공학석사 학위논문

사용자들의 공통관람 영화 유사도를 이용한 추천시스템의 성능향상

Improvement of Recommender System by common rated movie similarity of users

2014 년 7 월

서울대학교 대학원

산업공학과 데이터마이닝 전공

구 윤 성

사용자들의 공통관람 영화 유사도를 이용한 추천시스템의 성능향상

Improvement of Recommender System by common rated movie similarity of users

지도 교수 조 성 준

이 논문을 공학석사 학위논문으로 제출함

2014 년 7 월

서울대학교 대학원

산업공학과 데이터마이닝 전공

구 윤 성

구윤성의 공학석사 학위논문을 인준함

2014 년 7 월

위 원 장 박 종 헌 (인)

부위원장 조 성 준 (인)

위 원 이 재 욱 (인)

초 록

추천 시스템은 전세계적으로 많이 쓰이고 있고 중요한 분야이다. 영화, 음악 또는 책 등의 품목이 점점 많아지고 인터넷이 발달함에 따라 사용자들은 더욱 더 많은 콘텐츠들에 접근할 수 있다. 하지만 실제로 소비할 수 있는 콘텐츠는 한정적이다. 따라서 소비자들은 본인의 취향에 적합한 콘텐츠를 고르고 싶어한다. 때문에 추천시스템은 중요하고 유용하게 사용될 수 있다. 실제로 Amazon 이나 Netflix등의 서비스는 이미 이러한 시스템을 도입하여 소비자들이 직접 콘텐츠를 찾지 않아도 개별 사용자에게 알맞은 콘텐츠들을 추천해준다.

협력 필터링은 추천시스템에 가장 널리 쓰이고 가장 많이 연구되는 방법중의 하나이다. 그 중 사용자기반 협력 필터링은 특정 사용자가 흥미로워 할 수 있는 아이템을 찾기 위해 그와 유사한 사용자들을 찾아 그들의 정보를 이용하여 추천할 아이템을 찾는다. 이 때, 사용자들의 유사도를 구하는 것은 매우 중요한 과정이다. 본 연구에서는 사용자들의 평점을 이용하여 영화를 추천할 때 사용자들의 유사도를 구하는 새로운 방법을 제안한다.

먼저 사용자들의 평가가 있는 영화들은 사용자들이 과거에 선택해서 관람한 영화들이다. 즉 이 영화들은 사용자들이 자신의 취향에 적합할 것이라 생각하고 선택하였기 때문에 선택하지 않은 영화에 비해 중요하다. 따라서 두 사용자가 얼만큼의 영화를 공통으로 선택하여 평가하였는지를 고려하였다. 또한 공통으로 선택한 영화에 대하여 사용자들이 얼마나 만족하였는지 고려하였다. 사용자가 선택한 항목들의 평균보다 만족스러웠는지, 불만족스러웠는지 두 사용자간의 이러한 편향성이 일치하는지 고려하여 유사도를 계산하였다.

본 연구에서 제안한 유사도를 이용하여 협력 필터링을 수행한 결과 기존의 유사도를 구하는 방법인 피어슨 상관관계, 자카드 유사도보다 더 높은 예측력을 보이는 것을 알 수 있었다. 또한 적은 이웃만을 고려하여 예측했을 때에도 성능이 좋은 것을 확인하였다. 제안한 방법을 사용했을 때, 예측오차 값의 평균 MAE(Mean Absolue Error) 값은 0.7929로 기존의 유사도 보다 적은 값을 보였다.

주요어 : 협력 필터링, 사용자 유사도, 공통 평가, 편향성
학 번 : 2012-23307

목 차

표 목차

그림 목차

제 1 장 서 론

영화산업이 발전함에 따라 매년 영화의 개봉 수는 증가하고 그에 따라 누적되는 영화의 수는 폭발적으로 증가하고 있다. 최근 영화의 개봉편수는 아래의 [표 1] 과 같다(2012 한국 영화산업 결산).

구분	한국영화		외국영화		총 개봉편수
	제작편수	개봉편수	수입편수	개봉편수	
2004	82	74	285	194	268
2005	87	83	253	215	298
2006	110	108	289	237	345
2007	124	112	404	280	392
2008	113	108	360	272	380
2009	138	118	311	243	361
2010	152	140	383	286	426
2011	216	150	551	289	439
2012	229	175	773	456	631

[표 1] : 연도별 한국,외국영화 제작,개봉편수

이렇게 점점 많은 영화들이 개봉되고 또한 누적되고 있기 때문에, 사람들은 자신의 취향에 맞는 영화를 찾는 것이 쉽지 않다. 따라서 추천시스템을 이용하여 사용자의 취향에 맞는 영화를 찾아낸다면 매우 유용하게 이용될 수 있다.

실제로 영화 추천 시스템은 많이 사용되고 있다. 한국에도 영화 추천시스템을 이용한 서비스를 제공하는 업체가 있다. 대표적으로 '왓챠'에서 추천시스템을 이용하여 사용자들에게 영화를 추천해주고 있다.

[그림 1] : 영화추천 서비스 왓챠

영화 추천 서비스 '왓챠'는 컴퓨터나 스마트 폰을 통해서 사용할 수 있다. [그림 1] 처럼 해당 서비스는 사용자의 평점을 예측하기도 하고 사용자가 좋아할 영화를 추천해주기도 한다.

이렇게 사용자들에게 영화, 음악 등의 아이템들을 찾아 추천해주기 위해 협력 필터링(Collaborative Filtering)이 연구되어 왔다. 협력 필터링은 추천 시스템을 위해 가장 많이 연구되어온 방법 중 하나이다. 협력 필터링은 사용자들의 정보를 이용하여 그 사용자들에게 흥미있는 정보, 즉 영화나 음악 등의 아이템을 찾아낸다 (Wang et al. 2008).

콘텐츠 기반의 방식이 영화나 음악 등의 콘텐츠 자체의 정보를 이용하여 추천하는데 비해, 협력 필터링은 사용자가 영화나 음악 등의 아이템에 대하여 평가한 정보를 이용하여 추천해주는 시스템이다.

[그림 2] : 협력 필터링 이웃선정

[그림 2]을 보면 협력 필터링 기법을 간략하게 설명하고 있다. 먼저 사용자들의 평가정보를 이용하여 사용자들간의 유사도를 계산한다. 많은 항목들 중에서 사용자들이 공통적으로 평가를 한 항목을 바탕으로 그 항목들에 대한 평가가 얼마나 유사한지 계산한다. 이후 계산된 유사도를 이용하여 해당 사용자와 유사한 사용자들로 이웃을 선정한다. 즉, 동일한 아이템에 대하여 비슷한 평가를 한 사용자들을 찾아내어 해당 사용자의 이웃으로 선정한다. 이후 선정한 이웃들의 평점 정보를 바탕으로 해당 사용자의 평점을 예측하게 된다.

협력 필터링은 사용자기반, 아이템기반 두가지 방법으로 나뉘게 된다. 예측하고자 하는 평점에 대하여 해당 아이템과 유사한 아이템들의 정보를 바탕으로 예측하는 아이템기반 협력 필터링과 해당 사용자와 유사한 사용자들을 바탕으로 예측하는 사용자 기반 협력 필터링으로 나뉜다(Wang et al. 2006). 본 연구에서는 사용자기반 협력 필터링에 집중하였다.

