{v_a})^2 \sum_{i \in I_{a,b}} (v_{b,i} - \overline{v_b})^2}} \tag{3}$$

그림. 1에서 알 수 있듯이, 상관관계 계수의 특성상 두 사용자 a ,b가 여러 항목에 대해 유사한 평가값을 보인다면 그 두 사용자는 양의 유사도(+1)을 가지게 되고, 만약 반대의 평가값을 보인다면 음의 유사도(-1)를 가지게 될 것이다. 따라서 식(2)의 협업 필터링 알고리즘은 사용자 $a$와 양의 유사도를 가진 사용자의 평가값은 양으로 참조하고, 음의 유사도를 가진 사용자의 평가값은 음으로 참조하는 방식으로 되어 있다.



(a) w(a,b) ~ +1 (b) w(a,b) ~ -1 (c) w(a,b) ~ 0

그림. 1 상관관계 그래프
Fig. 1 Correlation graph

## 3. 시간적 변화를 고려한 협업 필터링 알고리즘

일반적으로 사람의 성향은 시간이 지나면서 변할 수 있다. 특히 대상 항목이 영화나 음악같이 유행에 민감한 항목이라면, 이에 대한 사용자의 성향도 급격히 변한다고 할 수 있다. 한편, 앞 절에서 언급된 협업 필터링 알고리즘은 시간의 변화에 대한 고려 없이 전체 데이터베이스를 대상으로 하고 있으므로 사용자 성향은 고정되게 평가될 수 밖에 없다. 이 경우 변화하는 사용자 성향에 대한 고려가 되어 있지 못하므로, 유행에 민감한 항목에 대한 평가값의 예측성능이 떨어질 수 있다. 한편, 사용자 성향의 변화는 결국 두 사용자 $a$와 $b$ 사이의 유사도 $w(a,b)$의 변화로 결부될 수 있다. 따라서 본 논문에서는 시간적 성향 변화가 고려된 새로운 유사도 산출방식을 제안한다.

식(3)에서 볼 수 있듯이, 기존의 유사도 $w(a,b)$ 산출 방식은 시간에 대한 고려 없이 전체 데이터베이스에서 사용자 $a$, $b$가 공통으로 평가 내린 항목에 대해 계산하고 있다. 본 논문에서 제시하고자 하는 방법은 전체 평가값중에서 추측하고자 하는 시점에서의 사용자 $a$, $b$가 공통으로 평가 내린 항목만으로 유사도 $w(a,b)$를 산출하는 방식이다. 이 방식은 추측되는 시점에서의 평가값만을 사용하므로 그 시점에서의 사용자 a, $b$의 성향 변화가 고려된 방식이라 할 수 있다.

그러나 사용자 $a$, $b$가 임의의 항목에 대해 평가를 내리는 시점이 서로 다르므로 특정한 시간대에서 사용자 $a$, $b$가 공통으로 평가 내린 항목의 수는 대체적으로 적다. 한편, 유사도 $w(a,b)$는 식(3)에서처럼 통계적인 상관 계수의 계산으로 산출되는데, 만약 대상이 되는 항목의 수가 너무 적다면 적절한 유사도 $w(a,b)$가 적절히 산출되기가 어렵다. 이와 같은 문제점을 해결하기 위해, 유사도 $w(a,b)$를 산출할 때, 특정한 시간대에서 두 사용자중 한명이라도 평가 내린 항목을 포함시키도록 한다. 이를 위해 처음부터 시점 $t$ 까지 사용자 $a$, $b$가 공통으로 평가를 내린 항목의 집합을 $I_{a,b}(t)$라 하고 $t-1$ 에서 $t$ 까지의 시간대에서 두 사용자가 평가 내린 항목의 집합을 다음과 같이 정의하자.

$$S_{a,b}(t) = I_{a,b}(t) - I_{a,b}(t-1) \tag{4}$$

식 (4)를 이용하여 시점 $t$ 에서의 두 사용자간의 유사도 $wt(a,b)$를 다음과 같이 정의한다.

$$w_t(a,b)=\frac{\sum_{i\in S_{a,b}(t)}(v_{a,i}-\overline{v_a})(v_{b,i}-\overline{v_b})}{\sqrt{\sum_{i\in S_{a,b}(t)}(v_{a,i}-\overline{v_a})^2\sum_{i\in S_{a,b}(t)}(v_{b,i}-\overline{v_b})^2}} \quad (5)$$

식(5)에서 정의되는 유사도$wt(a,b)$는 $t$-1 에서 $t$ 까지의 시간대에서 평가된 항목만을 사용하므로, 시간에 따라 변화된 사용자의 성향을 반영하는 방식이라 할 수 있다.

이를 의사 코드(Pseudo code)로 요약하면 다음 그림. 2 와 같이 나타낼 수 있다.

