협업 필터링을 통한 IPTV 프로그램 자동 추천

*김은희, 김문철
한국정보통신대학교
e-mail : ehkim@icu.ac.kr, mkim@icu.ac.kr

Automatic Recommendation of IPTV Programs using Collaborative Filtering

*Eun-Hui Kim, Mun-Churl Kim
Information Communication University

Abstract

A large amount of efforts are required to search user's preferred contents for the program contents being provided by IPTV services. In this paper, using collaborative filtering, an automatic recommendation method of IPTV program contents is presented by reasoning similar group preferences on IPTV program contents which constitutes personalized IPTV environments. The proposed method models the user's preference of IPTV program contents with the program attributes such as content, genres, channels actor/actress, staffs and calculates it using the watching history of program contents in different genres and watching times. Also, the proposed method considers timely changing user's preference and the preference oon the content itself, which improves the traditional collaborative filtering methods that can not recommend the non-consumed items.

I. 서론

본 논문은 협업 필터링 방법으로 프로그램 시청 선호도가 유사한 그룹들의 관심도 추론하여 프로그램 자동 추천 방법을 제시하고 이를 이용한 개인의 선호의 IPTV 프로그램 시청 환경을 제공하는 것을 제안한다. 본 논문에서 사용자의 IPTV 프로그램 선호도를 프로그램 속성(프로그램 자체, 장르, 채널, 배우, 출연진)으로 모델링 하고 사용자의 시청 프로그램 장르, 시청 시간 정보를 이용하여 사용자의 프로그램의 콘텐츠에 대한 관심도를 측정하는 방법을 제시한다.

또한 시청자들의 취향이 시간에 따라 변하는 것을 관심도 계산에 고려하고, 프로그램 콘텐츠 자체 속성별 관심도를 함께 고려함으로써 기존 추천 시스템이 시청자들이 관심을 보이지 않은(비구매) 프로그램 콘텐츠(아이템)에 대한 추천이 이루어지지 못했던 기존의 방법을 개선한다.

II. 본론

가. 시청자의 개별 아이템에 대한 선호도 측정

1. 제안된 프로그램 소비 시간을 고려한 선호도 계산

시청자 k의 프로그램(아이템) m의 자체 선호도 $a_k^1(m)$는 시리즈방영물의 경우, 한 회분의 시간과 전체 방영횟수에 대해 시청자가 시청한 시청길이의 합으로 관심도를 측정하되, 현재 날짜와t_{now} 시청날짜 t차이가 크면 작은 weight값을 주도록 (1)식을 적용한다.

$$a_k^1(m)=\frac{\sum_{t\in t_k} \text{시청길이}_k(m) \bullet e^{-\frac{|t_{now}-t|}{|window|}}}{\text{한회분방영길이}(m) \bullet \text{횟수}(m)} \quad (1)$$

즉, $|t_{now}-t|$가 window 시간크기이면, 순수 시청합계의 $e^{-1}\approx 0.3$의 weight값을 적용한다.

2. 제안된 아이템 속성을 고려한 관심도 측정

시청자 k의 아이템 m에 대한 총 관심도 $x_k(m)$는 아이템 속성(프로그램 자체, 장르, 채널, 배우, 출연진) j별 시청자의 선호도 $a_k^j(m)$의 가중치 합으로 계산된다.

$$x_k(m)=\sum_j \delta_j a_k^j(m)=\delta_1 a_k^1(m)+\cdots+\delta_N a_k^N(m) \quad (2)$$

여기서 $\sum_j \delta_j = 1$ 이며, $0 \le a_k^j(m) \le 1$이다. 만일, 시청자 k의 아이템 m에 대한 총관심도가 두 개의 속성, 아이템

자체 선호도($a_k^1(m)$)와 아이템 장르 선호도($a_k^2(m)$), 채널 선호도($a_k^1(m)$)로 표현하면 총 관심도는 다음과 같다.

$$x_k(m) = \delta_1 a_k^1(m) + \delta_2 a_k^2(m) + \delta_3 a_k^3(m) \quad (3)$$

장르 g_s 선호도 $a_k^2(m \in g_s)$은 각 장르에 속한 아이템 자체 선호도의 합으로 주어지며, 최대 장르 선호도 값으로 normalize된다. 따라서 장르 선호도 값도 0과 1사이의 값을 가진다. 즉, (4) 와 같다.

$$a_k^2(m \in g_s) = \frac{\sum_{m \in g_s} a_k^1(m)}{\arg \max_{j \in \{1,2,..N\}} \left(a_k^2(m \in g_j)\right)} \quad (4)$$

나. 사용자 프로파일 기반 관심도 유사 시청자 추출

본 논문은 semantic 시청자 프로파일(성별/나이)정보를 활용하여 이미 cluster화 되어있는 있는 그룹 내에 active user u_a와 시청 관심도 $x_k(m)$가 유사한 user들을 추출하였다. active user u_a와 user u간의 유사도$S_{u_a,u}$는 Pearson Correlation Coefficient를 사용하였다.