이렇게 협력 필터링에서 유사한 사용자들을 찾기 위해서는 각각의 사용자들의 평점을 이용한다. 각 사용자들의 평점간 피어슨 상관관계, 코사인 등을 이용하여 두 사용자의 평점이 얼마나 유사한지 계산하게 된다. 두 사용자가 모두 평가한 항목에 대하여 같은 항목에 대해 얼마나 비슷한 평가를 하는지를 보는 방법이다. 하지만 실제 데이터를 살펴보면 사용자가 평가를 하는 항목은 전체 평가할 수 있는 항목 중 극히 일부에 지나지 않는다.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	5	3	4	3	3	5	4	1	5	3
2	4	0	0	0	0	0	0	0	0	2
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	4	3	0	0	0	0	0	0	0	0
6	4	0	0	0	0	0	2	4	4	0
7	0	0	0	5	0	0	5	5	5	4
8	0	0	0	0	0	0	3	0	0	0
9	0	0	0	0	0	5	4	0	0	0
10	4	0	0	4	0	0	4	0	4	0
11	0	0	0	0	0	0	0	4	5	0
12	0	0	0	5	0	0	0	0	0	0
13	3	3	0	5	1	0	2	0	3	0
14	0	0	0	0	0	0	5	0	4	0

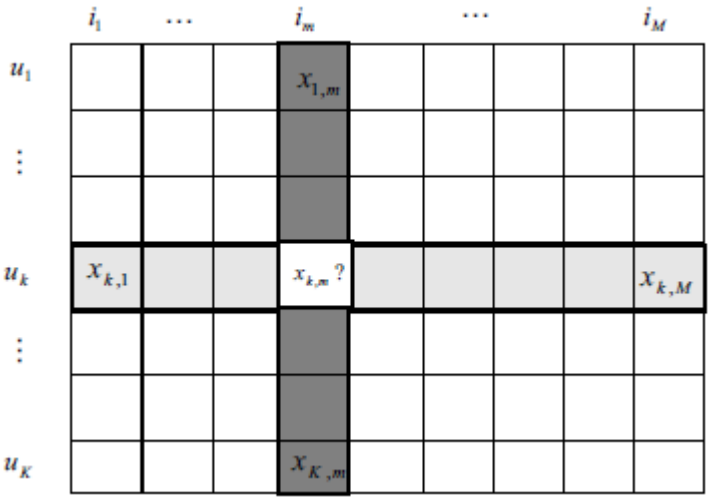
[그림 3] : 사용자-영화 행렬

위 [그림 3]는 실제 사용한 사용자-영화 행렬 중 일부분을 보여주고 있다. 1번 행을 보면 해당 사용자는 1부터 10에 해당하는 영화를 모두 보고 평가를 하였다. 반면 5번행의 경우 1부터 10까지의 영화 중 실제 평가를 한 영화는 1,2번 둘 밖에 존재하지 않는다. 따라서 위 행렬만 살펴볼 경우 1번 사용자와 5번 사용자의 유사도는 1,2번 영화의 평점을 이용하여 계산하게 된다.

본 연구에서는 협력 필터링 중, 사용자 기반 협력 필터링을 사용하여 유사도에 대해 논하였다. 협력 필터링을 수행하는 과정에서 특정 사용자와 유사한 사용자를 찾을 때 계산하는 새로운 유사도를 제안하였다. 특정 사용자와 유사한 사용자를 찾는데 있어서 공통으로 평가한 항목들이 얼마나 많은지를 이용하고 또한 그 항목들에 대하여 각각의 사용자들의 편향성이 비슷한지를 이용하여 유사도를 계산한다. 이를 이용하여 이웃을 선정한 뒤, 그 이웃들을 이용하여 평점을 예측하고 평가한다.

제 2 장 기존 연구

사용자 기반 협령 필터링을 수행하려면 먼저 사용자들과 그들의 영화 평점을 이용해 사용자−영화 행렬을 만든다.



[그림 4] : 사용자−영화 행렬

[그림 4]와 같이 사용자들의 평점을 이용해 사용자−영화 행렬을 만들게 된다. 이 때, 각각의 행은 사용자들을 나타내고 각각의 열은 영화를 나타내게 된다. 즉 k 번째 행의 m번째 열은, 사용자 k가 영화 m에 대해 평가한 점수를 나타내게 된다. 즉 $x_{k,m} = r$ 은 사용자 k 가 영화 m에 평점 r을 주었다는 것을 뜻하고 $x_{k,m} = \emptyset$ 은 해당 평점을 알지 못하는 것을 뜻한다.

위와 같이 사용자−영화 행렬 X가 주어졌을 때, 해당 행렬은 행벡터들로 분해될 수 있다.

$$X = [u_1, u_2, \dots, u_K]^T,\ u_k = [x_{k,1}, x_{k,2}, \dots, x_{k,M}]^T, k = 1, \dots, K$$

사용자−영화 행렬을 행벡터로 분해하면 각각의 행 u_k^T 는 사용자 k 의 모든 영화들에 대한 평점들에 대한 벡터이다. 따라서 해당 행벡터들의 분석을 통하여 사용자 기반 협력 필터링을 이용할 수 있다. 협력 필터링은 이처럼 사용자들의 평점만을 이용하기 때문

에 특정 배경지식이나 사용자와 영화에 대한 광범위한 정보를 필요로 하지 않는다(Hao Ji et al. 2013).

이렇게 행벡터로 분해한 후에 각각의 행을 이용하여 사용자들의 유사도를 구하게 된다. 유사도의 계산은 협력 필터링에서 가장 중요한 부분이다(Liu et al. 2014).

유사도	정의
피어슨 상관관계	$\mathrm{Sim}(u,u') = \frac{\sum_{i\in I(u)\cap I(u')}(x_{u,i} - \overline{x_u}) \cdot (x_{u',i} - \overline{x_{u'}})}{\sqrt{\sum_{i\in I(u)\cap I(u')}(x_{u,i} - \overline{x_u})^2} \cdot \sqrt{\sum_{i\in I(u)\cap I(u')}(x_{u',i} - \overline{x_{u'}})^2}}$
코사인	$\mathrm{Sim}(u,u') = \frac{\sum_{i\in I(u)\cap I(u')} x_{u,i}x_{u',i}}{\sqrt{\sum_{i\in I(u)\cap I(u')} x_{u,i}^2} \cdot \sqrt{\sum_{i\in I(u)\cap I(u')} x_{u',i}^2}}$
자카드	$\mathrm{Sim}(u,u') = \frac{\lvert I(u)\cap I(u')\rvert}{\lvert I(u)\cup I(u')\rvert}$

[표 2] : 협력 필터링에서의 사용자 유사도

협력 필터링은 먼저 [표 2]와 같은 유사도를 이용해 사용자의 이웃을 선정한다.

$$P(r_{u,i}) = \overline{r_u} + \frac{\sum_{u_a\in S(u)} Sim(u_a, u)\cdot(r_{u_a,i} - \overline{r_{u_a}})}{\sum_{u_a\in S(u)} Sim(u_a, u)}$$

$S(u)$: 사용자 u의 이웃

이웃을 선정하면 위와 같은 식을 이용하여 사용자의 평점을 예측한다. 즉 특정 사용자와 유사한 이웃들을 찾고, 해당 사용자는 보지 않았지만 이웃들이 보고 평가한 영화들의 평점을 이용하여 해당 사용자가 특정 영화를 보았을 때 어떤 평점을 주게 될 지 예측하게 된다. 어떠한 영화에 대하여 사용자의 평점은 없지만 이웃들의 그 영화에 대한 평점 경향이 평균보다 높게 나타나는지 낮게 나타나는지 살펴보고 해당 사용자와 유사한 이웃들이 해당 영화에 대해 본인들의 평균보다 높게 평점을 준다면 높은 평점을, 낮게 평점을

주면 낮은 평점을 예측한다.

이웃들의 평점을 이용할 때, 각각의 이웃에 대하여 유사도 만큼 예측에 영향을 준다. 즉 특정 사용자와 더욱 유사한 사용자의 평점이 다른 사용자들의 평점보다 예측에 많은 영향을 준다(Wang et al. 2006).

이와 같이 협력 필터링은 [표 2]와 같은 유사도를 이용하여 이웃을 선정하고 이웃들의 평점정보를 이용해 평점을 예측하게 된다. 본 연구에서는 협력 필터링을 이용할 때 기존의 유사도와 제안하는 유사도를 이용하여 평점을 예측, 비교 분석하였다.

제 1 절 피어슨 상관관계

피어슨 상관관계(Pearson Correlation)는 가장 많이 쓰이는 상관관계 계수이다. 피어슨 상관관계는 표준화된 두 변수가 공변하는 관계를 나타내는 통계량이다. 두 변수가 서로 얼마나 선형적인 관계인지 즉 하나의 변수가 증가할 때, 나머지 변수가 어떻게 변화하는지를 살펴보는 변수이다. 피어슨 상관관계를 구하는 식은 다음과 같다.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

이 때 피어슨 상관관계는 −1에서 1 사이의 값을 갖는다. 1이 갖는 의미는 두 변수가 완전히 선형적인 양의 상관관계에 있다는 것이고, −1은 완전히 선형적인 음의 상관관계를 갖는 다는 것을 의미한다.

