

전자상거래를 위한 대표 속성을 이용한 협력적 여과 방법

류영석*, 양성봉
연세대학교 컴퓨터과학과

A New Collaborative Filtering Method Using Representative Attributes for Electronic Commerce

Young-Suk Ryu, Sung-Bong Yang
Dept. of Computer Science, Yonsei University

요 약

전자상거래 분야에서 급속도로 증가하고 있는 정보들 중에서 사용자가 자신의 기호에 맞는 정보들만 선택하기 위하여 각 정보를 일일이 검토하기는 어려운 일이다. 이를 보완하기 위해 자동화된 정보 여과 기술이 사용되는데 대표적인 방법들로 내용 기반 여과(content-based filtering)와 협력적 여과(collaborative filtering)가 있다. 이 중 협력적 여과 기술은 높은 질(quality)을 가지는 정보들을 여과할 수 있지만 해당 정보의 속성은 고려하지 않는다는 단점을 가지는데, 본 논문에서는 이를 보완하여 정보의 대표 속성을 중심으로 선호도 예측을 수행하는 개선된 협력적 여과 방법을 제안한다. 그리고 기존의 협력적 여과 기술들과 예측의 정확성에 대하여 성능 비교 실험을 수행함으로써 제안한 방법의 타당성을 제시한다.

1. 서론

자동화된 정보 여과 기술은 많은 정보들 중에서 사용자에게 필요한 정보만을 추출하여 추천함으로써 사용자의 수고를 덜어줄 수 있다. 현재까지 제안된 정보 여과 기술은 크게 내용 기반 여과와 협력적 여과로 나눌 수 있다.

내용 기반 여과는 기존의 내용 추출(information retrieval) 개념을 사용하는 방식으로, 정보의 다양한 속성(attribute)에 대한 사용자의 선호도들의 집합(profile)을 구성하여 정보의 내용과 사용자의 선호도를 비교함으로써 기호에 맞는 정보만을 여과해 낼 수 있다. 속성에 대한 사용자의 선호도는 사용자의 명시적인 제시에 의하거나 사용자의 간접적인 행동을 통해 반영될 수 있다. 이러한 내용 기반 여과는 특정한 단어나 속성 값을 가지는 정보를 여과하기에 매우 효과적이다. 그러나 여과된 정보들의 질을 판별하기에는 적절하지 못하다[1]. 또한 사용자 프로파일에는 없는 속성항목을 가지는 정보에 대한 여과는 수행하기 어렵다.

협력적 여과는 정보 자체에 대한 사용자의 선호도

들의 집합을 구성한 후 다른 사용자들의 선호도와 비교하게 되며 이를 통해 정보를 여과하게 된다. 즉 동일한 정보에 대한 선호도의 상관관계를 통하여 현 사용자와 기존 사용자들 사이에 기호의 유사성의 정도를 결정하게 되고, 기존 사용자들과의 유사성과 그들의 선호도에 기반하여 높은 선호도가 예상되는 정보만 현 사용자에게 여과하여 제공하게 되는 것이다.

이러한 협력적 여과는 자동화된 프로세스로는 쉽게 분석될 수 없는 정보의 질을 기존 사용자들의 선호도를 통해 어느 정도 반영한다는 장점이 있으나 정보의 속성에 대한 사용자의 선호도는 고려하지 않는다는 문제점을 가지고 있다.

따라서 본 논문에서는 정보의 대표적인 속성에 대한 사용자의 선호도를 협력적 여과에 반영함으로써 예측의 정확성을 향상시키는 개선된 협력적 여과 방법인 CFRA(Collaborative Filtering with Representative Attributes)를 제안하고 그 성능을 평가한다.

본 논문의 구성은 다음과 같다. 2장에서는 협력적 여과 방법에 대한 관련 연구들을 분석하고 3장에서는 본 논문에서 제시하는 개선된 협력적 여과 방법을 설

명한다. 4장에서는 성능 평가를 위한 실험 환경 및 실험 결과를 기술한다. 끝으로 5장에서는 결론을 맺는다.

