스테레오타입 기반의 협업 여과 추천 시스템

이용준° 이세훈 이정현
한국전기연구원 지식정보그룹 인하공업전문대학 컴퓨터정보공학부 인하대학교 컴퓨터공학부
yjlee°@keri.re.kr, Seihoon@inhatc.ac.kr, jhlee@inha.ac.kr

A Recommender System using Collaborative Filtering with Stereotype Model

Yongjun Lee° Sehoon Lee Junghyun Lee
Korea Electrotechnology Research Institute, Knowledge & Information Service Group
School of Computer & Information Systems, Inha Technical College
School of Computer Science & Engineering, Inha University

요 약

본 논문에서는 협업 여과 추천의 사용자 정보 부족으로 발생하는 초기화 문제를 개선하기 위하여 스테레오타입 정보를 활용하여, 희소성 문제 해결 방안으로 스테레오타입 정보 기반의 사용자 성향 반영을 통한 계층적 구조를 가지는 가상 점수를 부여하여, 유사도 계산의 개선 및 추천의 정확도를 향상시킨다, 또한 항목의 속성을 분석하여 유사도가 높게 나타날 수 있는 항목을 선정하여 추천의 정확도를 향상시키고자 한다.

1. 서 론

협업 여과(collaborative filtering)는 다른 사람들의 평가를 의미적으로 수집하고 분석하여 정보를 찾는 시간을 줄이는 방법으로 많이 이용되고 있다. 그러나 초기 사용자에 대해서는 이웃과의 연관성을 계산할 수 있는 자료가 없다는 초기화 문제와 해당 영역의 사용자와 항목의 수가 커짐에 따라 이웃 계산의 정확도가 왜곡되는 희소성(sparcity)문제로 추천의 정확도가 낮아지므로 추천의 정확도를 향상시키기 위한 방안이 필요하다.

본 논문에서는 사용자 정보 부족으로 발생하는 초기화 문제를 개선하기 위하여 스테레오타입을 초기 사용자의 추천에 활용한다.

또한 희소성 문제를 해결하기 위하여 스테레오타입 정보를 이용, 사용자 성향 반영을 통한 계층적 구조를 가지는 가상 점수 부여를 통하여, 유사도 계산의 개선 및 추천의 정확도를 향상시키고, 항목의 속성을 분석하여 유사도가 높게 나타날 수 있는 항목을 선정하여 추천의 정확도를 개선하는 방안을 제안하고자 한다.

2. 관련 연구

추천시스템에 사용되는 추천 기법은 내용기반 여과(content-based filtering), 협업 여과(collaborative filtering)가 주류를 이루고 있다[1].

내용기반 여과는 사용자의 프로파일을 기반으로 사용자의 유형을 추출하고, 이를 이용하여 사용자의 유형에 맞는 제품을 추천한다. 그러나 사용자가 처음 시도하는 항목에 대해서는 설정된 정보가 없어 제품의 추천이 어렵다[2,3].

협업 여과는 내용기반여과의 문제점을 해결하기 위해 사이트나 개별 컨텐츠와 같은 대상이 되는 항목에 대하여 다른 사용자의 평가를 기반으로 사용자에게 추천을 생성하는 기술이다. 어떤 정보를 이미 보았거나 경험한 사람들의 행동과 의견을 가지고 그 정보를 아직 보지 못한 사람들에게 그 정보의 가치를 예측하여 주는 기법으로, 다른 사람들의 평가를 의미적으로 수집하고 분석하여 정보를 찾는 시간을 줄일 수 있다. 협업 여과는 Goldberg에 의해서 정보검색시스템에 적용하는 것을 시작으로 다양한 종류의 추천시스템에서 사용되고 있다[1,4,5,6,7].

사무 업무그룹과 같은 폐쇄그룹 사용자간의 정보 공유를 위하여 개발된 TAPESTRY[4], 유즈넷 사용자와 영화를 위한 익명의 협업 여과 기법을 제시한 GroupLens[5, 6], 음악 추천을 위한 Ringo[7]와 비디오 추천 시스템[8] 등 여러 분야에서 적용되고 있다.