따라서 위 협력 필터링의 유사도를 구하는데 있어서 사용자가 공통적으로 본 영화에 대하여 양의 상관관계가 클수록 유사도는 큰 값을 갖는다. 따라서 같은 영화에 대하여 높은평점를 주거나 낮은 점수를 준 사용자들간의 유사도 값이 크고 반대로 한 사용자가 높은 평점을 준 영화에 대하여 낮은 평점을 준 사용자는 유사도 값이 작게 나타나게 된다.

사용자−영화 행렬에서 각각의 행벡터는 각각 한명의 사용자의 평점들을 나타내게 된다. 이 때, 벡터의 차원은 총 영화의 개수가 되는데 이 영화의 수는 굉장히 많다. 즉 전체 벡터 중 실제 평점값을 갖고 있는 성분의 수는 극히 한정이 된다. 따라서 두 사용자의 유사도를 피어슨 상관관계를 사용하여 계산할 때, 전체 영화가 아닌 두 사용자가 모두 평가를 한 영화에 대해서만 고려하게 된다.

	영화1	영화2	영화3	영화4	영화5	영화6
사용자1		3		5	2	2
사용자2	1			3	3	

[표 3] : 사용자 평점 예시

[표 3]은 두명의 사용자가 6개의 영화에 대해 평가를 하고 평점을 매긴 예시이다. 6개의 영화가 주어졌지만 사용자들이 모든 영화에 대해서 보고 평가를 하진 않는다. 따라서 이런 사용자 평가점수의 피어슨 상관관계를 구하기 위해서는 모든 영화에 대해서 상관관계를 구하는 것이 아니라 두 사용자가 모두 평가한 항목에 대해서만 고려하게 된다. 즉 영화 4,5 만을 고려하여 상관관계를 계산하게 된다.

즉 임의의 사용자 u, u' 간의 피어슨 상관관계를 이용한 유사도 $\mathrm{Sim_{pc}}(u, u')$ 는 다음과 같이 정의된다.

$$\mathrm{Sim_{pc}}(u, u') = \frac{\sum_{i \in I(u) \cap I(u')} (x_{u,i} - \overline{x_u}) \cdot (x_{u',i} - \overline{x_{u'}})}{\sqrt{\sum_{i \in I(u) \cap I(u')} (x_{u,i} - \overline{x_u})^2} \cdot \sqrt{\sum_{i \in I(u) \cap I(u')} (x_{u',i} - \overline{x_{u'}})^2}}$$

여기서 $I(u)$ 는 사용자 u가 평점을 준 영화의 집합, $\overline{x_u}$ 는 사용자 u가 준 평점들의 평균 값을 나타낸다.

제 2 절 자카드 유사도

자카드 유사도는 피어슨 상관관계와는 달리 평가한 영화들의 점수를 이용하지 않고 어떠한 영화들에 평점을 주었는지를 고려하게 된다.

$$\mathrm{Sim_{jac}(u,u')} = \frac{|I(u) \cap I(u')|}{|I(u) \cup I(u')|}$$

$I(u) = \{\, i \,|r_{u,i} \neq \emptyset\}$

자카드 유사도는 위와 같이 정의 된다. 즉 두 사용자 u, u'에 대하여 둘이 공통으로 평가한 영화의 수를 둘이 평가한 모든 영화의 수로 나누어 주면 자카드 유사도가 계산된다. 자카드 유사도는 두 사용자의 유사도를 계산할 때, 두 사용자가 평가한 영화에 대해서 평점에 관계없이 둘 모두 평가한 영화의 비율이 두 사용자가 평가한 모든 영화에 대해 얼마나 되는지 보는 유사도이다.

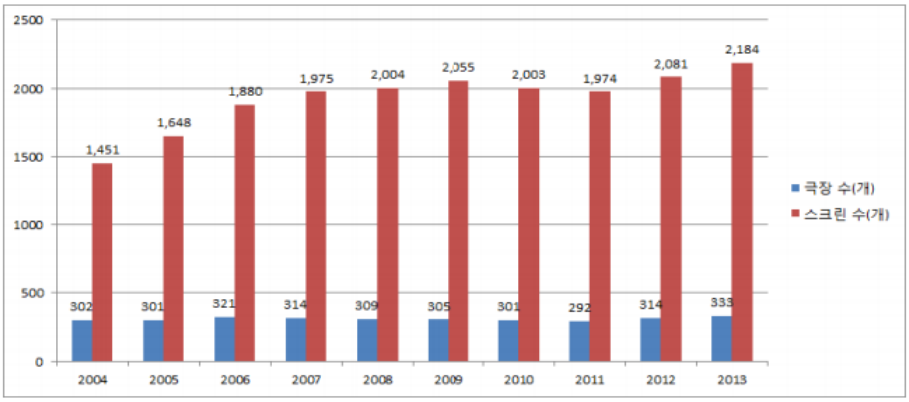
자카드 유사도는 피어슨 상관관계나 코사인을 이용한 유사도와는 다르게, 굉장히 적은 영화에 대해서만 공통으로 평가한 사용자들의 유사도를 높게 평가하지 않기 위해서 공통으로 평가한 영화의 비율이 얼마나 되는지를 고려하는 유사도이다(J. Bobadilla et al. 2010). 따라서 자카드 유사도와 본 연구에서 제안하는 유사도를 비교하여 분석하고자 한다.

제 3 장 연구 방법

협력 필터링에서의 유사도는 계속 연구되어 왔지만 여전히 단점들을 가지고 있다(Liu et al. 2014). 본 장에서는 이번 연구에서 제안된 사용자 유사도 계산과정을 기술한다. 본 연구에서 제안한 유사도는 두 사용자가 공통으로 평가를 한 항목이 얼마나 되는지와 공통으로 평가한 영화에 대해서 각각의 사용자의 편향성이 일치하는지 여부를 가지고 유사도를 계산한다. 제 1 절에서는 공통으로 평가한 영화의 개수를 이용하는 방법을 기술한다. 제 2 절에서는 각각의 영화에 대하여 사용자들의 편향성이 유사도에 어떻게 영향을 미치는지 기술한다. 제 3 절에서는 공통평가 영화와 편향성을 모두 고려하는 것에 대하여 논하였다.

제 1 절 공통평가 영화

[표 1]에서 볼 수 있듯이, 영화산업이 발전하면서 매년 개봉하는 영화의 편수는 점점 증가하고 그에 따라 누적영화의 수는 더욱더 증가하고 있다. 따라서 사용자가 영화를 선택할 수 있는 경우의 수가 상당히 많다.



[표 4] : 연도별 전국 극장, 스크린 수

[표 4] 를 보면 2013년도 전국 극장 수는 333개 스크린 수는 2184개로 나타나고 있다(2013 한국 영화산업 결산). 즉 평균적으로 하나의 극장에 6~7개의 스크린이 있다. 따라서 사용자는 영화를 보러 극장에 갔을 때, 어떤 영화를 볼 지 선택권을 갖게 된다. VOD를 생각해보면 이는 스크린 수의 제한도 없기 때문에 사용자가 선택할 수 있는 폭은 더욱 더 넓다.

사용자가 특정영화에 대하여 평가를 하고 평점을 주었다는 것은 그 영화를 보았다는 사실을 내포하고 있다. 사용자는 평가를 할 수 있는 영화가 많지만 그 중 일부 영화에만 평가를 하게 된다. 실제 사용된 데이터는 943명의 사용자와 1682개의 영화가 있다. 따라서 모든 사용자가 모든 영화에 대해 평가를 하게 된다면 약 160만개의 평점이 존재하게 된다. 하지만 실제로 평점의 개수는 10만개가 있다. 따라서 전체 중 약 6%의 평점이 실제로 존재한다.

