Journal of the Korean Data & Information Science Society
2013, **24**(4), 803–813

http://dx.doi.org/10.7465/jkdi.2013.24.4.803
한국데이터정보과학회지

필터링기법을 이용한 영화 추천시스템 알고리즘 개발에 관한 연구[†]

김선옥[1] · 이수용[2] · 이석준[3] · 이희춘[4] · 지선수[5]

[1]한라대학교 정보통신방송공학부 · [2]연세대학교 교양교직과 · [3]상지대학교 경영정보학과 · [4]상지대학교 컴퓨터데이터정보학과 · [5]강릉원주대학교 정보기술공학과

접수 2013년 5월 29일, 수정 2013년 6월 23일, 게재확정 2013년 7월 7일

요 약

전자상거래에서 상품의 구입은 오프라인에서 구매하는 방식과는 차이가 있다. 오프라인에서 상품 추천은 판매원의 추천에 의해 이루어지지만 온라인에서 상품 추천은 판매원이 상품 추천을 할 수가 없기 때문에 오프라인과는 다른 형태의 상품을 추천하게 된다. 추천시스템은 온라인 상거래에서 상품을 추천하는 방법으로 기존 상품을 구입한 고객의 선호도를 기반으로 상품을 구입하려는 고객의 선호도를 예측하여 추정된 선호도가 높은 상품을 고객에게 추천하는 방법이다. 협력적 필터링 알고리즘은 전자상거래의 상품추천 추천시스템에 사용되며 추정된 값들로 추천 상품 목록을 만들고 그 목록을 고객에게 추천을 하는 것이다. 이 논문에서 사용된 데이터집합은 Movielens 데이터집합인 100k 데이터집합과 1 million 데이터집합이며 일반화를 위해 2개의 데이터집합에서 유사한 결과를 도출하여 일반화시키고자 한다. 영화 추천시스템의 새로운 알고리즘을 제안하기 위해 기존의 알고리즘과 변형된 알고리즘에 의해 추정된 추정값들의 분포 특징을 분석과 응답자별로 분류해서 응답자별 분포의 특징을 분석하였다. 이 논문에서는 이웃기반 추천시스템 협력적 필터링 알고리즘을 개선하기 위해 기존의 알고리즘과 변형된 알고리즘을 바탕으로 새로운 알고리즘을 제안하였다.

주요용어: 상품추천, 전자상거래, 추천시스템, 협력적 필터링.

1. 서 론

인터넷의 대중화와 태블릿PC, 넷북, 스마트폰 등과 같은 디지털 기기의 발전은 고객으로 하여금 다양한 정보를 보다 편리하고 용이하게 접근할 수 있게 하였다. 즉, 고객은 시간과 공간의 제약없이 쉽게 필요로 하는 상품에 대한 정보를 폭넓고 다양하게 찾을 수 있게 되었다. 추천시스템은 고객의 정보를 바탕으로 자동화된 정보필터링 기술을 적용하여 고객이 필요로 하는 상품이나 고객에게 적합한 상품을 추천하는 시스템이다. 특히 전자상거래에서 많이 활용되고 있으며, 국외의 Amazon.com, eBay, CDnow.com, Levis 등과 국내의 쇼핑몰 업체 등에서 널리 적용되고 있다. 이러한 환경은 전자상거래

[†] 이 논문은 2012년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (2011-0009774).
[1] (220-712) 강원도 원주시 흥업면 한라대1길, 한라대학교 정보통신방송공학부, 교수.
[2] (220-710) 강원도 원주시 흥업면 매지리 234, 연세대학교 인문예술대학 교양교직과, 부교수.
[3] (220-702) 강원도 원주시 상지대길 83, 상지대학교 경영정보학과, 조교수.
[4] (220-702) 강원도 원주시 상지대길 83, 상지대학교 컴퓨터데이터정보학과, 교수.
[5] 교신저자: (220-711) 강원도 원주시 남원로 150, 강릉원주대학교 정보기술공학과, 교수.
E-mail: ssji@gwnu.ac.kr

시장을 더욱 활성화 시키는 계기가 되어, 전년대비 2010년 18.4%, 2011년은 17.5%, 2012년 12.7%로 해마다 증가 추이를 보이고 있다. 미래에도 추천시스템의 활용은 증대되리라고 기대된다. 추천시스템의 목적은 고객에게 적절한 추천목록을 제시함으로써 고객의 만족도를 높여 고객충성도를 높이는데 기여하는 시스템이다. 고객에게 추천목록을 제시하려면 상품에 대한 예측 선호도를 추정하고 추정된 예측 선호도에 의해 추정값이 높은 상품으로 추천 목록을 작성하게 된다. 추정된 값은 오차가 작도록 예측 선호도를 추정하는 것이 바람직하다.

2. 연구문제 및 관련연구

2.1. 연구문제

전자상거래 추천시스템은 고객의 선호에 맞는 상품을 추천하여 고객의 만족도를 높이고 고객의 충성도를 높여 기업의 성과를 높이는데 기여할 수 있는 시스템이다. 통계청의 자료에서 알 수 있듯이 전자상거래의 매출액은 해마다 증가 추이를 보이고 있다. 추천 목록을 작성하기 위해 상품 아이템에 대한 정확한 추정값을 계산하는 것은 추천시스템에서 필수 불가결한 것이다. 정확한 추정을 위해서는 좋은 알고리즘이 필요하다. 이 논문에서는 기존의 협력적 필터링 알고리즘의 특성을 분석하고 선형결합 된 새로운 알고리즘을 제안하며 제안 방법의 일반화를 위해 100k 데이터집합, 1 million 데이터집합 두 개의 데이터집합에 적용하여 분석하였다.