그러나 협업 여과 기법에도 몇 가지 문제점이 있으며[9], 초기화와 희소성의 문제점은 그 대표적인 예이다[3,10].

초기화 문제(first rate problem)는 사용자가 평가한 항목이 없어 유사한 성격의 이웃을 선정하기 어려움에 따른 문제이며, 희소성에 대한 문제는 사용자가 평가한 항목이 너무 적어서 평가한 항목만으로는 이웃과의 유사도(similarity) 계산의 오차가 커서 추천의 정확도가를 저하시킨다. 따라서 이러한 추천의 정확도를 개선하기 위한 방안이 필요하다.

스테레오타입은 사람들에서 동시에 발생하는 특성을 모아서 표현하는 것으로 실질적인 적은 수의 관찰을 기반으로 많은 수의 그럴듯한 추론을 할 수 있게 한다. 유사한 사용자를 군집화하여 동일시하는 스테레오타입은 사용정보가 완전치 못할 경우, 유사한 정보를 활용하여 응답 시간을 축소하고, 추천의 정확도 향상에 활용할 수 있다[11].

희소성 문제는 기존의 유사도 계산에 사용되는 피어슨 상관관계 방식이 유사도 계산에 참여하는 사용자와 이웃이 모두 평가한 항목에 대해서 계산토록 되어 있어서, 둘 중 한 사람이 평가를 반영치 않은 경우 유사도 계산에서 제외되어, 사용자의 유사도 계산의 정확도가 저하됨에 따라 발생하는 문제이다. 유사도 계산이 정확하지 않음에 따라 예측 결과의 신뢰도는 낮아진다. 실질적인 문제점인 평가 점수 부재에 대한 문제를 해결키 위하여 가상 점수를 반영하는 방식의 연구가 진행되었으나 단순한 평균값을 이용한 방식이었다[12,13,14]

3. 스테레오타입 기반 협업 여과 추천 시스템

유사한 사용자를 군집화하여 동일시하는 스테레오타입은 사용자 정보가 완전치 못할 경우, 유사한 정보를 활용하여 응답 시간을 축소하고, 추천의 정확도 향상에 활용할 수 있으며, 인구 통계 정보를 기반으로 하고 있어 구성이 간편하다. 본 논문에서는 사용자 정보 부족으로 발생하는 초기화 문제를 개선하기 위하여 스테레오타입을 이용한 사용자 모델을 정의하고 초기 사용자의 추천에 이 정보를 활용한다. 희소성 문제를 해결하기 위하여 스테레오타입 정보를 이용, 사용자 성향

반영을 통한 계층적 구조를 가지는 가상 점수 부여를 통하여 추천의 정확도를 향상시키며, 항목의 속성을 분석하여 유사도가 높게 나타날 수 있는 항목을 선정하여 추천의 정확도를 개선하고자 한다.

3.1 가상 점수 반영

항목 i 에 대한 사용자 a 와 u 의 유사도 계산에 주로 사용되는 피어슨 상관 관계식은 식 (1) 과 같이 표현 된다.

$$w_{a,u}=\frac{\sum_{i=1}^{m}(r_{a,i}-\overline{r_a})*(r_{u,i}-\overline{r_u})}{\sigma_a*\sigma_u} \quad (1)$$

where $r_{a,i}$, $r_{u,i}$: voting value of item i from user a and u

$\overline{r_a}$, $\overline{r_u}$: voting average for user a and u

σ_a, σ_u : standard deviations for user a and u

유사도 계산시 사용자의 점수 부여가 적어 $r_{a,i}$ 가 있으나, $r_{u,i}$ 가 없는 경우 $r_{u,i}$ 를 가상 평가값인 $r'_{u,i}$ 값으로 대체하여, 식 (1)을 식 (2)로 변형하여 사용하여 유사도 계산의 정확도를 향상 시킬수 있다.[14]

$$w'_{a,u}=\frac{\sum_{i=1}^{m}(r_{a,i}-\overline{r_a})*(r'_{u,i}-\overline{r'_u})}{\sigma_a*\sigma_{u'}} \quad (2)$$

where $r'_{u,i}$: vitual voting value

$\overline{r'_u}$: average for user u based on virtual voting value

$\sigma_{u'}$: standard deviation for user u based on virtual voting values

예측 계산식은 식(3)과 같이 정의된다.