2. 관련 연구

2.1. GroupLens

초기에 유즈넷 뉴스를 대상으로 개발되었으며 식 1의 Pearson 상관관계 계수를 통하여 사용자 a 와 사용자 k 사이의 유사성의 정도를 결정하고 이렇게 계산된 유사도를 이용하여 식 2와 같이 사용자 a 의 정보 i 에 대한 선호도의 예측을 하게 된다[2][3]. 식 1과 식 2에서 j 는 사용자 a 와 사용자 k 가 모두 선호도를 매긴 정보를 의미하고 $r_{a,j}$ 는 사용자 a 의 정보 j 에 대한 선호도, $r_{k,j}$ 는 사용자 k 의 정보 j 에 대한 선호도를 의미한다. 또한 \bar{r}_a 는 사용자 a 의 전체 정보에 대한 평균 선호도이고 \bar{r}_k 는 사용자 k 의 전체 정보에 대한 평균 선호도이다.

$$w_{a,k} = \frac{\sum_j (r_{a,j} - \bar{r}_a)(r_{k,j} - \bar{r}_k)}{\sqrt{\sum_j (r_{a,j} - \bar{r}_a)^2 * \sum_j (r_{k,j} - \bar{r}_k)^2}} \quad (1)$$

$$\hat{r}_{a,i} = \bar{r}_a + \frac{\sum_k w_{a,k} * (r_{k,i} - \bar{r}_k)}{\sum_k w_{a,k}} \quad (2)$$

2.2. Ringo

Ringo[4]는 음반과 서적 등을 추천하기 위하여 개발되었으며 Pearson 상관계수를 이용하여 현 사용자와 기존 사용자들간에 유사도를 결정 후 유사도가 일정 임계치를 넘는 기존 사용자들만 현 사용자의 이웃(neighborhood)으로 인정하고 이러한 이웃들만 식 2에 적용하여 선호도 예측을 한다. 또한 식 3과 같이 Pearson 상관계수에서 각 사용자의 전체 정보에 대한 평균 선호도 대신 선호도등급을 7단계로 구분하고 그 중간 값인 4를 일괄적으로 적용하는 Constrained Pearson 상관계수를 제안하였다.

$$w_{a,k} = \frac{\sum_j (r_{a,j} - 4)(r_{k,j} - 4)}{\sqrt{\sum_j (r_{a,j} - 4)^2 * \sum_j (r_{k,j} - 4)^2}} \quad (3)$$

2.3. Best n-neighbors

이 방법은 현 사용자와 기존 사용자들사이에 유사도를 결정 후 유사도가 가장 큰 n 명의 사용자들만 현 사용자의 이웃으로 인정하고 식 2를 통하여 선호도 예측을 하게 된다[5]. Pearson 상관계수로 유사도를 구했을 때 이를 Pearson_Max_n이라고 지칭하며, 특히 기존 사용자들을 모두 이웃으로 인정하는 경우는 Pearson_All_but_1으로 지칭한다[6].

2.4. Filterbot

유즈넷 뉴스에 대하여 정보의 특정 속성 값에 대하여 항상 정해진 선호도를 보이는 가상의 사용자들을 만들어 이들을 기존 사용자 그룹에 포함시켜 유사도를 계산하고 예측에 반영한다. 이러한 가상의 사용자로는 SpellCheckerBot, IncludedMsgBot, LengthBot이 있으며 이들은 각각 맞춤법, 첨부메시지 양, 기사의 길이에 따라 미리 정해진 선호도를 보인다[1].

3. CFRA

본 논문에서는 정보의 속성을 고려하지 않는 협력적 여과의 단점을 보완하여 좀더 효과적인 여과를 수행하기 위해 대표 속성(Representative Attribute)을 중심으로 선호도 예측을 수행하는 방법을 제안한다. 대표 속성이란 정보의 선호도에 가장 크게 영향을 미치는 속성을 의미한다. 대표 속성 값은 대표 속성이 가질 수 있는 값을 지칭한다.

식 1과 식 2를 살펴보면 기존의 협력적 여과방법은 사용자의 각 정보에 대한 선호도의 정도를 반영하여 예측을 수행하기 위해 전체 정보에 대하여 유사도를 계산하여 예측에 반영하게 된다. 그러나 전체 정보들을 모두 사용하여 유사도를 구하는 것은, 정보의 대표 속성 값들에 대해 사용자가 차별적인 선호도를 가지는 경우 이를 제대로 반영하지 못하는 단점을 가진다.