**Pseudo code of temporal weight algorithm**

```
for ( t=0 to t-1 months )
   if (a's item == b's item )
      add the item to I(t-1)
for ( t=0 to t months )
   if (a's item == b's item )
      add the item to I(t)

S(t) = I(t)-I(t-1)
Calculate weight using Eq.(5)
```

그림. 2 제안한 유사도 알고리즘 의사코드
Fig. 2 Pseudo code of proposed weight algorithm

## 4. 결 과

제안된 방식의 검증을 위해 Data-Mining Benchmark Data 로 널리 사용되고 있는 EachMovie Data Set [5]에 제안된 방식을 적용하였다. EachMovie 데이터는 영화에 대한 회원별 평가값이 (05)인 정수형 데이터로서 총 72916명의 1628개 영화에 대한 평가를 기반으로 총 2811983건의 평가 데이터로 이루어져있다. 표. 1과 표. 2에는 회원별 혹은 영화별로 평가값을 평가한 분포를 보여 주고 있다. 비교 검증의 신뢰성을 높이기 위해 200개 이상의 영화를 평가한 회원들 1753명을 대상으로 1000명이상 회원들이 평가한 영화 503개 중 임의의 선정된 회원과 영화를 추출하고 선정된 회원이 특정 영화에 대해 평가한 값을 이용하여 한 달 간격으로 실험을 실시 하였다.

표. 1 회원별 영화평가 분포표
Table. 1 A distribution table of voted movie per user

| 평가한 영화 수 | 해당 회원 수 |
|---|---|
| 100 개 미만 | 65,637 명 |
| 100 개 ~ 199 개 | 5,526 명 |
| 200 개 ~ 299 개 | 1,227 명 |
| 300 개 ~ 399 개 | 346 명 |
| 400 개 ~ 499 개 | 123 명 |
| 500 개 ~ 599 개 | 36 명 |
| 600 개 이상 | 21 명 |

표. 2 영화별 평가회원 분포표
Table. 2 A distribution table of voting user per movie

| 평가한 회원 수 | 해당 영화 수 |
|---|---|
| 100 명 미만 | 471 개 |
| 100 명 ~ 499 명 | 432 개 |
| 500 명 ~ 999 명 | 222 개 |
| 1000 명 ~ 2000 명 | 189 개 |
| 2000 명 이상 | 314 개 |

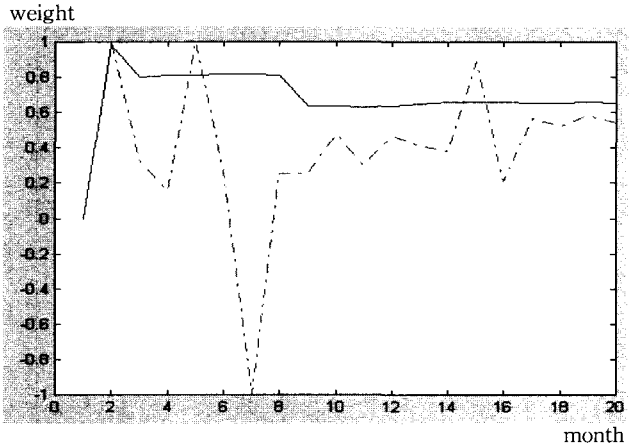



그림. 3 시간 흐름에 따른 상관 관계의 변화
Fig. 3 The variance of correlation with time

그림. 3은 임의의 선정된 두 사용자간의 유사도 $w(a,b)$의 변화를 한달 간격으로 보여주고 있다. 여기서 실선은 GroupLens방식의 산출된 유사도의 시간적 변화를 보여주고 점선은 제안된 방식에 의한 유사도의 변화를 보여주고 있다. 그림. 3에서 볼 수 있듯이 기존의 GroupLens방식에서는 유사도 $w(a,b)$가 고정적으로 되고 있으나, 제안된 방식에서는 시간의 흐름에 따라 성향의 변화가 나타남을 보여주고 있다.

GroupLens방식과 제안된 방식의 비교를 위해 실제 사용자가 평가값과 예측된 평가값을 이용한 식(6)의 MAE(Mean Absolute Error)평가와 시스템 성능 평가 방법인 Precision을 사용하였다.[6,7]

$$E=\frac{\sum|v_{a,x}-p_{a,x}|}{N} \quad (6)$$

$v_{a,x}$ : Actual rating of user a for item x

$p_{a,x}$ : Predicted rating of user a for item x

$N$ : Total number of the item x

표. 3은 MAE를 이용한 평균 오차를 비교한 결과이다. 표. 3에서 GL은 GroupLens방식으로 예측된 결과를 나타낸다. 그리고, GL(R)과 Propsed(R)은 예측치를 반올림하여 오차를 계산한 경우이다. 이는 정수형 원시 데이터의 결과에 대한 신뢰도를 높이기 위해 사용한 오차 계산 방법이다. 표. 3에서 알 수 있듯이 제안된 방법의 평균오차가 GroupLens방식에 비해 현저히 낮으며, 특히 정수형태의 원시 데이터를 고려하여 결과를 반올림한 경우에는 제안된 방식이 약 2배 정도의 더 낮은 평균 오차률을 보이고 있다.