$$P_{a,i}=\overline{r_a}+\frac{\sum_{u=1}^{n}(r_{u',i}-\overline{r_{u'}})*w'_{a,u}}{\sum_{u=1}^{n}w'_{a,u}} \quad (3)$$

3.2 속성 가중치 반영

유사도 계산시 사용자의 스테레오타입 정보가 k개의 속성을 가진다면, $P=\{P1, P2, \cdots, Pk\}$ 이다. 예를 들어 k = 3 인 경우 $P=\{P1, P2, P3\}$이다.

학습자료를 기반으로 각각의 사용자의 속성 P_i 을 반영하여 계산된 예측치 P_{P_i} 와 사용자가 부여한 평가 점수 r 과의 오차를 E_{P_i} 라 하면 오차 E_{P_i}가 가장 적은 속성 P_i를 사용자의 대표속성으로 간주하면

$$E_{P_i}=P_{P_i}-r \quad (4)$$

따라서 사용자 대표속성을 다음과 같이 표현하였다.

사용자 대표속성 = 속성 P_i where min{ E_{P_i}} (5)

3.3 속성 대표값 반영

사용자 성향은 지속적으로 변화할 수 있어 사용자의 대표 속성도 사용자의 선호도 변화를 고려하여 지속적으로 관찰하고, 최종 n개의 대표속성은 Queue 에 저장하여 관리한다.

사용자의 속성 중 사용자의 속성을 가장 잘 나타내는 속성을 추적하기 위하여 사용자 속성을 반영한 계산 예측치와 실 부여 점수를 비교하여 어느 속성을 반영한 경우가 사용자의 속성을 가장 잘 반영하는지를 추적한다. 예를 들면 큐의 크기가 100이고, 반영할 속성이 3개인 경우, 큐에 저장되어 있는 내용이 속성1인 경우 5, 속성2인 경우 3, 속성3 인 경우 2라면 이 사용자의 최다 속성을 속성1로 한다.

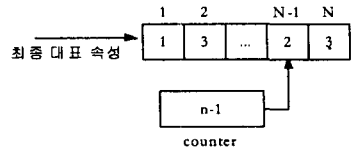

그림 1 Last-n 대표 속성

3.4 추천 항목의 클러스터링

협업 여과 방식의 추천 정확도에 영향을 미치는 중요한 요인 중 하나는 근접한 이웃을 찾는 것이다. 따라서 항목의 속성이 유사한 항목이 많을수록 유사한 사용자를 찾을 확률은 높아진다[15] 그러나 이러한 유사 속성을 따른 분류는 학습 자료를 축소하여 계산시간을 단축 시킬 수 있는 장점이 있으나, 경우에 따라서는 학습 자료의 축소로 인해 구성 자료가 너무 적어져 전체 항목을 대표하지 못해 실제적인 예측 계산 시 예측 오류가 커지게 된다. 따라서 충분한 학습 자료가 확보될 수 있는 경우에만 적용이 가능하다.

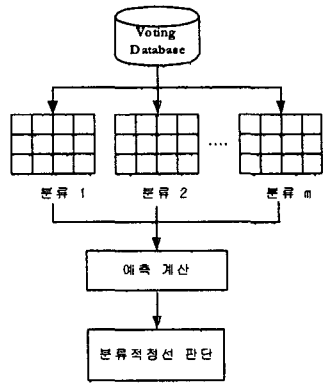

그림 2 항목 분류의 반영

전체 항목을C 라 하고 학습자료를 기반으로 이를 클러스터링 하여 $C=\{C_1, C_2, \cdots, C_i\}$ 으로 구성하면, 계산에 사용되는 행렬은 m_i*n_i where $m \le m_i$, $n \le n_i$ 로 적어진다. 전체 항목 C를 기반으로 계산된 오차를 E_c, 클러스터링을 기반으로 각각의 분류 항목으로 계산한 오차를 $E_{C_1}, E_{C_2}, \cdots, E_{C_i}$ 라 하면, 분류 항목이 $E_c > E_{C_i}$ 인 경우에는 분류 항목만으로 유사도를 계산하여, 예측 계산에 사용하여, 계산 속도 및 예측율을 향상 시켰다.