2.2. 관련연구

협력적 필터링과 추천시스템에 관한 선행연구를 살펴보면 협력적 필터링의 알고리즘의 성능향상과 관련된 연구로 Lee 등 (2006)의 연구가 있으며 이 연구는 기존의 알고리즘인 이웃기반 알고리즘인 NBCFA (neighborhood based collaborative filtering algorithm)를 분석하고 이웃 고객에 대한 선택기준을 사용하여 새로운 알고리즘인 CMA (correspondence mean algorithm)를 제안하였다. Kim (2010)은 추천시스템에서 최소자승법 (LSM; least square method)을 이용하여 추정값과 평가값의 오차를 최소화시킨 새로운 알고리즘을 제시하였다. 선호도 추정 정확도의 향상을 위해 Linden 등 (2003)은 선호도 상품을 분석하여 추천시스템의 성능향상에 대해 연구하였다. Kim 등 (2012)은 명시적 선호도 평가값만을 이용하는 협력적 필터링 방법에 콘텐츠 정보를 활용하여 영화에 대한 추정 정확도를 높이고자 하였으며, Lee 등 (2011)은 협력적 필터링 기법의 정확도 향상을 위하여 다양한 연령층이 선호하는 영화일 경우 더 많은 사람들이 선호할 것으로 판단하여 고객의 연령의 편차를 반영하여 추정치의 정확도를 높였다. Yu (2012)는 협력적 필터링 기법의 성과 측정을 위한 다양한 방향을 제시하고 있으며 Kim 등 (2011)은 추천시스템에서 정보 탐색시간 측면에서 사용자들의 클릭스트림 데이터의 인접성을 평가하여 상품추천에 이용하였다.

협력적 필터링에 의한 추정결과와 선호도 평가값과의 관계를 분석하여 선호도 추정의 정확도 향상시키기 위해 Lee 등 (2007a)은 응답자 선호도의 Run을 이용하였으며 Lee 등 (2007b)는 협력적 필터링에서 선호도 평가패턴과 예측오차의 관련성을 통하여 알고리즘을 적용하여 선호도를 추정하기 전 예측 오차에 대한 사전 평가 가능성에 대한 연구가 진행되었다. 또한 메모리기반 협력적 필터링 시스템에서 선호도 예측의 특성에 관한 연구가 진행되었다 (Yang 등 2008).

협력적 필터링 알고리즘의 궁극적 목적은 상품을 고객에게 추천하는 것이며 이를 위해 고객의 선호도를 추정하여 이를 이용한 상품 추천목록인 Top N 에 대한 연구가 진행되고 있다. Kim 등 (2010)은 응답자의 응답분포와 추천 순위의 관계에 대하여 연구하였으며 Wu 등 (2012)의 연구에서는 영화 추천 목록의 리스트를 확률적으로 추출하고 이를 평가하기 위한 지표를 개발하여 성능을 비교하였으며, Qinjiao

등 (2012)은 상품에 대한 방문, 구매와 같은 암묵적 선호도를 2진수로 표현하여 이를 추천 리스트 작성에 반영하여 성과를 높이기 위한 연구를 진행하였으며 Yanxiang 등 (2013)은 데이터 희소성에 따른 Top N 추천의 문제를 사용자 기반의 클러스터링 방법을 이용하여 완화시키기 위한 연구를 진행하였다.

2.3. 선호도 예측 알고리즘

협력적 필터링 선호도 추정 알고리즘으로 사용되는 기존의 수식 (2.1)은 이웃 기반 협력적 필터링 알고리즘으로 NBCFA이며 수식 (2.2)는 NBCFA에서 J에 대해 새로운 해석을 적용하여 변형한 알고리즘이다. 수식 (2.3), 수식 (2.4)는 Lee (2006)가 제안한 알고리즘으로 각각 CMA Type 1, CMA Type 2로 발표하였다. 일반적으로 CMA라고 하면 CMA Type 2를 의미한다.

1) 이웃 기반 협력적 필터링 (NBCFA; neighborhood based collaborative filtering algorithm)

수식 (2.1)은 협력적 필터링 알고리즘인 NBCFA이며 수식 (2.2)는 수식 (2.1)에서 $\bar{J}$를 $\bar{J}_{match}$로 대체한 수식이다.

$$\widehat{U}_x = \bar{U} + \frac{\sum\limits_{J \in Raters} \left(J_x - \bar{J}\right) r_{uj}}{\sum\limits_{J \in Raters} |r_{uj}|}, \text{ 여기서 } \bar{J} = \frac{\sum\limits_{i=1}^{n} J_i}{n}, \; i \neq x \tag{2.1}$$

$$\widehat{U}_x = \bar{U} + \frac{\sum\limits_{J \in Raters} \left(J_x - \bar{J}_{match}\right) r_{uj}}{\sum\limits_{J \in Raters} |r_{uj}|} \tag{2.2}$$

