정규논문 (Regular Paper)
방송공학회논문지 제18권 제1호, 2013년 1월 (JBE Vol. 18, No. 1, January 2013)
http://dx.doi.org/10.5909/JBE.2013.18.1.88
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

VOD 서비스 플랫폼에서 협력 필터링을 이용한 TV 프로그램 개인화 추천

한 성 희[a)‡], 오 연 희[a)], 김 희 정[a)]

Personalized TV Program Recommendation in VOD Service Platform Using Collaborative Filtering

Sunghee Han[a)‡], Yeonhee Oh[a)], and Hee Jung Kim[a)]

요 약

개인화된 추천을 제공하기 위한 협력 필터링은 추천 시스템에서 성공적으로 활용되어 온 기법이다. 그러나 협력 필터링이 주로 연구 및 적용된 분야들은 사용자로부터의 명시적 피드백이 존재하는 독립된 아이템들을 추천하는 것에 초점을 두고 있다. VOD 서비스 플랫폼에서 개인화된 TV 프로그램을 추천하기 위해서는 해당 도메인의 특성과 제한들을 고려하는 것이 필요하다. 본 논문에서는 TV 프로그램의 시리즈 속성을 이용하여, 선호를 판단하기 힘든 비명시적 피드백인 회별 프로그램 시청기록을 명시적이고 지속적인 프로그램 선호도로 변환하는 방법을 고안하였다. 데이터 수집과 최종 추천은 회별 프로그램 단위로 이루어지면서 협력 필터링 처리 단위는 프로그램으로 변경되어 TV 프로그램 VOD 추천 환경에 가장 적당한 형태로 협력 필터링을 변형 적용하였다. 실험 결과는 고안된 추천 시스템이 단순히 협력 필터링을 적용했을 때보다 높은 정확도와 더 적은 계산량을 가지는 것을 보여준다. 도메인 특화된 이러한 변형은 추천 시스템의 알고리즘 모듈로 구성되어 기존에 알려진 다양한 협력 필터링 기법과 결합하여 사용될 수 있다.

Abstract

Collaborative filtering(CF) for the personalized recommendation is a successful and popular method in recommender systems. But the mainly researched and implemented cases focus on dealing with independent items with explicit feedback by users. For the domain of TV program recommendation in VOD service platform, we need to consider the unique characteristic and constraints of the domain. In this paper, we studied on the way to convert the viewing history of each TV program episodes to the TV program preference by considering the series structure of TV program. The former is implicit for personalized preference, but the latter tells quite explicitly about the persistent preference. Collaborative filtering is done by the unit of series while data gathering and final recommendation is done by the unit of episodes. As a result, we modified CF to make it more suitable for the domain of TV program VOD recommendation. Our experimental study shows that it is more precise in performance, yet more compact in calculation compared to the plain CF approaches. It can be combined with other existing CF techniques as an algorithm module.

Keywords : Recommender System, TV Program, VOD, Collaborative Filtering, Implicit Feedback

Ⅰ. 서 론

오늘날의 방송사들은 전파를 통한 동시 방송이라는 고유의 전송 형식을 넘어서 다양한 IP 기반의 서비스 플랫폼을 통하여 사용자들에게 콘텐츠를 서비스하고 있다. 방송사는 정해진 채널 편성표에 따라 매일 새로운 콘텐츠를 생성해 내며 이렇게 생성된 방송 콘텐츠는 실시간으로 방송 또는 스트리밍 방식으로 서비스되고 방송 종료 후 VOD의 형태로 지속적으로 소비된다. 방송사는 콘텐츠 소비 형태와 단말의 종류에 따라 여러 서비스 플랫폼을 운용한다. 이렇게 매일 쏟아져 나오는 많은 양의 콘텐츠를 서비스하는 플랫폼에서는 사용자의 취향을 고려하여 통찰력 있는 콘텐츠 배치를 제공하는 것이 중요할 수밖에 없다. 따라서 KBS에서도 방송 콘텐츠의 메타데이터와 사용자의 소비 이력을 분석한 방송 콘텐츠 추천 검색 시스템[1]을 구축 중에 있다. 방송사는 방송 프로그램에 대하여 상대적으로 잘 정립된 콘텐츠 메타데이터 체계와 데이터를 보유하고 있으며 이러한 콘텐츠 메타데이터는 추천에 있어 콘텐츠의 특성을 발견함에 있어 중요한 데이터로 사용된다. 그러나 콘텐츠와 사용자 취향의 숨은 연관성을 찾아내는 데 있어 콘텐츠 메타데이터만으로는 충분하지 못한 경우가 있으며 사용자 소비 이력 정보를 활용하는 것이 개인화된 추천 시스템의 정확도 향상을 가져올 수 있다[3]. 사용자 소비 이력 정보를 활용하여 사용자와 아이템 쌍에 대하여 숨겨진 연관도를 구하는 연구는 크게 이웃 모델(Neighborhood Model)에 기반한 협력 필터링(Collaborative Filtering)과 잠재 요소 모델(Latent Factor Model)에 속하는 다양한 접근법들이 있다[2]. 두 모델 모두 사용자가 독립적인 아이템에 대하여 명시적으로 내린 평가(Explicit Rating) 데이터를 기반으로 동작한다. 그러나 KBS의 VOD 서비스 플랫폼에서 TV 프로그램들을 추천하는 환경에서는 위의 전제와 다른 특징과 제한이 존재한다.

