제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

하이브리드 컨텐츠 추천시스템의 설계 및 구현

왕지현, 임명은, 윤보현
한국전자통신연구원, 휴먼정보처리연구부
e-mail : jhwang@etri.re.kr

The Design and Implementation of Hybrid Contents Recommender

JiHyun Wang, Myung-Eun Lim, Bo-Hyun Yun
Dept. of Human Information Processing
Electronics and Telecommunications Research Institute

요 약

본 논문은 협업에 의한 추천 방법과 내용에 의한 추천 방법을 혼합한 하이브리드 추천 방법을 제시한다. 일반적으로 '영화'정보와 같이 아이템에 대한 설명이 부족하거나 실제 영화의 내용과는 차이가 있는 컨텐츠의 경우에는 '주연', '감독', '줄거리'와 같이 실제 아이템의 내용이 아닌 부수적인 정보를 통해 평가값을 예측하는 방법보다 협업에 의한 평가값의 예측을 통해 더 낳은 추천을 제공할 수 있다. 이에 따라 본 연구는 내용에 기반한 추천방법에 의존하지 않고 사용자의 유사 선호 경향이 있는 타 사용자의 평가값들을 사용하여 추천하며, 협업에 의해 추천될 수 없는 아이템들에 대해 내용기반 추천 방법을 사용하는 하이브리드 컨텐츠 추천 시스템을 설계, 구현하였다.

1. 서론

협업에 의한 추천 시스템(Collaborative-Based Recommender)은 추천 대상이 되는 아이템들에 대한 사용자들의 선호 정도에 따라 유사한 선호 경향을 갖고 있는 타 사용자에게 평가가 좋은 아이템을 추천해 주는 시스템이다. 사용자들의 평가에 의해 아이템을 추천하기 때문에 "word-of-mouth"에 의한 아이템 추천을 자동화하는 시스템이라 할 수 있다.[8] 협업에 의한 추천 대상이 되는 아이템들은 주로 아이템에 대한 설명(description)이 부족하여 주어진 아이템 정보만으로는 아이템의 특성을 분석하기 어렵거나 현재 기술 수준상 분석하기 어려운 아이템들이 대상이 된다. 예를 들어, '이미지'나 '동영상', '음악' 파일과 같은 컨텐츠들이 대상이 될 수 있다.

협업에 의한 추천 방법은 아이템 내용 자체로부터 추천을 하는 것이 아니기 때문에 다음과 같은 주요 문제들이 발생하게 된다.

- 데이터 부족(Data Sparseness)

  협업기반 추천 방법은 유사 선호도를 갖고 있는 사용자들간에 공통으로 평가한 아이템들(co-rated Items)에 의해 추천을 하게 된다. 그러나 일반적으로 사용자들의 평가 수에 비해 아이템들의 수가 절대적으로 많기 때문에 사용자들간에 공통으로 평가된 아이템들의 수가 거의 없거나 매우 적은 경우가 많다. 이러한 문제를 해결하기 위해 차원 축약(Dimension Reduction) 방법을 이용하여 유사 선호도 사용자 그룹을 찾거나[1,6], 내용기반 필터를 사용하여 유사한 정보를 포함한 컨텐츠들이 유사한 선호도로서 평가될 것이라는 가정 하에 부족한 평가 정보를 채우는 방법을 사용한다.[2,4]

- 새 아이템 추천(Cold-Start)

  어느 사용자도 평가하지 않은 아이템은 아이템의 선호도를 알 수 없기 때문에 어느 누구도 추천할 수 없다. 이 문제는 협업기반 추천 방법의 근원적인 문제로 인식이 되어 왔고 이에 대한 해결책으로서 내용기반 필터를 협업기반 필터와 혼합한 하이브리드(Hybrid)에 의한 추천 방법이 근래에 등장하기 시작했다.[2,3]

본 논문은 위의 대표적인 주요 문제들을 해결하는 하이브리드에 의한 컨텐츠 추천 시스템을 설계, 구현하고 순수 협업기반 필터링 방법(Pure Collaborative-

based Filtering) 또는 순수 내용기반(Pure Contents-based Filtering) 필터링 방법보다 나은 성능을 보임을 실험을 통해 나타낸다.