4. 실험 및 평가

실험은 GroupLens Research Project[16]에서 제공한 MovieLens 데이터 집합을 이용하여 실험을 하였다. 사용자의 스테레오타입 정보로는 user-id, age, gender, occupation, zip code 등이 포함되어 있으며, 영화는 19개의 장르로 구분되어 있고, 중복 장르 선택이 가능토록 되어있다. 총 100,000개의 데이터 집합 중에서 학습자료로 80,000개의 정보를 실험자료로 20,000개의 정보를 구분하여 사용하였다.

사용자 성향 분석에 따른 정확도는 실험 자료 20,000개의 자료를 사용하고 queue를 이용하여 사용자 속성을 동적으로 반영하였으며, 항목 분류에 따른 정확도는 실험 자료 중 10,000개의 자료를 이용하여, 항목의

특성을 분석하고, 분석된 결과를 기반으로 나머지 10,000개의 자료를 이용하여 사용자 속성 방식과 같이 진행하여, 예측 결과의 향상 여부를 확인하였다.

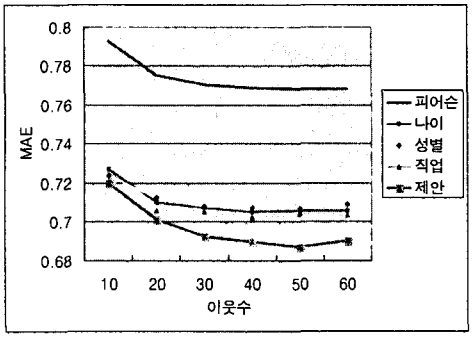

그림 3 사용자 정보 반영에 따른 실험 결과

사용자 성향 분석과 항목 분류를 반영한 예측 정확도를 비교한 것을 표 1 에 나타내었다. 피어슨 상관 관계를 이용한 협업 여과 방식이 이웃의 유사도를 근간으로 하고 있어, 분류 클러스터인 장르내의 이웃의 수가 적어지면 현저하게 예측 결과가 낮아짐을 알 수 있다.

표 1 항목 분류에 따른 예측 정확도의 변화

구분	사용자수	영화수	점수건수	Coverage	피어슨방식	제안방식
액션	938	251	20,566	0.984	0.755	0.682
모험	901	135	11,131	0.991	0.777	0.709
드라마	943	725	31,989	0.975	0.768	0.699
로맨스	943	247	15,547	0.988	0.803	0.736
공포	789	92	4,187	0.997	0.893	0.813
환타지	512	22	1,090	0.999	1.049	0.958

이러한 현상은 제안 방식에서도 결과가 동일하게 나타났다. 즉 가상 점수를 반영하여 이웃의 유사도를 향상시킬 수는 있으나, 클러스터내의 이웃의 수 자체가 너무 작으면 예측 정확도가 낮아짐을 확인할 수 있었다.

표 2 사용자 정보를 반영한 실험 결과(이웃의 수 n= 50 인 경우)

구분	MAE	ROC-4			
		Sensitivity	Specificity	Accuracy	Error Rate
피어슨	0.768	0.700	0.296	0.523	0.477
나이반영	0.706	0.769	0.344	0.583	0.417
성별반영	0.707	0.790	0.372	0.607	0.393
직업반영	0.705	0.767	0.339	0.580	0.420
성향반영	0.687	0.797	0.365	0.608	0.392
항목분리	0.675	0.798	0.368	0.610	0.390

항목 분류에 따른 효과가 있는 항목을 하나의 군으로 구성하여 분리하여 유사도를 계산하고, 항목 분리에 따른 효과가 없는 항목의 기존의 방식을 적용하는 방식도 예측 효과를 높일 수 있는 방법이다. 표 2 는 ROC-4 에 대한 비교이다. 개인성향과 항목을 분리하여 반영한 겨우 기존의 피어슨 방식에 비하여 MAE는 14%, ROC-4의 정확도는 17% 향상됨을 확인할 수 있었다.