여기서, $\widehat{U}_x$는 검정 데이터집합의 x상품에 대한 선호도 예측 대상 고객 U의 선호도 추정값이며, $\bar{U}$는 선호도 예측 대상 고객이 평가한 모든 상품에 대한 선호도 평가값의 평균이다. J_x는 x상품에 대한 이웃 고객의 선호도 평가값이며, 이웃 고객 J는 $\widehat{U}_x$를 계산하기 위한 x상품에 대해 선호도를 평가한 고객들로 구성된다. $\bar{J}$는 이웃 고객 J가 평가한 선호도 평가값의 평균으로 이때 x상품에 대한 선호도 평가값은 제외시킨다. $\bar{J}_{match}$는 고객 u와 이웃 고객 j가 공통으로 표기한 아이템들에 대한 고객 j의 선호도 평균이다. 또한, r_{uj}은 예측 대상 고객 U와 이웃 고객 J의 선호도 유사정도를 나타내는 유사도 가중치이다. 본 연구에서 분류의 정확도를 분석하기 위하여 NBCFA를 이용하여 고객의 선호도를 추정하였으며 유사도 가중치 r_{uj}는 피어슨 상관계수를 이용하였다.

2) 이웃 기반 대응평균 알고리즘 (CMA; correspondence mean algorithm)

협력적 필터링 이웃 기반 알고리즘인 NBCFA 알고리즘에서는 $\bar{U}$는 특정 고객이 나타낸 선호도 전체의 평균을 나타내고 있다. 상관관계를 나타내는 가중치 r_{uj}는 특정 고객 u와 j의 u와 j가 공통으로 평가한 항목으로만 구하게 된다. 여기서 $\bar{U}$는 고객 u의 선호도 전체의 평균을 이용하게 된다. CMA 알고리즘은 $\bar{U}$를 고객 u와 이웃 고객 j가 공통으로 표기한 선호도의 평균들의 평균인 $\bar{U}_{match}$를 이용한다. 고객 u와 이웃 고객 j가 공통으로 표기한 선호도를 이용하여 고객 u의 선호도를 예측하기 때문에 이웃 고객 j의 선호도도 고객 u와 공통으로 표기한 선호도만을 이용하는 것이 필요할 것으로 판단되어 $\bar{J}$를 고객 u와 공통으로 표기한 아이템들에 대한 선호도인 $\bar{J}_{match}$를 이용한다. 수식 (2.3)은 CMA Type

1이며 수식 (2.4)는 CMA Type 2이다.

$$U_x = \overline{U}_{match} + \frac{\sum\limits_{J \in Raters} (J_x - \bar{J}) r_{uj}}{\sum\limits_{J \in Raters} |r_{uj}|}, \text{ 여기서 } \bar{J} = \frac{\sum\limits_{i=1}^{n} J_i}{n}, \ i \neq x \tag{2.3}$$

$$U_x = \overline{U}_{match} + \frac{\sum\limits_{J \in Raters} (J_x - \bar{J}_{match}) r_{uj}}{\sum\limits_{J \in Raters} |r_{uj}|} \tag{2.4}$$

여기서, $\overline{U}_{match}$는 사용자 u와 이웃 고객 j가 공통으로 표기한 아이템에 대한 고객 u의 선호도 평균들의 평균이다. $\bar{J}_{match}$는 고객 u와 이웃 고객 j가 공통으로 표기한 아이템들에 대한 고객 j의 선호도 평균이다. $\bar{J}$는 이웃 고객 J가 평가한 선호도 평가값의 평균으로 이때 x상품에 대한 선호도 평가값은 제외시킨다.

2.4. 예측 정확도 평가척도

협력적 필터링에서 예측의 정확도를 평가하기 위하여 응답 추정값의 절대평균오차 (MAE; mean absolute error)를 이용하여 MAE를 계산한다. MAE가 크면 전체 시스템의 예측 정확도가 낮아지고 MAE가 작으면 예측 정확도가 높아진다. 다음은 MAE의 계산식이다.

$$MAE = \frac{1}{n} \sum_{j=1}^{n} \left| R_{uj} - \widehat{R}_{uj} \right|$$

여기서, R_{uj}는 아이템 j에 대한 고객 u의 실제 선호도 평가값이고, $\widehat{R}_{uj}$는 아이템 j에 대한 고객 u의 선호도 평가값의 추정값이다.

3. 연구방법과 분석절차 및 분석 방법

3.1. 실험 데이터집합의 구성

본 연구는 미네소타 대학의 GroupLens 연구소에서 공개하는 MovieLens 데이터집합에서 100k 데이터집합과 1 million 데이터집합을 이용하여 실험하였다. 100k 데이터집합은 총 943명의 응답자들이 1,682편의 영화에 대해 자신의 선호정도를 5점 척도로 표기한 선호도 평가값으로 구성되어 있고 총 평가값의 수는 100,000개로 구성되어 있으며 각 응답자는 적어도 20개 이상의 영화에 응답을 하였다. 1 million MovieLens 데이터집합은 총 6,040명의 사용자들이 3,952편의 영화에 대해 자신의 선호정도를 5점 척도로 표기한 선호도 평가값들로 구성되어 있으며 총 평가값의 수는 1,000,209개로 구성되어 있다. 본 연구에서는 실험을 위한 분석절차는 다음과 같다.