표. 3 평균 오차 비교
Table. 3 The comparison of MAE (GL : Grouplens)

| Algorithm | GL | GL(R) | Proposed | Proposed(R) |
|---|---|---|---|---|
| MAE | 0.810 | 0.800 | 0.647 | 0.467 |

표. 4는 허용오차 ±0.5로 하여 임의의 영화에 대해 회원의 평가값과 예측치간에 일치하는 정도(예측 정확도)를 평가한 결과이다. 표에서 GL(1)과 Proposed(1)은 실제 평가값과 예측 평가값의 허용 오차범위 ±1내에서 측정된 결과이다. 표 5. 에서 알 수 있듯이 제안된 방식의 예측 정확도가 기존의 GroupLens방식에 비해 월등함을 보여주고 있다.

표. 4 예측 정확도에 대한 비교
Table. 4 The comparison of precision

| Algorithm | GL | GL(1) | Proposed | Proposed(1) |
|---|---|---|---|---|
| Precision | 33% | 64% | 67% | 78% |

## 5. 결 론

급속히 확대되고 있는 B2C등의 전자 상거래 환경에서 고객 데이터 등을 활용하여 고객 관리를 하는 CRM(Customer Relationship Management)의 역할이 증대되고 있다. 추천 시스템은 이러한 CRM에서 중추를 이루는 기본 프로세서의 하나로 마켓팅등으로 응용이 확대되고 있다. 본 논문에서는 추천 시스템에서 주로 활용되는 협업 필터링 알고리즘에 대해 논의하고, 사용자 성향의 시간적 변화를 고려한 새로운 방식의 알고리즘을 제안하였다. 제안된 방법은 두 사용자간의 유사성을 나타내는 상관계수의 계산을 변형하여, 사용자 성향의 시간적 변화를 수용하도록 하였다. 제안된 방식은 Data Mining Benchmark Data로 널리 사용되고 있는 EachMovie Data Set에 적용되어 기존의 방법보다 월등한 결과를 보여주고 있다. 이는 적용된 데이터가 유행에 민감한 영화에 관한 것이서, 사용자 성향의 시간적 변화가 크기 때문에 기존의 고정된 사용자 성향을 바탕으로한 협업 필터링 알고리즘으로는 예측력이 떨어진다. 따라서 영화 또는 음악과 같이 사용자 성향의 변화가 큰 항목에 대해서는 본 논문에서 제안된 방법이 보다 더 정확한 예측을 하리라 기대한다.

## 참 고 문 헌


[1] Resnick, P. and Varian, H. , "Recommender systems", *communitions of the ACM*, Pages 56-58, 1997.

[2] J. Ben Schafer, Joseph Konstan, and John Riedl, "Recommender systems in E-commerce", *In Proceedings of the ACM Conference on Electronic Commerce*, Pages 158-166, 1999.

[3] John S. Breese, David Heckerman and Carl Kadie, "Empirical analysis of predictive algorithms for collaborative filtering", *In Proceeding of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, Pages 43-52, 1998.

[4] Resnick, P. and et al., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews", *In Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work*, Pages 175-186, 1994.

[5] http://www.research.digital.com/SRC/eachmovie/

[6] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl, "An Algorithm Framework for Performining Collaborative Filtering", *In Proceedings of the 1999 Conference on the Research and Development in Information Retrieval*, 1999.

[7] Raymond J. Mooney, Loriene Roy, "Content-Based Book Recommending Using Learning for Text Categorization", *In Proceedings of the fifth ACM Conference on ACM 2000 digital libraries*, Pages 195-204, 2000.


## 저 자 소 개


**박영용(Park, Young-Yong)**
2001년 : 세종대학교 컴퓨터공학과 졸업 (학사)
2003년 : 세종대학교 전자공학과 졸업(석사)
2003년~현재 : (주)롯데정보통신 근무

관심분야 : Data Mining, 신경회로망, 퍼지이론
Phone : 011-9733-0729
E-mail : graceyy@orgio.net

**이학성(Lee, Hak-Sung)**
1989년 : 한국과학기술원 전기 및 전자공학과 졸업(학사)
1991년 : 한국과학기술원 전기 및 전자공학과 졸업(공학 석사).
1996년 : 한국과학기술원 전기 및 전자공학과 졸업(공학 박사).
1996년~1998년 : LG 종합 기술원.
1998년~2000년 : LG 이노텍.
현재 : 세종대학교 전자공학과 조교수.

관심분야 : 학습 제어, 지능제어, 로봇 제어