본 논문의 구성은 다음과 같다.

2 장은 협업 및 하이브리드 추천 시스템들의 최근 관련연구를 알아보고 3 장은 본 논문의 접근 방법 및 실험을 위한 대상 도메인과 데이터 셋(dataset)을 기술하며, 4 장 및 5 장은 하이브리드 추천 시스템의 구성 모듈인 협업기반 필터와 내용기반 필터를 상세히 설명하고 6 장은 개별 필터들을 하이브리드하는 방법을 설명하며 7 장은 실험 및 결과를, 마지막 8 장은 결론으로써 끝을 맺는다.

2. 관련연구

[1]은 영화 정보 내의 '관련 영화 추천', 영화 리뷰 내의 '인용된 영화 제목'등에 의해 하나의 영화가 다른 영화들을 추천하는 '영화 vs. 영화'의 추천 프로파일 매트릭스(matrix)를 구성한다. 프로파일 매트릭스의 데이터 부족 문제를 해결하기위해 SVD(Singular Value Decomposition)를 이용하여 차원 축약(Dimension Reduction)을 수행한다. [1]에 따르면 SVD 를 적용하는 것이 하지 않은 것보다 성능이 뛰어남을 보여주고 있다.

[2]는 영화 정보를 추천하기 위해 평가값을 예측할 때는 내용기반 필터링 방법을 사용하고 유사 선호도 그룹을 찾을 때는 협업에 의한 필터링 방법을 사용하는 하이브리드 방식을 구현하였다. 실제 사용자가 평가한 평가값을 바탕으로 Naïve Bayes Classifier 에 의한 내용기반 필터를 학습하고 사용자가 평가하지 않은 아이템들은 학습된 내용기반 필터로 평가값을 예측한다. 이와 같이 평가값을 예측하여 사용자 프로파일내의 직접 평가하지 않은 아이템들의 평가값을 채움으로써 1. 서론에서 거론된 '데이터 부족' 문제와 '새 아이템 추천' 문제를 해결한다. [2]에 따르면 차원 축약 등 어떠한 방법을 사용하지 않은 순수 협업기반 필터링 방법이 순수 내용기반 필터링 방법 보다 나은 성능을 보여주고 있고, 내용기반 필터링 방법으로 평가값을 예측하여 협업으로 유사 선호도 그룹을 찾는 하이브리드 방법이 이보다 더 낳은 성능을 나타냄을 보여주고 있다.

[3]은 문서 필터링을 위해 LSI(latent semantic indexing)를 사용한 내용기반 필터와 협업기반 필터를 혼합한 하이브리드 방법을 제안하였다. 문서 내의 키워드의 발생 빈도에 따른 가중치(weight)가 '키워드 vs. 문서' 매트릭스의 각 셀(cell)에 저장된다. 그리고 사용자의 문서에 대한 평가값을 정규화한 '평가값' 매트릭스를 곱하여 각 문서에 대한 사용자의 선호 프로파일 매트릭스를 만든다. 그리고 나서 이 프로파일 매트릭스에 대해 SVD 를 수행하여 Rank-k 의 Approximation 을 구한다. 사용자의 프로파일을 나타내는 Centroid 를 구하기 위해 사용자와 관련성이 높은 문서들에 대한 키워드 벡터들을 구하여 LSI 공간내에서 새로운 문서들에 대한 순위를 결정한다.