3.2. 분석절차 및 분석방법

1. 100k 데이터집합과 1 million 데이터집합에 대해 수식 (2.1), (2.2), (2.3), (2.4)으로 선호도 추정값을 계산한다. 선호도 추정 방식은 추정하려고 하는 1개의 값을 제외한 나머지 자료는 훈련 데이터집합으로 선정하여 모든 데이터집합에 대해 선호도를 추정하였다. 즉 100k 데이터집합은 응답자료 1,000,000개에 대해 선호도 추정값을 구하였다. 또한 1 million 데이터집합은 응답자료 10,000,209개에 대해 선호도 추정값을 구하였다. 그러나 이웃 기반의 협력적 필터링 추정방식은 추정을 하고자 하

는 영화에 대해 선호도를 평가한 이웃이 존재하여야 하고 또 이웃 간 유사도 가중치가 계산되어야 하는 조건에 부합되지 않아 추정값을 구할 수 없는 경우가 있어서 100k 데이터집합은 1,000,000개 응답값 중에서 99,859개를 추정할 수 있어 추정 비율은 99.859%로 나타났으며 1 million 데이터집합은 10,000,209개중 10,000,076개를 추정 할 수 있어 추정 비율은 99.987%로 나타났다.

2. 100k 데이터집합과 1 million 데이터집합을 수식 (2.1), (2.2), (2.3), (2.4)으로 추정 계산된 추정값의 분포특징을 분석하고 MAE를 계산하였다.

3. 100k 데이터집합의 응답자 943명을 응답자별로 분류하여 응답분포의 특징과 추정값의 분포 특징을 분석하였으며 1 million 데이터집합의 응답자 6,040명을 응답자별로 분류하여 응답분포의 특징과 추정값의 분포 특징을 분석하였다.

4. 수식 (2.1), (2.2), (2.3), (2.4)로 추정된 100k 데이터집합과 1 million 데이터집합의 추정값을 각각의 응답자 943명, 6,040명을 응답자별로 구분하여 추정값의 오차를 계산하여 과소 추정값과 과대 추정값으로 나누어 이항 검정으로 분석하였다. 이항 검정은 유의수준 α=0.05, 0.01, 0.005, 0.001 로 설정하여 수식 (2.1), (2.2), (2.3), (2.4)의 과소 추정값과 과대 추정값의 분포를 분석하였다.

5. 100k 데이터집합과 1 million 데이터집합에서 추정에 사용된 수식 (2.1),(2.2),(2.3),(2.4)의 특징을 파악하고 새로운 알고리즘에 사용될 수식을 선정하였다.

6. 선정된 수식의 조합에 의해 100k 데이터집합과 1 million 데이터집합에 대해 MAE를 최소로 하는 해를 구하였다.

7. MAE를 최소로 하는 수식에 의해 다시 추정값을 구하여 MAE를 구하고 기존 수식의 MAE와 비교 분석하였다.

4. 실험 및 분석 결과

100k 데이터집합인 경우 수식 (2.1), (2.2), (2.3), (2.4)를 사용하여 100,000개의 평가값에 대해 추정한 결과 추정값을 계산 할 수 없는 경우를 제외한 99,859개를 추정할 수 있으며 추정가능 비율은 99.859%로 나타났다. 또한 1 million 데이터집합인 경우 수식 (2.1), (2.2), (2.3), (2.4)를 사용하여 1,000,209개의 평가값에 대해 추정한 결과 추정값을 계산할 수 없는 경우를 제외한 1,000,076개를 추정할 수 있으며 추정가능 비율은 99.987%로 나타났다.

Table 4.1 Estimation ratio of 100k dataset and 1 million dataset

classification	number of dataset	number of estimation	ratio
MAE of 100k dataset	100,000	99,859	99.859%
MAE of 1 million dataset	1,000,209	1,000,076	99.987%

Table 4.2는 100k 데이터집합, 1 million 데이터집합에서 추정값의 MAE 결과를 나타내고 있다. 결과에서 수식 (2.4)가 가장 작으며 MAE는 각각 0.5819, 0.5863으로 나타났다. MAE가 크게 나타난 것은 수식 (2.2)의 MAE로 각각 0.6351,0.6425로 가장 크게 나타났다. 각 수식은 100k 데이터집합과 1 million 데이터집합에서 동일한 순의 정확도를 보이고 있다.

Table 4.2 MAE of 100k dataset and 1 million dataset

classification	equation (2.1)	equation (2.2)	equation (2.3)	equation (2.4)
MAE of 100k dataset	0.6228	0.6351	0.5870	0.5819
MAE of 1 million dataset	0.6262	0.6425	0.5960	0.5863

Table 4.3은 100k 데이터집합과 1 million 데이터집합의 각각의 추정값 98,859개와 1,000,076개의 추정값에서 (2.1), (2.2), (2.3), (2.4)의 추정 특성을 확인하기 위해 평가값에 대한 과대 및 과소 추정결과를 분류하였다. 결과에서 100k 데이터집합에서 수식 (2.1), (2.2), (2.4)는 오차가 음인 경우가 많이 나타났으며 수식 (2.3)은 양의 오차가 많이 나타났다. 1 million 데이터집합도 100k 데이터집합과 유사한 결과를 보이고 있어 동일 유형의 데이터집합에 대해 각 수식별 과소 및 과대 추정특성이 나타날 가능성을 보이고 있다. 특히 수식별 추정 결과가 매우 유사함을 보이고 있어 각 수식별 추정경향이 안정적임을 알 수 있다.