사용자는 영화를 선택할 때 영화를 임의로 선정하지 않고 자신의 취향에 맞는 영화를 선택할 것이다. 영화를 감상하기 전에는 영화의 전체내용은 알 수 없다. 하지만 영화의 장르, 배우, 줄거리 등의 정보는 알 수 있다. 따라서 전체적인 영화에 대한 정보는 알 수 있기 때문에 이를 보고 자신의 취향에 더욱 적합한 영화를 선택하여 보게 될 것이다.

평점이 낮은 영화는 사용자가 해당 영화를 보고 만족하지 못했다는 것을 나타낸다. 즉 평점이 낮은 영화보다는 평점이 높은 영화가 취향에 맞고 사용자에게 만족을 주었다는 것을 의미한다. 따라서 사용자는 영화를 선택할 때 자신이 보고 더욱 만족할 만한 더 높은 평점을 줄 것 같은 영화를 선택할 것이다.

평점	개수
0	830
1	6054
2	11272
3	26934
4	33892
5	21018

[표 5] : 점수 별 평점개수

실제 데이터의 평점을 살펴보면 [표 5]와 같다. 총 10만개의 평점 중 절반이 넘는 54910개의 평점이 4,5점에 분포되어 있다. 실제로 0~5점의 평점이 고르게 분포되어 있지 않고 더 높은 평점에 많은 영화가 분포되어 있다는 사실을 알 수 있다. 즉 평점이 고르게 분포되어 있지 않다. 모든 평점이 존재하지 않고 평점이 없는 영화에 대해서 어떠한 규칙이 존재한다면 이를 무시하였을 때, 추천 시스템의 정확도는 떨어지게 된다(Steck et al. 2010). 또한 대부분

의 사용자는 그들이 만족한 영화에 대해서만 평점을 주는 경향이 있다(Pradel et al. 2012).

사용자는 영화를 선택할 때 임의로 선택하여 영화를 보는 것이 아니다. 따라서 평점이 없는 영화는 사용자가 자신의 취향에 맞지 않는다고 판단하여 보지 않았다고 해석할 수 있다. 따라서 영화의 평점이 실제로 낮다고 하더라도 사용자가 평가를 하여 평점을 주었다는 것은 다른 영화들 보다는 상대적으로 자신의 취향에 맞는 영화라고 할 수 있다. 사용자는 전체 영화 중에서 일부의 영화에 대해서만 평점을 준다. 모든 영화를 보고 평가한다는 것은 사실상 불가능하기 때문이다. 따라서 평가가 되어있는, 평점이 존재하는 영화는 평점이 없는 영화보다는 상대적으로 사용자가 만족할 만한 영화라고 할 수 있다. 즉 사용자가 평가한 영화의 목록만을 이용해도 사용자의 취향을 파악할 수 있을 것이다.

사용자들은 각자 자신의 취향에 적합하다고 생각한 영화를 선택하고 해당 영화를 감상 후 평가를 한다. 즉 어떤 두 사용자가 하나의 영화에 공통으로 평가했다면 그 평점에 관계없이 두 사용자 모두 해당영화를 선택했다는 것을 의미한다. 하지만 피어슨 상관관계의 경우 이렇게 공통으로 평가한 영화가 얼마나 많은지 고려하지 않고 해당 영화들의 평점만을 고려한다. 두 사용자의 공통 평가 영화의 수가 많거나 적거나 해당 영화들의 평점의 선형관계만 고려한다.

두 사용자가 공통으로 평가한 영화가 많다는 것은 해당 두 사용자의 영화 선택이 많이 겹쳤다는 것을 의미한다. 즉 취향이 비슷하다고 할 수 있다. 따라서 두 사용자의 유사도를 계산할 때, 공통으로 평가한 영화의 수를 고려해 보고자 한다.

자카드 유사도 역시 두 사용자가 공통으로 평점을 준 영화들의 수를 고려한다. 하지만 공통으로 평가한 영화의 수를 두 사용자가 평가를 한 영화의 총 수로 나누어 준다. 하지만 영화를 고르는 취향에 관계없이 영화 자체를 많이 보는 사람과 그렇지 않은 사람

이 있다. 사용한 데이터를 살펴보아도 20개의 평가만 한사람도 있지만 500개 이상의 영화에 대하여 평점을 준 사용자들도 있다.

	영화1	영화2	영화3	영화4	영화5
사용자1	3	5			
사용자2	3	5	1	2	
사용자3	3	5			3

[표 6] : 사용자 예시

위 [표 6]의 사용자 예시를 보면 사용자 1은 영화 1,2 에 대해서만 평점을 가지고 있다. 사용자 2,3은 모두 영화 1,2에 대하여 평점을 가지고 있다. 따라서 사용자 1과 두 사용자 2,3과의 공통으로 평가한 영화의 개수는 2개로 동일하다. 하지만 각각 사용자 1,2 사용자 1,3이 평가한 영화의 합집합의 개수는 4개 3개가 된다. 즉 자카드를 이용한 유사도는 사용자 1,3 이 사용자 1,2 보다 더 크게 나타나게 된다. 하지만 예시를 보면 영화 1,2 에 대한 평점은 사용자 1,2,3이 모두 같다. 따라서 사용자 1,2 보다 사용자 1,3이 더 유사하다고 평가하기는 어렵다.

제안하는 방법은 다음과 같다.

$$Sim_{com}(u, u_a) = n(I_u \cap I_{u_a})$$

u, u_a : 사용자

I_u, I_{u_a} : 사용자 u, u_a 가 선택한 영화의 집합

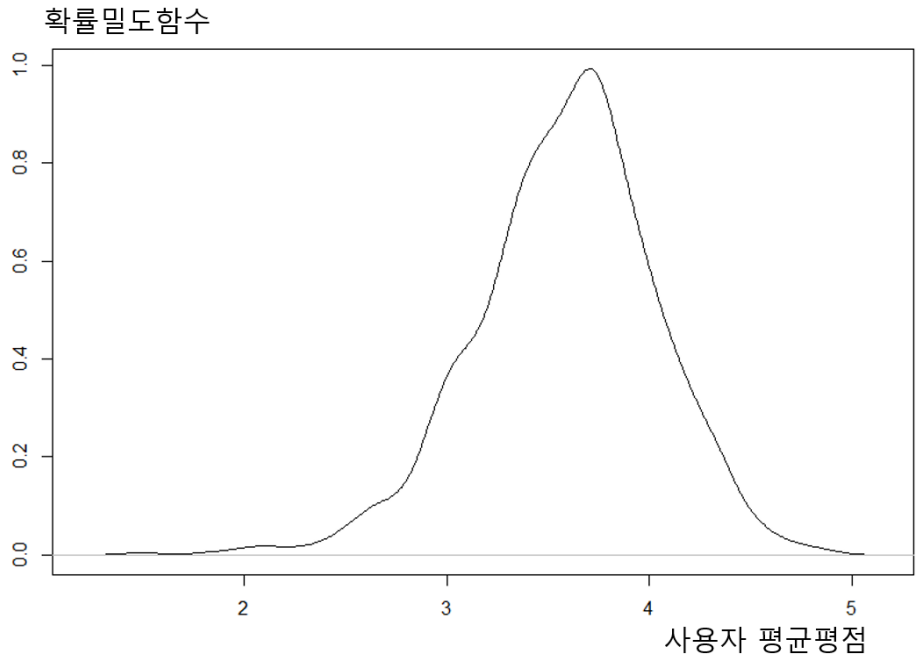
$n(I_u)$: 사용자 u가 평가한 영화의 개수

즉 두 사용자의 유사도를 구할 때, 두 사용자가 공통으로 평가한 영화의 개수만을 고려하는 것이다. 제안된 유사도를 사용하면 [표 6]의 경우 사용자 1,2 사용자 1,3의 경우 모두 2라는 같은 값의 유사도를 갖게 된다. 따라서 Sim_{com} 유사도를 사용하면 공통으로 평가한 영화의 평점에 관계없이 공통으로 평가한 영화의 수가 많을수록 높은 유사도를 갖게 된다.