그러므로 본 논문에서는 이를 보완하기 위해 각 대표 속성 값에 한정하여 유사한 이웃을 찾아내 이를 예측에 이용하는 방법을 제안하였고 이를 CFRA 알

고리즘이라고 명하였다.

먼저 예측하려는 정보의 대표 속성 값과 같은 대표 속성 값을 가지는 정보들에 대해서만 유사한 기호를 가지는 이웃들을 뽑아낸다. 이를 위하여 식 1의 Pearson 상관계수 식을 식 4와 같이 변형하였다.

$$w'_{a,k} = \frac{\sum_s (r_{a,s} - \overline{r_{a,s}})(r_{k,s} - \overline{r_{k,s}})}{\sqrt{\sum_s (r_{a,s} - \overline{r_{a,s}})^2 * \sum_s (r_{k,s} - \overline{r_{k,s}})^2}} \quad (4)$$

식 4에서 s 는 예측하려는 정보의 대표 속성 값과 같은 대표 속성 값을 가지는 정보를 의미하고 $w'_{a,k}$ 는 예측하려는 정보의 대표 속성 값에 대한 사용자 a 와 사용자 k 사이의 기호의 유사도를 의미한다. $\overline{r_{a,s}}$ 는 정보 s 와 같은 대표 속성 값을 가지는 정보들에 대한 사용자 a 의 평균 선호도를 의미하고 $\overline{r_{k,s}}$ 는 정보 s 와 같은 대표 속성 값을 가지는 정보들에 대한 사용자 k 의 평균 선호도를 의미하는데 이를 사용한 이유는 평균 선호도도 예측하려는 정보의 대표 속성 값과 같은 대표 속성 값을 가지는 정보에 한하여 구하는 것이 의미가 있기 때문이다.

식 2의 경우도 식 5와 같이 변형되게 되는데, 식 4와 마찬가지로 평균 선호도의 대상이 예측하려는 정보의 대표 속성 값을 가지는 정보들로 한정되어 예상 선호도를 구하게 된다.

$$\hat{p}'_{a,i} = \overline{r_{a,i}} + \frac{\sum_k w'_{a,k} \times (r_{k,i} - \overline{r_{k,i}})}{\sum_k w'_{a,k}} \quad (5)$$

그러나 일반적으로 협력적 여과 방법에서 정보에 대하여 사전에 선호도 입력한 횟수가 모든 정보들에 대한 사용자들의 선호도 프로파일을 완전히 구성시킬 만큼 많지 않으므로, 이러한 대표 속성을 이용한 방법을 적용하게 되면 상대적으로 적은 자료에 근거하여 선호도 예측을 하게 되어 예측 값의 신뢰도가 떨어지게 된다. 그러므로 이를 해결하기 위하여 예측에 사용할 수 있는 자료의 수가 임계 값 이상일때만 변경한 식들을 적용하고 그렇지 않을 경우는 Pearson_Max_50의 예측 값을 따르게 하였다. 또한 변경한 식들을 적용할 경우 예측 값을 계산한 후 근거한 자료의 수에 따라 최종적으로 예측 값을 보정하게 된다. 이러한 CFRA 알고리즘을 정리하면 표 1과 같다.

```

If (# of the items rated by target user that have
representative attribute values equal to target value >=
δ 1)
For (each source user in source user set)
If (# of the items rated by source user that have
representative attribute values equal to target value
>- δ 1)
If (# of co items that have representative attribute values
equal to target value >- δ 2)
compute similarity of tendency between target user
and source user with Equation 4 ;
possible_neighbor_count1 ++ ;
Else
pa,i = predicted preference with Pearson_Max_50 ;
final predicted preference p'a,i = pa,i ;
end prediction ;
If (possible_neighbor_count1 >- δ 3)
p'a,i = predicted preference with Equation 5 using the
best fifty neighborhoods who are highly similar
to target user ;
pa,i = predicted preference with Pearson_Max_50 ;
possible_neighbor_count2 = the number of possible
neighborhoods with Pearson
Max 50 ;
final predicted preference
p'a,i =  $\frac{p'_{a,i} \cdot possible\_neighbor\_count1 + p_{a,i} \cdot possible\_neighbor\_count2}{possible\_neighbor\_count1 + possible\_neighbor\_count2}$  ;
Else
pa,i = predicted preference with Pearson_Max_50 ;
final predicted preference p'a,i = pa,i ;
end prediction ;
    
```