Table 4.3 Number of minus error and plus error of datasets

classification		equation (2.1)	equation (2.2)	equation (2.3)	equation (2.4)
100k dataset	minus error	53335 (53.4%)	60011 (60.1%)	45423 (45.5%)	52694 (52.8%)
	plus error	46524 (46.6%)	39848 (39.9%)	54436 (54.5%)	47165 (47.2%)
1 million dataset	minus error	535688 (53.6%)	618988 (61.9%)	438517 (43.8%)	529689 (53%)
	plus error	464388 (46.4%)	581288 (38.1%)	561559 (56.2%)	470387 (47%)

Table 4.4는 100k 데이터집합 100,000개의 자료 중 추정 가능한 99,859개의 추정값과 오차를 계산하여 943명의 응답자 중 음의 오차가 더 많이 나타난 응답자와 양의 오차가 더 많이 나타난 응답자를 수식 (2.1), (2.2), (2.3), (2.4)에 따라 분류한 표이다. 100k 데이터집합에서 99,859개의 추정값의 오차를 응답자별로 분류한 결과 수식과 오차의 음, 양의 형태는 관련성이 있는 것으로 나타났다 (p=0.000). 수식 (2.1), (2.2), (2.4)는 음의 오차인 경우가 많이 나타났고 수식 (2.3)은 양의 오차가 많이 나타났다. 따라서 수식 (2.4)는 평가값 보다 과대 추정치가 많으며 수식 (2.3)은 과소 추정치가 많이 나타나고 있다.

Table 4.4 Chi-squaretest of equation and error type in 100k dataset (users 943)

classification	equation (2.1)	equation (2.2)	equation (2.3)	equation (2.4)
minus error	618 (65.54%)	854 (90.56%)	352 (37.33%)	686 (72.75%)
plus error	325 (34.46%)	89 (9.44%)	591 (62.67%)	257 (27.25%)
statistic	χ^2= 622.621, p=0.000			

Table 4.5는 1 million 데이터집합 10,00,209개의 자료 중 추정 가능한 10,000,076개의 추정값과 오차를 계산하여 6,040명의 응답자 중 음의 오차가 더 많이 나타난 응답자와 양의 오차가 더 많이 나타난 응답자를 수식 (2.1), (2.2), (2.3), (2.4)에 따라 분류한 표이다. 1 million 데이터집합에서 10,000,076개의 추정값의 오차를 응답자별로 분류한 결과 수식과 오차의 음, 양의 형태는 관련성이 있는 것으로 나타났다 (p=0.000). 수식 (2.1), (2.2), (2.4)는 음의 오차가 많이 나타났고 수식 (2.3)은 양의 오차가 많이 나타났다. 이는 100k 데이터집합과 유사한 결과를 보이고 있어 수식이 데이터의 규모와 관계없이 일관적인 추정 경향이 있음을 알 수 있다.

Table 4.5 Chi-square test of equation and error type in 1 million dataset (users 6040)

classification	equation (2.1)	equation (2.2)	equation (2.3)	equation (2.4)
minus error	3913 (64.78%)	5732 (94.90%)	1642 (27.19%)	4591 (76.01%)
plus error	2127 (35.22%)	308 (5.10%)	4398 (72.81%)	1449 (23.99%)
statistic	χ^2=6550.214, p=0.000			

Table 4.6은 100k 데이터집합에서 응답자 943명을 응답자별로 구분하여 평가값과 추정값의 오차를 계산한 후 943명의 응답자별로 오차의 음, 양에 대해 이항검정을 실시한 결과이다. 이항검정은 943명 각각에 대해 유의 수준 α=0.05, 0.01, 0.005, 0.001 수준에서 검정하였다. 표에서 보는 것처럼 수식

(2.1), (2.2), (2.4)인 경우에는 오차가 음인 응답자가 많이 나타났으며 수식 (2.3)에서는 오차가 양인 응답자가 많이 나타났다.

Table 4.6 Binomial test of users error in 100k dataset

Sig. level	classification	users of equation (2.1)	users of equation (2.2)	users of equation (2.3)	users of equation (2.4)
α=0.05	minus error>plus error	172(71.37%)	365(99.18%)	28(12.84%)	95(92.23%)
	minus error<plus error	69(28.63%)	3(0.82%)	190(87.16%)	8(7.77%)
α=0.01	minus error>plus error	115(77.70%)	264(99.62%)	13(9.03%)	42(87.5%)
	minus error<plus error	33(22.30%)	1(0.38%)	131(90.97%)	6(12.5%)
α=0.005	minus error>plus error	90(78.95%)	231(99.57%)	11(9.02%)	34(89.47%)
	minus error<plus error	24(21.05%)	1(0.43%)	111(90.98%)	4(10.53%)
α=0.001	minus error>plus error	57(81.43%)	154(99.35%)	6(7.32%)	16(84.21%)
	minus error<plus error	13(18.57%)	1(0.65%)	76(92.68%)	3(15.79%)

Table 4.7은 1 million 데이터집합에서 응답자 6,040명을 응답자별로 구분하여 평가값과 추정값의 오차를 계산한 후 6,040명의 응답자별로 오차의 음, 양에 대해 이항검정을 사용하여 분석하였다. 이항검정은 6,040명 각각에 대해 유의 수준 α=0.05, 0.01, 0.005, 0.001 수준에서 검정하였다. 표에서 보는 것처럼 수식 (2.1), (2.2), (2.4)인 경우에는 오차가 음인 응답자가 많이 나타났으며 수식 (2.3)에서는 오차가 양인 응답자가 많이 나타났다.