제 2 절 편향성

제 1 절에서는 사용자들의 평점 점수에 대해서는 고려하지 않고 얼마나 많은 영화에 대해서 두 사용자가 모두 평가를 했는지만을 고려하였다. 본 절에서는 공통으로 평가한 영화에 대해서 사용자들의 평점 성향이 얼마나 비슷한지를 이용하여 유사도에 이용하고자 한다.

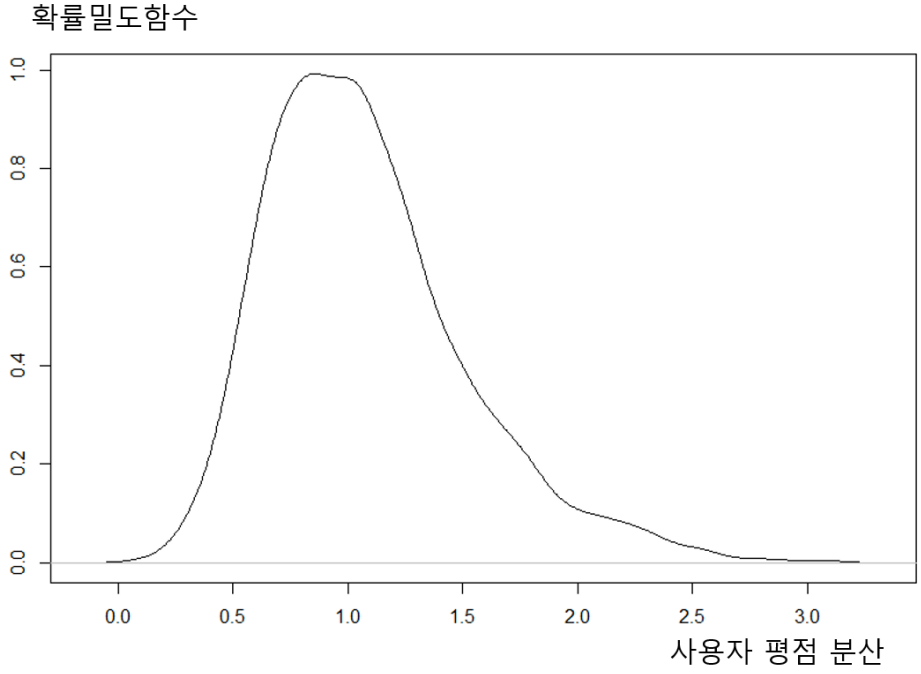
같은 취향을 가지고 있는 사용자라 하더라도 평점을 주는 성향은 다를 수 있다. 전반적으로 높은 평점을 주는 사용자가 있을 수도 있고 상대적으로 낮은 평점을 주는 사용자가 있을 수도 있다. 사용한 데이터의 각 사용자의 평균평점의 확률밀도 함수를 살펴보면 다음과 같다.



[그림 5] : 사용자 평균평점 확률밀도함수

[그림 5]를 보면 가장 많은 사용자들의 평균평점이 3.7점 정도로 분포되어 있는 것을 알 수 있다. 하지만 평균이 3점 미만인 사용자도 존재하고 4점 이상인 사용자도 존재하게 된다.

또한 사용자마다 평점을 주는 범위가 다를 수 있다.



[그림 6] : 사용자 평점 분산 확률밀도함수

[그림 6]를 보면 사용자 개개인의 평점의 분산을 알아 볼 수 있다. 약 1 정도의 분산을 갖고 있는 사용자가 가장 많았지만 그보다 더 큰 분산을 갖는 사용자도 있다.

따라서 같은 평점이라도 사용자마다 다른 의미를 가질 수 있다. 어떤 사용자에게는 4점이 평균보다 높은 점수일 수 있지만, 평균이 4점 이상인 사용자도 존재한다. 또한 자신의 평균보다 1점 더 높은 평점을 주었다고 가정했을 때, 분산이 큰 사용자에게는 1점 차이가 분산이 작은 사용자보다는 상대적으로 큰 의미를 갖지 않는다. 따라서 각 사용자의 평균과 분산, 표준편차를 이용하여 각각의 영화에 대해 사용자의 편향성을 분석해 보았다.

$$z_{u,i} = \frac{r_{u,i} - \overline{r_u}}{\sigma_u}$$

$r_{u,i}$: 사용자 u가 영화 i에 준 평점

$\overline{r_u}$: 사용자 u의 영화 평점의 평균

σ_u : 사용자 u의 영화 평점의 표준편차

각 사용자의 영화에 대한 편향성을 위와 같이 계산하였다. 이 후 두 사용자가 공통으로 평가한 영화에 대하여 두 사용자의 편향성이 얼마나 일치하는지 계산하였다.

$$Sim_{bias}(u, u_a) = \sum_{i \in (I_u \cap I_{u_a})} z_{u,i} * z_{u_a,i}$$

즉, 특정 영화에 대해서 두 사용자 모두 본인의 평균평점보다 높은 점수를 주거다 두 사용자 모두 평균평점보다 낮은 점수를 주면 유사도가 증가한다. 반대로 두 사용자의 편향성이 반대면, 한 영화에 대하여 한 사용자는 평균평점보다 높은 점수를 한 사용자는 평균평점보다 낮은 점수를 주었다면 유사도는 감소하게 된다.

제 3 절 공통 평가 영화와 편향성

제 1절 에서는 두 사용자가 공통으로 평가한 영화의 개수에 대해서 서술하였고, 제 2절 에서는 두 사용자의 편향성을 가지고 유사도를 계산하는 방법을 서술하였다. 본 연구에서는 위에서 설명한 두 가지 방법을 합쳐서 사용하였다.

$$Sim_{cb}(u, u_a) = Sim_{com}(u, u_a) + Sim_{bias}(u, u_a)$$

위와 같이 두 사용자가 공통으로 평가한 영화의 수와 편향성 점수를 합쳐서 유사도를 계산하였다.

두 사용자의 유사도를 계산할 때, 두 사용자가 공통으로 평가한 영화의 수와 해당평점의 편향성을 같이 고려해 주었다. 따라서 공통으로 평가한 영화가 많더라도 해당 영화에 대한 편향성이 다르다면 상대적으로 작은 유사도를 갖게 된다. 즉 같은 영화를 골랐지만 영화를 본 이후의 만족도가 서로 다르다면, 같은 영화에 대하여 한 사용자는 평균보다 만족했지만 다른 사용자는 불만족스러웠다면 Sim_{bias}는 음의 값을 갖게 된다. 즉 공통으로 평가한 영화의 개수보다 작은 유사도를 갖게 된다.

제 4 장 실험 및 결과

본 장에서는 이번 연구의 실험 결과와 그에 대해 기술한다. 제 1 절에서는 사용한 데이터에 대해 서술한다. 제 2 절에서는 제안한 유사도의 성능을 분석하기 위해 데이터를 어떻게 설정하여 실험하였는지를 서술하고, 제 3 절에서는 실험 결과에 대해서 서술하였다.

제 1 절 데이터 설명

본 연구에서는 MovieLens 의 데이터를 사용하여 제안한 방법을 실험하고 기존의 방법들과 비교해 보았다.

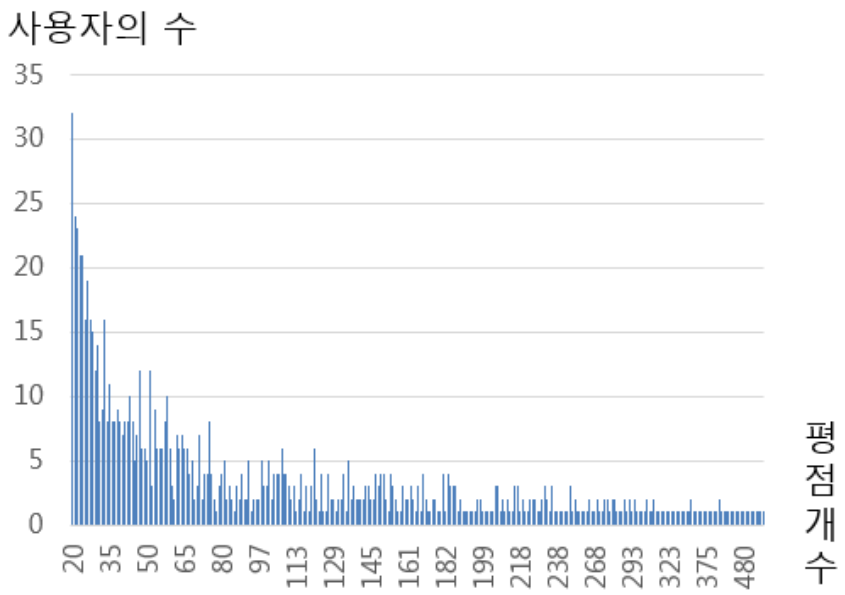
사용된 데이터는 943 명의 사용자와 1682 개의 영화로 구성되어 있으며 총 10 만개의 평점을 가지고 있다.