5. 결론

사용자의 스테레오타입 정보를 이용하여 가상 점수를 반영하면, 기존의 피어슨 방식에 비하여 예측 결과가 향상됨을 알 수 있다. 그러나 어떤 인구 통계 정보를 적용할 것인가를 선택하는 작업은 문제 영역에 종속적이어서 매우 어렵다. 따라서 본 논문에서는 사용자 별로 예측 계산 결과에 영향을 주는 사용자의 속성을 분석하여 동적으로 반영하고, 이를 예측 계산에 반영할 수 있도록 지속적으로 사용자 정보를 갱신하는 방식을 적용하였다. 개인성향과 항목을 분리하여 반영한 경우, 기존의 피어슨 방식에 비하여 MAE는 14%, ROC-4의 정확도는 17% 향상됨을 확인할 수 있었다.

참고문헌

[1]Ansari,A., Essegaier,S. and Kohli,R., "Internet Recommendation Systems," Journal of Marketing Research Vol.37, pp. 363-375, 2000.

[2] Basu,C., Hirsh,H. and Cohen,W, "Recommendation as Classification : Using Social and Content-based Information in Recommendation," Proc. of the Fifteenth National Conference on Artificial Intelligence(AAAI-98), pp.714-720, 1998[4] Goldberg,D., Nichols,D., Oki,B.M., and Terry,D, "Using Collaborative Filtering to Weave an Information Tapestry", Communications of the ACM, 35(12) pp61~70, 1992.

[3] Pazzani,M.J. ,"A Framework for Collaborative, Content-Based and Demographic Filtering", Artificial Intelligent Review, pp394~408, 1999.

[4] Goldberg,D., Nichols,D., Oki,B.M. and Terry,D., "Using Collaborative Filtering to Weave an Information Tapestry," Communications of the ACM, Vol.35 No.12, pp. 61-70, 1992.

[5] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, K., and Riedl, J, "GroupLens:Applying Collaborative Filtering to Usenet News", Communications of the ACM, 40(3), pp77~87, 1997.

[6]Rensnick,P., Iacovou,N., Suchak,M., Nergstorm,P. and Riedl.,J. "GroupLens : An Open Architecture for Collaborative Filtering of Netnews," Proc. of CSCW '94, pp. 175-186, 1994

[7] Shardanand, U.; and Maes, P., "Social information filtering: Algorithms for automating word of mouth", Proceedings of ACM CHI'95, Conference on Human Factors in Computing Systems, pp210~217, 1995.

[8] Basu, C., Hirsh, H., and Cohen, W, "Recommendation as Classification : Using Social and Content-based Information in Recommendation", In Recommender System Workshop, pp11~15, 1998.

[9] Burke, R., "Hybrid Recommender Systems : Survey and Experiments," User Modeling and User Adapted Interaction. 12(4), pp 331~370, 2000.

[10] Ungar, L. H., and D. P. Foster. "Clustering methods for collaborative filtering. Recommender Systems," Paper for 1998 Workshop. Technical Report WS-98-08. AAAI Press, 1998.

[11] Rich, E. Users are Individuals: Individualizing User Models. International Journal of Man-Machine Studies 18: 199-214. 1993.

[12]Breese,J., Heckerman,D. and Kadie,C., "Empirical Analysis of Prediction Algorithms for Collaborative Filtering," Proc. of the 14th Conference on Uncertainly in Artificial Intelligence, pp.43-52, 1998

[13] Melville,P., Mooney,R., Nagarajan, R, " Content-Boosted Collaborative Filtering for Improved Recommendations" Proceedings of the eighteenth National Conference on Artificial Intelligence, pp187~192,2002.

[14] 이용준, 이세훈, 왕창종, "인구통계정보를 이용한 협업여과추천의 유사도 개선 기법," 정보과학회논문지-컴퓨팅의 실제 제9권 제5호, 2003.

[15] Sarwar,B., Karypis,G., Konstan,J. and Riedl,J., "Item based collaborative filtering recommendation algorithms," Proc. ofthe 10th International World Wide Conference, pp.285-295, 2001.

[16] http://www.grouplens.or