$$S_{u_a,u} = \frac{\sum_{i \in L(u_a)} \left(x_u(i) - \overline{x_u}\right)\left(x_{u_a}(i) - \overline{x_{u_a}}\right)}{\sqrt{\sum_{i \in L(u_a)} \left(x_u(i) - \overline{x_u}\right)^2} \sqrt{\sum_{i \in L(u_a)} \left(x_{u_a}(i) - \overline{x_{u_a}}\right)^2}} \quad (5)$$

다. 아이템별 사용자 관심도 모델링의 적용

논문[1]에서 active user u_a의 item i_m에 대한 관심도 측정 식(6)을 적용하여 모델링 하였다.

$$RSV_{u_a}(i_m) = \sum_{\substack{i_b \in L(u_a) \\ \cap\, c(i_b,i_m) > 0}} \log\left(1 + \frac{(1-\lambda)P(i_b|i_m,r)}{\lambda P(i_b|r)}\right) + \log P(i_m|r) \quad (6)$$

여기서, λ는 Maximum likelihood model $P(i_b|i_m,r)$ 과 background model $P(i_b)$사이의 smoothing 변수이다. 이때, 각 식의 조건 부 확률 계산은 해당 아이템에 관심을 보인user의 수를 count하는 방법을 사용하였다. 즉 식 (8)에서 분자 $C(i_b,i_m)$은 i_b와i_m 두 아이템에 동시 관심을 보인 user수를 합한 개수를 이야기 한다. 식 (7)의 분모인 C(r)은 Similarity 측정 결과 관심을 보인 아이템이 유사한 Top K 사용자 수를 카운트 한 수를 의미한다.

$$P(i_b|r) = \frac{C(i_b,r)}{C(r)} \quad (7)$$

$$P(i_b|i_m,r) = \frac{C(i_m,i_b)}{C(i_m)} \quad (8)$$

Ⅲ. 실험

본 실험은 2002/12/1~2003/5/31 6개월간 1995명의 user들의 6개의 채널, 2429개의 프로그램 시청을 조사한 Nelson Korea 자료를 이용하였다. active user 선별 후, 2003/3/31일을 시청 일로 선정, 2002/12/1~2003/3/31기간을 Training Set으로 사용하고, 2003/3/31이후 1달 데이터를 Test Set로 사용하였다.

성능 분석은 Precision, 즉 추천아이템 중 실제 사용자가 관심을 보인 아이템의 비율을 사용하였다. 그림 1과 같이 $(\delta_1,\delta_2,\delta_3) = (0.65,0.25,0.1)$서 성능향상을 확인하였다.

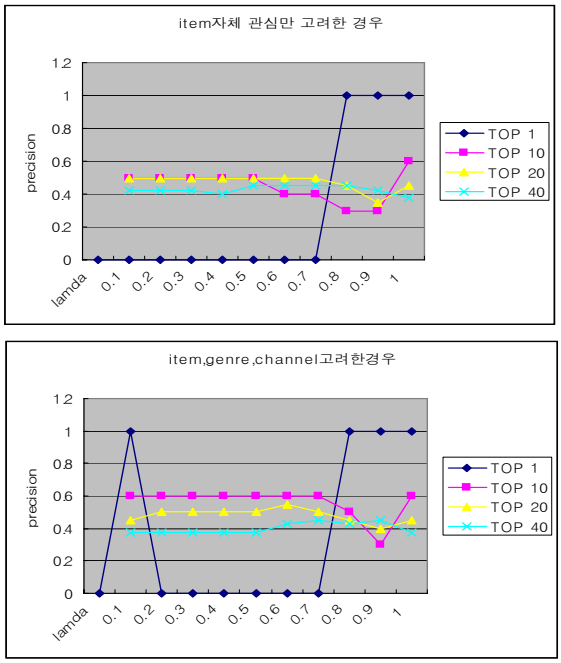

그림 1 아이템 속성을 고려한 경우 성능향상

또한, 시청 기준 일을 2003/04/30로 하면, 시청 기준 일을 2003/3/31로 선정한 것에 비해 Windowing 효과(식(1))를 확인할 수 있었다. 즉, 시청 분석 자료 수집 기간이 길어지면 시간에 따른 Weighting 방법이 효과를 보인다.

Ⅳ. 결론 및 향후 연구 방향

실험을 통해, program의 속성별, 시청 시간의 window를 적용할 때, 추천 아이템의 개수가 많아짐에도 불구하고 사용자의 개인선호도에 일치하는 아이템들이 잘 선별되는 결과를 얻을 수 있었다. 향후 추가 속성(채널, 배우, 출연진)을 고려하고, 요일별 분석을 통한 정교한 분석을 추가할 예정이다. 실시간 방송 데이터 외에 VOD형 콘텐츠에 대한 분석을 진행할 필요가 있다.

참고문헌

[1] *J. Johan Pouwelse , Jenneke Fokker, Arjen P. de Vries & Marcel J. T. Reinders, 13 Jan 2007, LLC 2007, Personalization on a peer-to-peer television system.*