사용자ID	영화ID	평점
196	242	3
186	302	3
22	377	1
244	51	2
166	346	1
298	474	4
115	265	2

[표 7] : 초기 데이터 예시

제공되는 데이터는 [표 7]와 같이 사용자, 영화, 해당평점 3 차원의 벡터들로 이루어져 있다.

각각의 사용자들은 최소 20 개의 영화에 대해 평가를 하였다. 각각 평점 개수 당 해당 사용자의 수는 다음과 같다.



[표 8] : 평점 개수 별 사용자의 수

[표 8]를 보면 가장 평점을 적게 준 사용자는 32 명이고 20 개의 영화에 대하여 평가를 하였다. 또한 400 개 이상의 영화에 대하여도 평가를 한 사용자도 확인 할 수 있다.

해당 데이터를 이용하여 사용자-영화 행렬을 만든다. 이후 사용자-영화 행렬을 이용하여 각각 사용자들의 유사도를 계산하고 그 유사도를 바탕으로 이웃들을 선정한 뒤 평점을 예측하게 된다.

제 2 절 실험 설정

가. 테스트 평점 선택

실험은 해당 데이터를 기존에 제안된 유사도인 피어슨 상관관계와 자카드 유사도를 이용하고 제안한 방법과 비교하였다.

실험 결과를 분석하기 위해 기존에 제공된 10 만개의 평점 중 일부를 지운 뒤 예측한 결과와 비교하여 분석하였다.

	영화 1	영화 2	영화 3	영화 4	영화 5	영화 6
사용자1	0	3	5	0	1	0
사용자2	3	1	0	0	4	5
사용자3	0	0	2	5	4	4
사용자4	3	5	0	0	0	1
사용자5	1	0	0	0	4	4

	영화 1	영화 2	영화 3	영화 4	영화 5	영화 6
사용자1	0	?	5	0	1	0
사용자2	?	1	0	0	4	5
사용자3	0	0	2	?	?	4
사용자4	3	?	0	0	0	1
사용자5	1	0	0	0	4	?

[그림 7] : 실험 분석을 위한 데이터 설정

[그림 7]와 같이 일부 평점을 지운 뒤 예측된 평점과 비교하였다. 사용된 데이터는 [표 8]에서 알 수 있듯이 사용자가 최소 20 개의 영화에 대하여 평가하고 평점을 주었다. 따라서 본 연구에서는 943 명의 사용자에 대하여 각각 임의로 15 개의 평점을 지우고 나머지 평점들을 이용한 뒤 실제 평점과 예측된 평점을 비교하여 분석하였다. 따라서 전체 10 만개의 평점 중 14145 개의 평점 약 14%의 평점을 지운 뒤 해당 평점을 예측하였다.

나. 이웃 선정

본 실험에서는 세가지 유사도를 이용하여 이웃을 선정하였다.

피어슨 상관관계

$$\mathrm{Sim_{pc}(u,u')} = \frac{\sum_{i\in I(u)\cap I(u')}(x_{u,i} - \overline{x_u}) \cdot (x_{u',i} - \overline{x_{u'}})}{\sqrt{\sum_{i\in I(u)\cap I(u')}\left(x_{u,i} - \overline{x_u}\right)^2} \cdot \sqrt{\sum_{i\in I(u)\cap I(u')}\left(x_{u',i} - \overline{x_{u'}}\right)^2}}$$

자카드 유사도

$$\mathrm{Sim_{jac}(u,u')} = \frac{|I(u) \cap I(u')|}{|I(u) \cup I(u')|}$$

제안한 유사도

$$Sim_{cb}(u,\ u_a) = n(I_u \cap I_{u_a}) + \sum_{i\in(I_u\cap I_{u_a})} z_{u,i} * z_{u_a,i}$$

위와 같이 세가지 방법을 사용하여 사용자-사용자 유사도를 계산하였다. 이 후 유사도가 큰 순서대로 k 명의 이웃을 선정하고 그 이웃들을 이용하여 평점을 예측하였다.

$$S(u) = \{u' \mid sim(u,u') \text{ 가 큰 K 명의 } u'\}$$

위와 같이 이웃을 정의하여 각각의 사용자 마다 세가지 유사도를 이용하여 유사한 이웃들을 선정하였다.

다. 결과분석

결과를 분석하기 위해서 먼저 예측된 평점과 실제 평점의 차이인 Mean Absolute Error(MAE)를 이용하였다. MAE 를 구하는 방법은 다음과 같다.

$$MAE = \frac{\sum|r_{u,i} - \widehat{r_{u,i}}|}{N}$$

위 식에서 알 수 있듯이 MAE 는 실제 평점과 예측평점의 차이의 평균을 나타낸다. 따라서 MAE 값이 작을수록 예측한 평점이 정확하다는 것을 의미한다. 즉 예측한 평점이 실제 평점과 정확하게 일치한다면 MAE 는 0 이라는 값이 나올 것이다. MAE 를 이용하여 각각의 유사도 별 최소의 MAE 값이 나오는 이웃의 수를 구하였다.

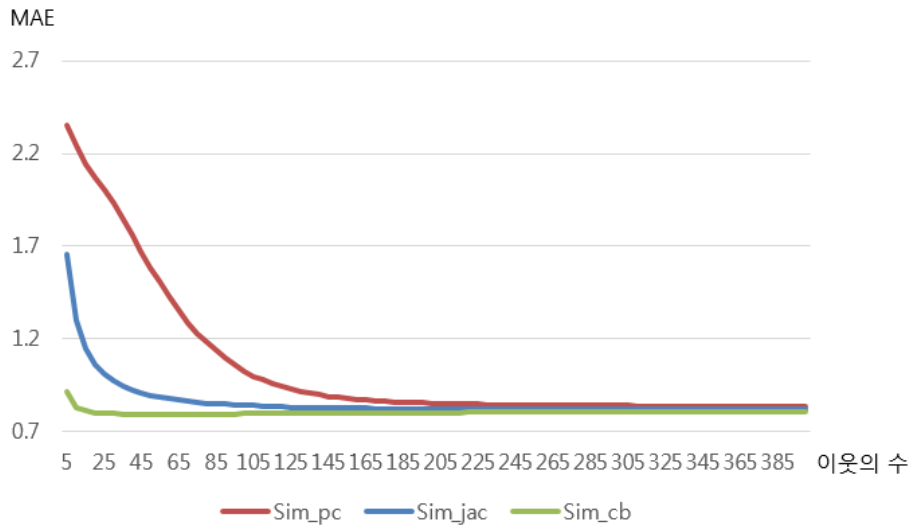
또한 예측한 평점 중 상위 10 개의 영화 중에 실제로 평가를 위해 임의로 지운 영화 중, 높은 평점 (4,5 점)을 갖는 영화가 얼마나 포함되는지 구하였다. 즉 추천시스템을 이용하여 상위 10 개의 영화를 추천하였을 때, 그 중 실제로 사용자가 고평가를 한 영화가 얼마나 있는지 구하였다.

마지막으로 4 점 이상의 평점을 갖는 영화 중에서 예측 값의 오차가 0.5 미만인 영화들이 얼마나 되는지 구하였다. 즉 평점이 4 점인 영화는 예측 값이 3.5 에서 4.5 사이 평점이 5 점인 영화는 예측 값이 4.5 를 넘는 비율을 구하여 성능을 평가하였다. 또 오차가 1 이상인 영화의 비율이 얼마나 되는지도 구하였다.