Table 4.7 Binomial test of users error in 1 million dataset

Sig. level	users	users of equation (2.1)	users of equation (2.2)	users of equation (2.3)	users of equation (2.4)
α=0.05	minus error>plus error	1421(70.24%)	3739(99.15%)	244(10.22%)	1004(87.46%)
	minus error<plus error	602(29.76%)	32(0.85%)	2144(89.78%)	144(12.54%)
α=0.01	minus error>plus error	966(71.77%)	2804(99.36%)	122(7.11%)	543(88.73%)
	minus error<plus error	380(28.23%)	18(0.64%)	1595(92.89%)	69(11.28%)
α=0.005	minus error>plus error	844(71.77%)	2483(99.52%)	94(6.19%)	433(88.55%)
	minus error<plus error	332(28.23%)	12(0.48%)	1424(93.81%)	56(11.45%)
α=0.001	minus error>plus error	597(71.24%)	1878(99.42%)	52(4.54%)	253(87.54%)
	minus error<plus error	241(28.76%)	11(0.58%)	1093(95.46%)	36(12.46%)

Table 4.6, Table 4.7의 결과에서 100k 데이터집합의 응답자 943명의 오차와 1 million 데이터집합의 응답자 6,040명에 대한 오차의 음, 양값에 대해 이항검정한 결과는 수식 (2.1), (2.2), (2.4)는 음의 오차가 많이 나타났으며 수식 (2.3)에서는 양의 오차가 많은 것으로 나타났다. 이것은 100k 데이터집합, 1million 데이터집합에 대해 수식 (2.1), (2.2), (2.3), (2.4)가 가지고 있는 추정값 계산 방법에 유사성이 있어 동일한 추정경향을 가지고 있다. 또한 이것은 MAE의 크기와 무관하게 오차가 음, 양으로 나타나 수식 (2.1), (2.2), (2.3), (2.4)의 추정경향을 고려하여 새로운 알고리즘 도출 가능성을 가지고 있다.

Table 4.8은 100k 데이터집합, 1 million 데이터집합에 대한 평가값과 추정값을 피어슨 상관계수로 분석한 결과로 상관계수의 크기는 수식 (2.1), (2.2), (2.3), (2.4)의 순으로 나타났다. 수식 (2.1)인 경우 100k 데이터집합, 1 million 데이터집합의 상관계수가 각각 0.712, 0.704로 가장 낮았으며 수식 (2.4)인 경우는 각각 0.760, 0.751로 가장 높게 나타났다. 특이한 점은 MAE가 가장 크게 나타났던 수식 (2.2)가 상관계수에서는 수식 (2.1)보다 더 높게 나타났다. 이것은 Top N 추천에서 N의 크기를 크게 하면 수식 (2.1)에 비해 수식 (2.2)의 Top N 적합률을 안정적으로 유지할 수 있다는 것을 의미하는 것이다.

Table 4.8 Pearson's correlation coefficient of rating value and estimation value

classification	100k dataset	1 million dataset
equation (2.1)	0.712	0.704
equation (2.2)	0.721	0.719
equation (2.3)	0.754	0.742
equation (2.4)	0.760	0.751

이상의 결과로 100k 데이터집합과 1 million 데이터집합에서 수식 (2.3), (2.4)가 MAE가 작으며 평가값과 추정값의 상관계수가 크게 나타났다. 또한 수식 (2.1), (2.2), (2.4)는 오차가 음인 경우가 많이 나타났다. 여기에서 수식 (2.1), (2.2), (2.3), (2.4)을 결합한 형태로 나타날 수 있어서 다음과 같은 결합 알고리즘을 제안한다.

$$\theta = \min\left[\sum_{k=1}^{n} |R_k - f_k(\widehat{R}_i, \widehat{R}_j)|/n\right],\ i,j = 1,2,3,4\ (i \neq j) \tag{4.1}$$

여기서, $f(\widehat{R}_i, \widehat{R}_j) = \theta \cdot \widehat{R}_i + (1-\theta) \cdot \widehat{R}_j, (-\infty \leq \theta \leq \infty)$이고, $\widehat{R}_i$는 수식 (i)의 추정값이며, k는 추정값의 개수이다.

수식 (4.1)에서 θ는 평가값 R과 추정값 $\widehat{R}_i$와 $\widehat{R}_j$에 가중치 θ, $1-\theta$를 곱한 선형결합식과의 절대오차의 평균을 최소로 하는 값을 의미한다.

Table 4.9는 100k 데이터집합에서 수식 (4.1)에 대한 MAE가 최소일 때 θ값의 표이다. 수식 (4.1)에서 θ가 -0.529인 $\widehat{R}_2$과 $\widehat{R}_4$의 선형결합일 때 MAE가 0.5738로 가장 작게 나타났다. 한편 Table 4.2에서 수식 (2.4)일 때 MAE 0.5819에 비해 더 작게 나타났다. 이것은 수식 (2.1),(2.2),(2.3),(2.4)에서 MAE가 가장 작게 나타난 수식 (2.4)에 비해 예측정확도가 더 높다는 것을 의미한다. 즉 $f(\widehat{R}_2, \widehat{R}_4)$가 다른 수식에 비해 MAE를 최소로 하는 알고리즘으로 판단된다.