3. 접근방법 및 대상 도메인

협업에 의한 필터링 방법은 유사한 선호도를 갖고 있는 다른 사용자들로부터 아이템을 추천 받는 것이기 때문에 유사 선호 경향이 있는 다른 사용자들을 찾기 위해 사용자들이 이전에 평가한 다른 아이템들 중에서 공통으로 평가한 아이템들로부터 유사 선호 사용자 그룹을 찾게 된다. 따라서 두 사용자 간에 공통으로 평가한 아이템이 없으면 두 사용자의 선호 경향이 유사한지 판별할 수 없게 된다. 이와 같은 데이터 부족문제를 해결하기 위해 [2]는 기존에 평가한 영화와 유사한 내용의 영화는 사용자가 유사하게 평가할 것이라는 가정을 한다. 이에 따라 평가하지 않은 아이템들의 평가값을 내용기반 필터로 미리 예측함으로써 예측된 평가값과 실제 사용자가 평가한 값들로부터 두 사용자의 유사 선호 경향을 분석한다.

아이템에 대한 정보가 부족한 도메인에서는 아이템을 추천하기 위해 협업에 의한 추천 방법이 내용에 의한 추천 방법보다 더 효율적이다. 협업에 의한 필터링에 있어서 평가값의 예측은 해당 사용자와 선호도가 유사한 사용자들에의해 의존적이기 때문에 정확한 유사 선호 사용자 그룹을 찾는 것이 중요하다. 따라서 본 논문은 영화에 대한 컨텐츠(description)가 실제 영화에 대한 사용자의 평가와 차이가 크기 때문에 영화 컨텐츠에 의해 유사 사용자를 찾는 방법이 평가수가 적더라도 적은 평가를 바탕으로 한 유사 사용자를 찾는 방법보다 오히려 잘못된 유사 선호 그룹을 찾을 가능성이 높을 것이라는 가정을 한다. 이에 따라 본 논문의 데이터 부족 문제는 프로파일 매트릭스의 SVD 적용에 따른 차원 축약 방법을 사용하며 유사 선호 그룹 내의 어느 사용자도 평가하지 않은 아이템들에 대한 문제는(Cold-Start Problem) 내용기반 필터에 의한 유사한 컨텐츠의 영화 평가값을 사용한다.

실험에 사용된 데이터 셋은 널리 사용되는 EachMovie 데이터 셋을 이용하였고 영화의 줄거리를 얻기 위해 VideoKorea (www.videokorea.com)사이트로부터 일부 영화의 줄거리를 추출하여 사용했다. EachMovie 데이터 셋은 72,916 명의 사용자가 1,628 개의 영화에 대해 평가한 2,811,983 라인의 로그 데이터로 구성되어 있다. 사용자 평가값은 '0.0', '0.2', '0.4', '0.6', '0.8', '1.0'의 6-scale 의 값으로 되어 있으며 숫자가 적을수록 선호하지 않거나 좋아하지 않은 평가를 나타내고 숫자가 높을수록 선호하거나 좋아하는 평가를 나타낸다.

4. 협업기반 필터링

협업기반 필터링은 다음과 같은 주요 단계를 거친다.

<4.1> 사용자 프로파일 구성 및 차원 축약

<4.2> 유사 선호도 그룹 찾기
<4.3> 평가값 예측

4.1 사용자 프로파일 및 차원 축약

사용자의 아이템들에 대한 평가값은 '사용자 vs. 아이템'간의 M * N 의 프로파일 매트릭스에 저장된다. 매트릭스의 각 행(Row)은 열(column)을 나타내는 아이템들에 대한 각 사용자의 사용자 프로파일 벡터(Vector)로서 표현된다.

M * N 의 매트릭스는 데이터 부족 현상에 의해 0 의 값이 대부분이기 때문에 SVD 를 적용하여 매트릭스의 Rank-k 의 Approximation 을 구한다. SVD 를 적용한 매트릭스는 오리지날 매트릭스와 가장 유사한 매트릭스가 되며 Approximated 매트릭스의 각 사용자 프로파일 벡터들은 0 에 근접한 임의의 값들로 채워지게 된다. 다시 말하면, SVD 적용에 의해 Approximated 된 각 사용자 프로파일 벡터들은 각 아이템들에 대한 선호도를 나타내는 차원 공간(Dimension Space)안에 위치하게 된다. SVD 에 대한 자세한 설명은 [9,10]을 참조하기 바란다.