제 3 절 실험 결과

가. MAE

먼저 세가지의 유사도를 이용하여 k 개의 이웃을 선정했을 때의 예측 오차값 MAE 값을 비교하였다.



[표 9] : 전체 예측 영화 MAE

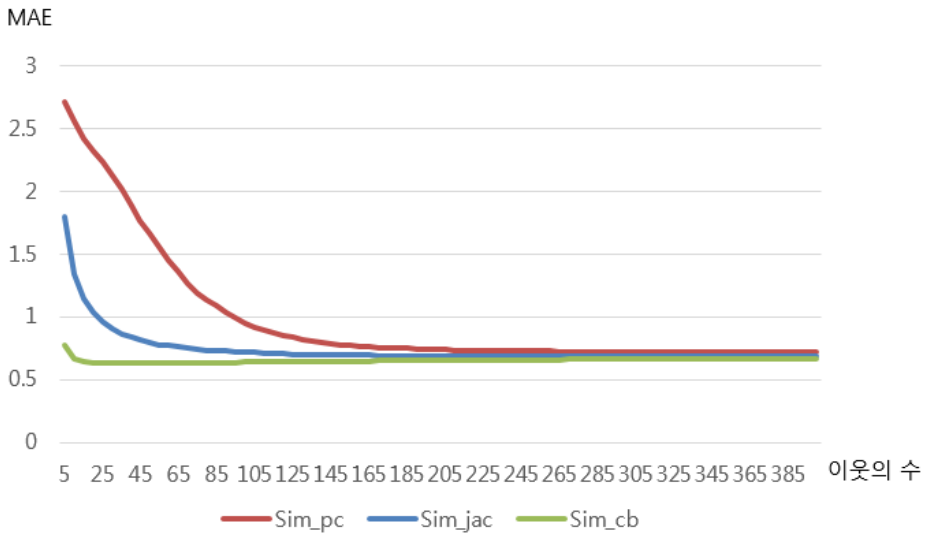
위 [표 9]를 보면 알 수 있듯이 본 연구에서 제안한 유사도를 이용했을 때, 작은 이웃의 수에서도 작은 MAE 값을 보였다. 각각 오차가 최소가 될 때의 MAE 값과 이웃의 수를 살펴보면 다음과 같다.

	Sim_{pc}	Sim_{jac}	Sim_{cb}
MAE	0.8358	0.8204	0.7929
이웃의 수	400	265	70

[표 10] : 최소 MAE

위 [표 10]를 보면 알 수 있듯이, 제안한 방법이 가장 작은 예측 오차 을 가졌다. 또한 상대적으로 매우 적은 이웃을 이용했을 때에도 좋은 예측 성능을 보였다.

평점이 4 점 이상인 영화에 대해서만 고려하였을 때의 MAE 도 확인해 보았다.



[표 11] : 고평점 영화의 MAE

	Sim_{pc}	Sim_{jac}	Sim_{cb}
MAE	0.7170	0.6887	0.6300
이웃의 수	400	260	35

[표 12] : 고평점 영화의 최소 MAE

4 점 이상의 평점을 가진 영화에 대해서는 세가지 유사도 모두 전체를 고려하였을 때보다 높은 정확도를 보였다. 제안한 유사도를 사용하였을 때, 피어슨 상관관계와 자카드 유사도를 사용하였을 때보다 훨씬 적은 이웃을 사용하였을 때 최소값을 가졌다. 또한 예측 오차도 제안한 방법이 상대적으로 작은 것을 확인할 수 있다.

나. Top 10 추천 리스트

임의로 선택한 14145 개의 영화 중에서 평점인 4 이상인 영화는 8198 개였다. 그 중 협력 필터링을 이용한 예측 평점 중 상위 10 개에 해당하는 영화의 비율이 얼마나 되는지 알아보았다.

	Sim_{pc}	Sim_{jac}	Sim_{cb}
Top 10	11.52%	16.53%	14.04%
이웃의 수	400	260	35

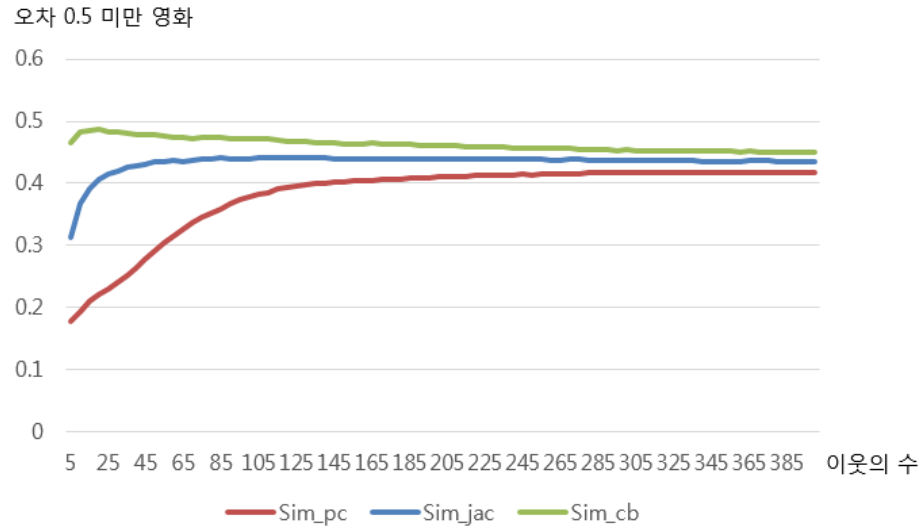
[표 13] : 추천 영화 Top10

MAE 값을 이용해 예측정확도가 가장 높은 이웃일 때 추천 Top10 영화에 실제로 사용자가 높은 평점 4 점 이상을 준 영화가 포함 될 확률은 [표 13]와 같았다. 즉 평가를 위해 임의로 선택한 영화 중 평점이 4 점 이상인 영화는 총 8198 개였다. 그 중 피어슨 상관관계를 사용하였을 때는 약 12%의 영화가 자카드를 사용하였을 때는 약 16% 제안한 방법을 사용하였을 때는 약 14%의 영화가 Top10 에 포함 되었다.

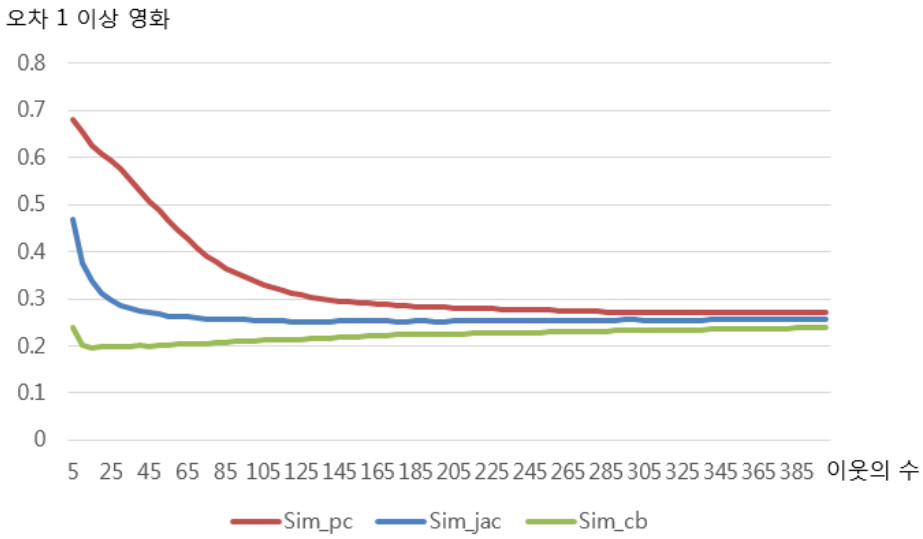
제안한 유사도를 사용하였을 때 Top10 에 포함된 실제 사용자가 평점 4 점이상을 준 영화의 비율은 피어슨 상관관계를 사용하였을 때보다는 크지만 자카드 유사도를 사용하였을 때보다는 작은 값을 가졌다.

다. 예측 오차 0.5 미만인 영화

임의로 평점을 지운 영화 중 평점이 4 점 이상인 8198 개의 영화에 대해서 예측 값과 실제 평점의 차이가 0.5 미만인 영화의 비율과 오차가 1 이상인 영화의 비율을 알아보았다.