Table 4.9 Minimum MAE and $\theta$of equation (4.1) in 100k dataset

classification	θ	MAE
$f(\widehat{R}_1, \widehat{R}_2)$	0.943	0.6227
$f(\widehat{R}_1, \widehat{R}_3)$	-0.189	0.5860
$f(\widehat{R}_1, \widehat{R}_4)$	-0.418	0.5777
$f(\widehat{R}_2, \widehat{R}_3)$	0.079	0.5867
$f(\widehat{R}_2, \widehat{R}_4)$	-0.529	0.5738
$f(\widehat{R}_3, \widehat{R}_4)$	0.326	0.5803

Table 4.10은 1 million 데이터집합에서 수식 (4.1)에 대한 MAE가 최소일 때 θ값의 표이다.

수식 (4.1)에서 θ가 -0.472인 $\widehat{R}_2$과 $\widehat{R}_4$의 선형결합일 때 MAE가 0.5793로 가장 작게 나타났다. 한편 Table 4.2에서 수식 (2.4)일 때 MAE 0.5863에 비해 더 작게 나타났다. 이것은 수식 (2.1), (2.2), (2.3), (2.4)에서 MAE가 가장 작게 나타난 수식 (2.4)에 비해 예측정확도가 더 높다는 것을 의미한다. 즉 $f(\widehat{R}_2, \widehat{R}_4)$가 다른 수식에 비해 MAE를 최소로 하는 알고리즘으로 분석되었다.

Table 4.10 Minimum MAE and $\theta$of equation (4.1) in 1 million dataset

classification	θ	MAE of $f(R_i, R_j)\|\theta$
$f(\widehat{R}_1, \widehat{R}_2)$	0.927	0.6261
$f(\widehat{R}_1, \widehat{R}_3)$	-0.020	0.5959
$f(\widehat{R}_1, \widehat{R}_4)$	-0.388	0.5827
$f(\widehat{R}_2, \widehat{R}_3)$	0.183	0.5933
$f(\widehat{R}_2, \widehat{R}_4)$	-0.472	0.5793
$f(\widehat{R}_3, \widehat{R}_4)$	0.269	0.5849

Table 4.11은 100k 데이터집합, 1 million 데이터집합에서 각각 θ=-0.529, θ=-0.472 일 때 $f(R_2, R_4)$의 추정오차의 절대치와 수식 (2.4)의 추정오차의 절대치를 짝을 이룬 t 검정한 결과이다. 검정 결과 100k 데이터집합, 1 million 데이터집합에서 모두 통계적으로 유의적인 결과를 얻었다. 이것은 기존의 알고리즘에 비해 MAE가 작게 나타나 예측 정확도를 향상시키는 알고리즘으로 사용할 수 있다.

Table 4.11 Paired t-test of equation (2.4) between $f(R_2, R_4)|\theta$in 100k dataset and 1 million dataset

classification	absolute errorof equation (2.4)	absolute errorof $f(R_2, R_4)\|\theta$	t value	p value
	mean (s.d)	mean (s.d)		
100k dataset (θ=-0.529)	.5819 (.4712)	.5738 (.4795)	21.949	.000
1 million dataset (θ=-0.472)	.5863(.4695)	.5793 (.4779)	65.759	.000

5. 결 론

Movielens 데이터집합을 기반으로 하여 협력적 필터링 알고리즘의 비교 결과 100k 데이터집합, 1 million 데이터집합에서 모두 수식 (2.4)의 MAE가 가장 작게 나타났다. 이것은 기존의 연구결과와 일치한다. 따라서 이 논문에서는 MAE를 작게 하는 알고리즘을 제안하기 위하여 이웃기반 협력적 필터링 알고리즘인 수식 (2.1), (2.2), (2.3), (2.4)의 MAE와 오차를 분석하여 기존 알고리즘에 가중치를 고려하여 선형결합식으로 표현할 수 있는지 분석하였다. 실험결과 제안한 선형결합식은 100k 데이터집합, 1 million 데이터집합에서 모두 수식 (2.4)보다 MAE가 작은 것으로 나타났으며 제안 선형결합식의 오차와 수식 (2.4)의 오차를 비교한 결과 통계적으로 유의한 차이가 있음을 확인하였다. 이 결과는 영화의 선호도 예측에서 예측 정확도가 더 향상되었음을 의미하며 기존의 협력적 필터링 알고리즘에 비해 더 우수한 알고리즘이라 할 수 있다. 또한 수식 (2.1), (2.2), (2.3), (2.4)의 추정값과 선호도 평가값과의 상관관계를 분석한 결과 수식 (2.4)에 의한 추정값과의 상관관계가 가장 높게 나타났다. 추정값과 평가값의 상관관계가 높음은 상품추천 Top N 의 적합률을 높일 수 있고 또한 Top N 추천에서 N의 크기가 변하더라도 안정적인 Top N 적합률을 유지할 수 있음을 의미한다. 본 연구에서 제안한 선형결합식은 기존 연구결과에서 성능이 가장 좋은 수식 (2.4)의 추정특성을 보완하여 더 우수한 결과를 유도하고 있고 수식 (2.4)를 활용하기 때문에 추정값과 평가값과의 상관관계를 더 높일 수 있을 것으로 기대되며 이를 위한 추후 연구로 개선된 알고리즘에 의해 영화 추천시스템의 추천목록을 작성하는 Top N 적합율을 높이는 연구와 Top N에서 N의 크기와 무관하도록 Top N 적합률을 유지하는 연구가 필요할 것으로 생각된다.