4.2 유사 선호도 그룹 찾기

벡터 공간 안에 위치한 각 사용자 프로파일 벡터들의 유사도는 Cosine 유사도 계산을 사용하여 구한다.

$$Weight(a,b) = \cos(\vec{a},\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{\|a\|_2 \times \|b\|_2} \qquad 식(1)$$

사용자와 사용자간의 유사도 값은 M * M 의 매트릭스에 저장되며 특정 임계값(Threshold)보다 큰 유사 선호도를 갖는 사용자들을 유사 선호도 그룹으로(Neighbors) 판정한다. 계산된 유사도 값은 아이템 추천에 대한 사용자의 신뢰도를 나타내는 해당 사용자의 가중치(Weight)로서 사용된다.

4.3 평가값 예측

사용자의 특정 아이템에 대한 평가 예측값은 유사 선호도 그룹내의 해당 아이템을 평가한 사용자들(Neighbors)의 평가값에 대한 Weighted Mean 으로써 구한다.

$$p_{a,i} = \overline{r_a} + \frac{\sum_{u=1}^{n} \left(r_{u,i} - \overline{r_u}\right) \times W_{a,u}}{\sum_{u=1}^{n} W_{a,u}} \qquad 식(2)$$

$P_{a,i}$: 사용자 a 의 아이템 i 에 대한 평가 예측값
r_a : 사용자 a 의 아이템 평균 평가값
r_u : 유사 선호도를 갖는 사용자 u 의 아이템 평균 평가값
$r_{u,i}$: 유사 선호도를 갖는 사용자 u 의 아이템 i 에 대한 평가값
$W_{a,u}$: 사용자 a 와 사용자 u 의 유사도

5. 내용기반 필터링

유사 선호도 그룹내의 어떠한 사용자도 평가하지 않은 아이템은 협업에 의해 추천할 수 없는 아이템이다. 이와 같은 아이템은 해당 아이템과 내용이 유사한 다른 아이템들의 평가값을 사용한다. 내용이 유사한 아이템은 평가값도 유사할 것이라는 가정에 의한 것이다.

내용기반 필터는 Naïve Bayes Classifier 를 사용하여 사용자의 '0.0'부터 '1.0'까지의 6 클래스로의 평가값에 대해 영화의 줄거리에 등장하는 키워드들의 발생 확률에 대한 학습을 수행한다.

Naïve Bayes Classifier 의 기본 가정은 각 키워드가 발생하는 확률은 문서 클래스에 종속적이지만 키워드의 발생 위치나 문맥에는 독립적이다라는 가정에 기반을 두고 있다.

$$P(c_k \mid D) = \frac{P(c_k)}{P(D)} \prod_{i=1}^{|D|} P(w_i \mid c_k) \qquad 식(3)$$

D : 문서, Document
c_k : 문서 클래스
W_i : 문서내에 발생하는 i 번째 키워드

본 논문은 영화의 '장르'에 따라 '줄거리'에 등장하는 키워드들의 빈도수가 다르다고 가정하고 이에 따라 '장르'별 '줄거리'에 등장하는 키워드들의 발생확률을 종속적으로 고려한다

$$P(c_k \mid M) = \frac{P(c_k)}{P(M)} \prod_{i=1}^{|S_m|} \prod_{j=1}^{|S_n|} P(w_{mi}, w_{nj} \mid c_k) \qquad 식(4)$$

$$C_{NB} = \arg\max_{c_k \in C} P(c_k \mid M)$$

M : 영화, Movie
S_m : 장르, S_n : 줄거리
w_{mi} : 장르에서 발생하는 i 번째 키워드
w_{nj} : 줄거리에서 발생하는 j 번째 키워드

내용기반 필터의 기본 알고리즘은 다음과 같다.