[표 14] : 예측 오차 0.5 미만인 영화



[표 15] : 예측 오차 1 이상인 영화

[표 14], [표 15]에서 알 수 있듯이 제안한 방법을 사용하였을 때, 평점이 4점 이상인 8198개의 영화에 대하여 더 많은 영화의 예측 오차가 0.5 이하였고 더 적은 영화가 예측 오차가 1 이상이었다.

라. 결과 분석

피어슨 상관관계, 자카드 유사도와 제안한 유사도를 비교하였을 때, 제안한 방법은 상대적으로 작은 이웃을 선택하였을 때 좋은 성능을 보였다. 피어슨 상관과계와 자카드 유사도는 사용자별 선정하는 이웃의 수가 작았을 때, 낮은 성능을 보였다.

제안한 유사도를 사용하여 영화를 추천했을 때, Top10 영화에 포함된 실제 4 점이상의 평점을 준 영화는 자카드 유사도에 비해 적었다.

MAE 를 이용하여 예측 평점의 정확도를 알아 보았다. 제안한 유사도가 피어슨 상관관계나 자카드 유사도를 이용했을 때보다 예측 정확도가 높음을 확인할 수 있었다. 또한 4 점 이상의 평점인 영화들에 대해서 제안한 방법이 0.5 이하의 오차를 갖는 경우가 많았다. 또한 오차가 1 이 넘는 영화는 제안한 영화가 가장 적었고 자카드 유사도 피어슨 상관관계를 사용했을 때 더 많았다. 따라서 알려지지 않은 평점을 예측하는 데 있어 제안한 유사도를 이용하여 협력 필터링을 수행하였을 때 더 높은 정확도를 보였다.

제 5 장 결 론

본 연구에서는 두 사용자의 유사도를 계산할 때 두 사용자가 공통으로 평가한 영화의 수가 많으면 유사한 사용자라고 판단하였다. 또한 공통으로 평가한 영화에 대해서 사용자의 편향성, 즉 평점을 본인의 평균보다 높게 주었는지 낮게 주었는지의 정도를 계산하여 비슷한 편향성을 보이는 사용자에게 더 높은 유사도를 주었다.

제안한 유사도를 사용하여 추천을 할 때, 기존의 방법들보다 적은 이웃을 선정하고도 높은 정확도를 보였다. 이웃들의 평점을 이용하여 특정 사용자의 평점을 예측할 때 각각의 이웃은 전체 이웃과의 유사도 합 중에서 본인의 유사도 만큼 비중을 얻는다. 선정하는 이웃의 수가 많아지게 되면 전체 이웃과의 유사도의 합은 커지게 된다. 따라서 중요한 이웃의 정보가 소실될 수 있다.

영화를 선택하는 사용자의 취향은 다양하다. 특정 취향은 많은 사람들이 공유할 수 있지만 특정 취향은 소수의 사람들 만이 공유하고 있을 수 있다. 따라서 적은 이웃을 사용하여서 예측을 할 수 있다면 소수의 사람이 공유하는 취향에 대해서도 비교적 정확한 예측을 할 수 있다. 이웃의 수가 적기 때문에 특정 사용자의 이웃을 선정할 때 상대적으로 유사하지 않은 사용자가 포함되는 경우를 줄일 수 있다.

제 6 장 향후 연구 방향

본 연구는 사용자기반 아이템기반의 협력 필터링 중에서 사용자기반 협력 필터링만을 고려하였다. 제안한 방법을 이용하여 아이템기반 협력 필터링을 수행하고 그에 따른 결과분석을 통해 성능을 검증해보아야 한다.

본 연구는 MovieLens 에서 제공한 데이터를 이용해서 성능을 검증하였다. 이 외에도 다양한 데이터, 영화 뿐 아니라 다른 아이템에 대한 추천시스템에도 제안한 방법을 적용하여 검증해 볼 수 있다.

참고 문헌

Hao Ma, Irwin King and Michael R.Lyu : Effective missing data prediction for collaborative filtering, SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007

Bruno Pradel, Nicolas Usunier, Patrick Gallinari : Ranking With Non-Random Missing Ratings: Influence Of Popularity And Positivity on Evaluation Metrics, Recsys'12 Proceedings of the sixth ACM conference on Recommender systems, 2012

H. Steck : Training and testing of recommender systems on data missing not at random, KDD '10 Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010

Jun Wang, Arjen P.de Vries, Marcel J.T. Reinders : Unifying User-based and Item-based Collaborative Filtering Approached by Similarity Fusion, SIGIR '06 Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006

Hao Ji, Jinfeng Li, Changrui Ren, Miao He : Hybrid Collaborative Filtering Model for improved Recommendation, Service Operations and Logistics and Informatics, 2013

Yue Shi, Alexandros Karatzoglou, Linas Baltrunas, : CLiMF_Learning to Maximize Reciprocal Rank with Collaborative Less is More Filtering, RecSys '12 proceedings of the sixth ACM conferenceon Recommendersystems, 2012

Hyung Jun Ahn : A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem, Information Sciences, volume 178, ELSEVIER, 2008

Haifeng Liu, Zheng Hu, Ahmad Mian, Hui Tian, Xuzhen Zhu : A new user similarity model to improve the accuracy of collaborative filtering, Knowledge-Based Systems, Volume 56, ELSEVIER, 2014

Xiaoyuan Su, Taghi M. Khoshgoftaar : A survey of collaborative filtering techniques, Journal Advances in Artificial Intelligence Volume 2009, Article No.4, ACM, 2009

Yue Shi, Martha Larson, Alan Hanjalic : Exploiting user Similarity based on Rated-movie pools for improved user-based collaborative filtering, RecSys '09 porceedings of the third ACM conference on Recommender systems, pages 125-132, 2009

Jun Wang, Arjen P. De Vries and Marcel J. T. Reinders : Unified Relevance Models for Rating Prediction in Collaborative Filtering, ACM Transactions on Information Systems, Volume 26, Issue 3, 2008

Xiaoyuan Su and Taghi M. Khoshqoftaar : A Survey of Collaborative Filtering Techniques, Advances in Artificial Intelligence, Volume 2009, 2009

J. Bobadilla, F. Serradilla, J. Bernal : A new collaborative filtering metric that improves the behavior of recommender systems, Knowledge-Based Systems, Volume 23, ELSEVIER, 2010

MovieLens Data : http://grouplens.org/datasets/movielens/

2013 한국영화산업 결산 – 한국 영화산업 사상 최고의 호황, 매출 1 조 8 백억원 기록 : 영화진흥위원해 정책연구부, 2014

2012 년 한국 영화산업 결산 – 영화진흥위원해 영화정책센터, 2013

Abstract

Improvement of Recommender System by common rated movie similarity of users

Younsung Koo

Department of Industrial Engineering

The Graduate School

Seoul National University

Recommender System is very important and used globally. As the internet service has grown and the size of contents such as movies, musics and books has been enormous, the choice of contents to consume become harder. However, contents that people really can consume is limited. So people want to choose contents that suits their taste. Therefore the recommender system is very important and can be useful. In real world, Amazon and Netflix already adopt recommender systems and recommend adequate contents to each consumer and it has been shown very effective.

Collaborative filtering is widely studied and used to recommend contents. User-based collaborative filtering finds similar users to recommend interesting items. Therefore calculating similarity is very important part of collaborative

filtering. In this paper, the new user similarity is proposed to recommend movies.

People give ratings to movies they watched in the past. The rated movies were selected by users. When they chose those movies, they thought that it could be their favorite. It means, other movies that has no ratings were less attractive to the users. To calculate similarity between two users, I considered the number of common rated movies. Also, the bias score how much higher score does the movie have than the user's average score, is considered.

By using proposed similarity through collaborative filtering, the error of predicted score was less than using Pearson correlation and Jaccard index for similarity. By using proposed similarity, I got 0.7929 as the value of mean absolute error. Also, proposed similarity needed less neighbor than other similarities.

Keywords : Collaborative filtering, User similarity, Common rated movie, User bias

Student Number : 2012-23307