표 1 CFRA 알고리즘

- target user : 예측을 받는 사용자
- target item : 예측하려는 정보
- target value : 예측하려는 정보가 가지는 대표 속성 값
- source user : 기존 사용자
- source user set : 기존 사용자 군
- δ 1, δ 2, δ 3 : 임계 값

4. 실험 및 성능 평가

4.1. 실험 환경

실험에는 EachMovie[7] 데이터틀 사용하였다. 이 데이터에는 영화에 관한 사용자의 선호도가 0, 0.2, 0.4, 0.6, 0.8, 1.0 의 6개 단계의 수치로 표현되어 있다. 본 실험에서는 최소 100회 이상 선호도 입력을 한 사용자 4788명을 추출하여 이 가운데 1000명을 기존

사용자군으로 두고 나머지 사용자들 중에 무작위로 테스트 사용자 100명을 선택하여 총 1628개의 영화 중 테스트 사용자가 선호도를 표시한 임의의 5개의 영화에 대하여 선호도를 예측하고 실제 선호도와 비교하였다. 영화의 대표 속성은 장르로 가정하고 실험을 수행하였고 각 영화의 장르는 EachMovie 데이터에 구분되어 있는 것을 따랐다. 즉 장르는 액션, 애니메이션, 외국예술, 고전, 코미디, 드라마, 가족, 공포, 로맨스, 스릴러의 10가지 값 중 하나를 가질 수 있다.

4.2. 평가 기준

예측의 정확성을 평가하기 위하여 MAE(Mean Absolute Error)를 사용하였다. 아래 식은 MAE를 나타낸 것이며 여기서 N 은 총 예측 횟수, ϵ_i 는 예측된 선호도와 실제 선호도간의 오차를 나타낸다.

$$|\bar{E}| = \frac{\sum_{i=1}^N |\epsilon_i|}{N} \quad (6)$$

4.3. 실험 결과 및 분석

실험을 통하여, 제안한 방법과 기존의 Pearson_Max_50, Pearson_All_but_1을 사용한 방법간에 예측의 정확성에 대한 비교를 수행하였다. CFRA에는 튜닝을 통하여 임계치 $\delta_1=20$, $\delta_2=10$, $\delta_3=5$ 가 사용되었다.

실험 결과는 표 2와 같다. 표 2를 살펴보면 제안한 CFRA가 기존의 방법들보다 예측의 오차가 적음을 알 수 있다. 즉 CFRA를 적용하였을 때 가장 좋은 성능을 보임을 확인할 수 있다.

Methods	MAE
Pearson_All_but_1	0.217028
Pearson_Max50	0.208839
CFRA	0.207524

표2. 실험 결과 - 예측의 정확성 비교

5. 결론

본 논문에서는 정보의 속성을 고려하지 않는 협력적 여과를 개선하기 위해 대표 속성을 이용하여 각

대표 속성 값에 대한 선호도를 예측에 이용하는 CFRA 방법을 제안하였다.

제안한 여과 방법에 대해 예측의 정확성을 확인하기 위하여 영화 정보에 대하여 장르를 대표 속성으로 두고 실험을 수행하였으며 그 결과 기존 여과 방법들에 비해 성능의 향상이 있음을 확인할 수 있었다.

[참고문헌]

[1] Badrul M. Sarwar, Joseph A. Konstan, Al Borchers, Jon Herlocker, Brad Miller, and John Riedl, "Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system," *Proceedings of 1998 Conference on Computer Supported Collaborative Work*, 1998.

[2] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl, "GroupLens: An open architecture for collaborative filtering of netnews," *Proceedings of ACM CSCW'94 Conference on Computer Supported Cooperative Work*, pp. 175-186, 1994.

[3] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl, "GroupLens: Applying collaborative filtering to Usenet news," *Communications of the ACM*, 40(3), pp. 63-65, 1997.

[4] Upendra Shardanand and Patti Maes, "Social information filtering: Algorithms for automating "word of mouth"," *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, pp. 210-217, 1995.

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

[6] John S. Breese, David Heckerman, and Carl Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pp. 43-52, 1998.

[7] P. McJones, EachMovie collaborative filtering dataset, <http://www.research.digital.com/SRC/eachmovie/>, 1997.