각 클래스 c_k에 대해, $P(c_k)$와 $P(w_{mi}, w_{nj}|c_k)$를 계산한다.

1. $Docs_k \leftarrow$ 전체 영화 집합 M 에서 클래스가 c_k인 영화들.
2. $P(c_k) \leftarrow \frac{|Docs_k|}{|M|}$
3. $Genre_k \leftarrow Docs_k$의 모든 '장르'들을 하나의 문서로 구성.
4. $Voc_G \leftarrow$ 전체 영화 집합 M 의 모든 '장르'에서 등장하는 서로 다른 키워드들

Voc_G에 있는 각 키워드, w_{mi}에 대해,
5. $n_{GK} \leftarrow Genre_k$에서 키워드, w_{mi}의 발생 횟수
6. $P(w_{mi}) \leftarrow \frac{n_{GK}}{|Genre_k|}$
7. $Text_k \leftarrow$ '장르'가 w_{mi}인 영화들의 '줄거리'들을 하나의 문서로 구성
8. $Voc_T \leftarrow Text_k$에서 발생하는 서로 다른 키워드들

Voc_T에 있는 각 키워드, w_{nj}에 대해,
9. $n_{Tk} \leftarrow Text_k$에서 키워드, w_{nj}의 발생 횟수
10. $P(w_{mi}, w_{nj} \mid c_k) = P(w_{nj} \mid w_{mi}, c_k) \times P(w_{mi} \mid c_k) \leftarrow \frac{n_{T_k}+1}{|Text_k|+|Voc_T|} \times \frac{n_{GK}+1}{|Genre_k|+|Voc_G|}$

$P(c_k)$: 임의의 영화 하나를 선택했을 때 클래스 c_k 의 영화일 확률

$P(w_{mi}, w_{mj} \mid c_k)$: 클래스 c_k 의 영화 M 의 '장르', S_{mi}와 '줄거리', S_{mj}로부터 임의의 하나의 키워드를 각각 하나씩 선택했을 때 선택된 키워드들이 w_{mi}와 w_{mj}가 될 확률.

6. 하이브리드 추천 방법

본 논문의 하이브리드 추천 방법은 영화의 선호도를 결정하기에는 영화에 대한 설명(description)이 실제 영화의 내용과는 차이가 크기 때문에 내용기반 필터에 의해 유사 선호도 그룹을 찾는 것보다 SVD 를 사용한 협업기반 필터에 의해 유사 선호도 그룹을 찾는다. 유사 선호도 그룹 내에 해당 아이템을 평가한 사용자들이 있으면 협업 기반 필터를 적용하여 평가값을 예측하고, 어느 누구도 해당 아이템을 평가하지 않았다면 내용기반 필터를 적용하여 이전에 사용자가 평가한 유사한 내용의 아이템에 대한 평가값으로 평가값을 예측한다.

$$p_{ij} = \begin{cases} r_{ij} & \text{If 사용자 i 가 영화 j 를 평가했다면,} \\ c_{ij} & \text{Otherwise, If Neighbor 가 영화 j 를 평가했다면,} \\ b_{ij} & \text{Otherwise} \end{cases}$$

r_{ij} : 사용자 i 가 실제로 평가한 값
c_{ij} : 협업기반 필터에 의해 예측된 평가값
b_{ij} : 내용기반 필터에 의해 예측된 평가값

7. 실험 및 고찰

EachMovie 로그 데이터의 전체 사용자들 중 40 개 이상 평가한 사용자들 중에서 임의로 10%를 뽑아서 테스트 사용자 셋(Test User Set)으로 하고 나머지 사용자들을 학습 사용자 셋(Training User Set)으로 한다. 학습 사용자들은 테스트 사용자들에 대한 유사 선호 사용자들을 구성하기 위해 사용된다.

테스트 사용자들에 대해 최소 40 개 이상씩을 평가했기 때문에 전체 평가 개수의 75%를 학습하고 나머지 25%의 평가값을 예측하여 MAE 와 ROC-4 를 측정한다.