References

Kim. J. H. and Byeon. H. S. (2011). A product recommendation system based on adjacency data. *Journal of the Korean Data & Information Science Society*, **22**, 19-27.

Kim. S. H., Oh. B. H., Kim. M. J. and Yang. J. H. (2012). A movie recommendation algorithm combining collaborative filtering and content information. *Journal of Korean Institute of Information Scientists and Engineers: Software and Applications*, **39**, 261-268.

Kim, S. O. (2010). The research of new algorithm to improve prediction accuracy of recommender system in electronic commerce. *Journal of Korean Data & Information Science Society*, **21**, 185-194.

Kim. S. O. and Lee H. C. (2010). A study of distribution of response and rank of recommendation in collaborative filtering. *Journal of the Korean Data Analysis Society*, **12**, 2071-2080.

Lee, H. C. and Lee. S. J. (2006). On the precision of the prediction of the nearest neighbor algorithm and adjusted algorithm for user-based recommender system. *Journal of the Korean Data Analysis Society*, **8**, 1893-1904.

Lee, S. H. and Park, S. H. (2011). Accuracy improvement of a collaborative filtering recommender system using attribute of age. *Journal of the Korea Safety Management & Science*, **13**, 169-177.

Lee. S. J., Kim. S. O. and Lee H. C. (2007a). The relationship of prediction accuracy and the run of abnormal users' ratings in collaborative filtering. *Journal of the Korean Data Analysis Society*, **9**, 2043-2054.

Lee, S. J., Kim, S. O. and Lee, H. C. (2007b). A study on the interrelationship between the prediction error and the rating's pattern in collaborative. *Journal of Korean Data & Information Science*, **18**, 659-668.

Linden, G., Smith, B. and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, **7**, 76-80.

Qinjiao, M., Boqin, F. and Shanliang, P. (2012). A study of top-n recommendation on user behavior data. *2012 IEEE International Conference on Computer Science and Automation Engineering*, 2012 International Conference, 25-27 May, 582-586.

Wu, Q., Li, L., Li, H., Tang, F., Barolli, L. amd Luo. Y. (2012). Recommendation of more interests based on collaborative filtering, *2012 IEEE 26th International Conference on Advanced Information Networking and Applications*, 2012 International Conference, 26-29 March, 191-198.

Yang, G. M., Lee, H. C. and Park, Y. S. (2008). The feature of preference prediction by memory-based collaborative filtering algorithm. *Journal of the Korean Data Analysis Society*, **10**, 591-601.

Yanxiang, L., Deke, G., Fei, C. and Honghui, C. (2013). User-based clustering with top-n recommendation on cold-start problem. *2013 Third International Conference on Intelligent System Design and Engineering Applications*, 2013 International Conference, 16-18 Jan, 1585-1589.

Yu, S. J. (2012). A comprehensive performance evaluation in collaborative filtering. *Journal of the Korea Society of Computer and Information*, **17**, 83-90.

Journal of the Korean Data & Information Science Society
2013, **24**(4), 803–813

http://dx.doi.org/10.7465/jkdi.2013.24.4.803
한국데이터정보과학회지

A study of development for movie recommendation system algorithm using filtering[†]

Sun Ok Kim[1] · Soo Yong Lee[2] · Seok Jun Lee[3] · Hee Choon Lee[4] · Seon Su Ji[5]

[1]School of Information Communication & Broadcasting Engineering, Halla University
[2]College of Humanities & Arts, Yonsei University
[3]Department of MIS, Sangji University
[4]Department of Computer Data & Information, Sangji University
[5]Department of Information Technology & Engineering, Gangneung-Wonju National University
Received 29 May 2013, revised 23 June 2013, accepted 7 July 2013

Abstract

The purchase of items in e-commerce is a little bit different from that of items in off-line. The recommendation of items in off-line is conducted by salespersons' recommendation, However, the item recommendation in e-commerce cannot be recommended by salespersons, and so different types of methods can be recommended in e-commerce. Recommender system is a method which recommends items in e-commerce. Preferences of customers who want to purchase new items can be predicted by the preferences of customers purchasing existing items. In the recommender system, the items with estimated high preferences can be recommended to customers. The algorithm of collaborative filtering is used in recommender system of e-commerce, and the list of recommended items is made by estimated values, and then the list is recommended to customers. The dataset used in this research are 100k dataset and 1 million dataset in Movielens dataset. Similar results in two dataset are deducted for generalization. To suggest a new algorithm, distribution features of estimated values are analyzed by the existing algorithm and transformed algorithm. In addition, respondent'distribution features are analyzed respectively. To improve the collaborative filtering algorithm in neighborhood recommender system, a new algorithm method is suggested on the basis of existing algorithm and transformed algorithm.

Keywords: Collaborative filtering, e-commerce, items recommendation, recommender system.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0009774).
[1] Professor, School of Information Communication & Broadcasting Engineering, Halla University, Gangwon-do 220-712, Korea.
[2] Associate professor, College of Humanities & Arts, Yonsei University, Gangwon-do 220-701, Korea.
[3] Assistant professor, Department of MIS, Sangji University, Gangwon-do 220-702, Korea.
[4] Professor, Department of Computer Data & Information, Sangji University, Gangwon-do 220-7-2, Korea.
[5] Corresponding author: Professor, Department of Information Technology & Engineering, Gangneung-Wonju National University, Gangwon-do 220-711, Korea. E-mail: ssji@gwnu.ac.kr