	MAE	ROC-4
Pure Contents-Based Filter	1.082	0.6133
Pure Collaborative Filter	1.045	0.6394
Hybrid Filter	0.9784	0.6548

테이블 1 : 실험 결과

MAE 는 예측한 값과 실제 사용자가 평가한 값과의 차이를 정규화 한 것이고 ROC-4 는 0.8, 1.0 은 Like, 나머지 평가값들은 Hate 로 평가하여 Like 의 개수를 전체 개수로 나눈 값이다.

8. 결론

본 논문의 하이브리드 추천 방법은 컨텐츠 정보가 부족한 영화 정보와 같은 아이템을 추천할 때 데이터 부족 문제(Data Spaseness)를 해결하기 위해 협업기반 필터를 사용하며 새 아이템 추천 문제(Cold-Start Problem)에 대해서는 내용기반 필터를 사용한다.

내용기반 필터를 사용하는 기존의 연구가 '주연', '감독', '장르'와 같은 아이템 정보(description)가 유사한 영화는 사용자가 이전에 평가한 영화와 유사하게 평가할 것이라는 가정에 기반을 두고 있지만 본 연구에서는 유사 선호 그룹 내에서 해당 영화를 평가한 사용자가 있을 경우에는 그 사용자 또는 사용자들이 평가한 평가값을 평가값 예측에 사용함으로써 내용기반에 의한 예측을 사용하지 않고 협업에 의한 예측을 사용한다. 이와 같은 가정에 대한 배경은 실제 영화를 본 사용자의 평가와 내용만을 가지고 평가한 사용자의 평가값이 차이가 크기 때문이다. 7 장의 실험은 내용기반 필터를 사용한 추천보다 협업에 의한 추천이 더 낮은 성능을 보이고 있음을 나타내고 있으며, 두가지 방법을 혼합한 하이브리드 방법이 각 방법만을 사용한 경우보다 더 낮은 성능을 보임을 나타내고 있다.

참고문헌

[1] Miles Efron and Gary Geisler. "Is it all About Connections? Factors Affecting the Performance of a Link-Based Recommender System". *2001 Who's Who in Recommender Systems*, Workshop in conjunction with ACM 2001 SIGIR

[2] Prem Melville, Raymond J.Mooney and Ramadass Nagarajan. "Content-Boosted Collaborative Filtering". *2001 Who's Who in Recommender Systems*, Workshop in conjunction with ACM 2001 SIGIR

[3] I. Soboroff and C.Nicholas. "Combining content and collaboration in text filtering". In T. Joachims, editor, Proceedings of the IJCAI'99 Workshop on Machine Learning in Information Filtering, pp86-91, 1999

[4] John S.Breese, David Heckerman and Carl Kadie. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering". Technical Report of Microsoft Research, 1998

[5] Badrul Sarwar, George Karypis, Joseph Konstan and John Riedl. "Analysis of Recommendation Algorithms for E-Commerce". ACM Conference Electronic Commerce. 2000

[6] Daniel Billsus and Michael J. Pazzani. "Learning Collaborative Information Filters". Proceedings of 15th International Conference on Machine Learning, 1998

[7] Nathaniel Good, J.Ben Schafer, Joseph A.Konstan, Al Borchers, Badrul Sarwar, Jon Herlocker, and John Riedl. "Combinining Collaborative Filtering with Personal Agents for Better Recommendations". Proceedings of AAAI/IAAI Conference, 1999

[8] Upendra Shardanand. "Social Information Filtering for Music Recommendation". Master's thesis, MIT. 1994

[9] Scott Deerwester, Susan T.Dumais, George W. Furnas, Thomas K.Landauer, Richard Harshman. "Indexing by Latent Semantic Analysis". Journal of the American Society for Information Science, 41(6), pp391-407, 1990

[10] M.W.Berry, S.T.Dumais and G.W.O'Brien. "Using Linear Algebra for Intelligent Information Retrieval". SIAM Review 37(4), pp573-595