콘텐츠의 특징으로는 거의 모든 콘텐츠가 [프로그램 - 회별 프로그램]이라는 시리즈 구조로 되어 있다는 점이다.[1] 한 시리즈 내의 모든 에피소드들은 정도 차이는 있으나 시리즈 내에서 공통적으로 갖는 일관성이 존재한다. 그 공통점은 줄거리의 연속성, 동일 출연자, 프로그램 의도와 같은 요소일 수 있으며 시리즈가 매회 다른 출연자에 다른 주제를 다루고 있을 때조차도 매회의 에피소드들은 전체 시리즈에 대한 일관성을 가진다. 예를 들면, '인간 극장'이라는 프로그램은 매회 다른 출연자들이 각각 자신의 삶에 대한 다양한 이야기를 풀어놓지만 전체적으로 타인의 삶에 대한 관심과 공감에 초점을 둔다는 면에서 일관성을 가진다. 시리즈 구조는 또한 사용자들의 시청 패턴에도 영향을 미친다. 실시간 방송일 경우는 말할 나위 없이, VOD 서비스 플랫폼에서도 보통 시청자들은 자신이 본 회별 프로그램의 다음 순서를 순차적으로 보는 것을 가장 선호하게 된다. 즉, 콘텐츠의 시리즈 구조는 기존에 소비이력 기반의 기법들이 전제하는 독립적인 아이템에 대한 가정과는 다른 조건을 갖게 된다.

서비스 플랫폼 특이성에 의한 제한은 사용자들로부터 받을 수 있는 아이템에 대한 평가가 비명시적인 이진값(Implicit & Binary Data)이라는 사실이다. 현재 KBS가 VOD 콘텐츠를 제공하는 서비스로는 KBS 홈페이지, 플레이어 K, 콘팅 등이 있으며 각 서비스들은 모두 사용자 평가와 같이 적극적인 데이터 수집 과정이 존재하지 않는다. 사용자가 가입시와 프로그램 시청 후에 사용자로부터 적극적으로 선호에 대한 데이터를 수집하는 것은 사용자의 선호에 대한 모호성을 쉽게 없애주는 장점이 있지만 사용자를 번거롭게 하는 UI에 의한 거부감을 불러 일으킬 수 있다. 따라서 추천 시스템에서 활용할 수 있는 사용자 피드백은 어떤 사용자가 어떤 콘텐츠를 소비했는지의 유무에 대한 것만 존재한다. 또한 VOD 플랫폼에서는 매일 생성되는 새로운 콘텐츠들이 서비스 시스템에 업데이트되어 실시간 TV 추천과는 달리 방송 날짜로부터 상당 기간 동안 지속적

a) KBS 기술연구소 (KBS Technical Research Institute)
‡ Corresponding Author : 한성희 (Sunghee Han)
E-mail: shhan9@kbs.co.kr
Tel: +82-2-781-5232 Fax: +82-2-781-5299
· Manuscript received October 31, 2012 Revised December 14, 2012 Accepted December 24, 2012

1) 향후 이 논문에서 "프로그램"은 시리즈를 가리키며 "회별 프로그램"은 에피소드를 가리킨다. KBS에서 생산되는 모든 콘텐츠들은 프로그램과 회별 프로그램에 대하여 고유한 식별값을 메타데이터로 가지고 있다.

으로 서비스된다.

본 논문에서는 개인화된 추천을 위하여 사용자 소비 이력을 이용하는 기법 중 가장 널리 활용되는 아이템 기반 협력 필터링(Item Based CF : IBCF)을 기본적으로 사용하되, 위와 같은 추천 도메인에 대한 분석 결과에 특화된 방식으로 적용한다. 협력 필터링은 사용자들의 과거 소비 이력을 바탕으로 유사한 사용자나 아이템을 도출함으로써 사용자의 미래의 선호를 예측하는 방법이다. 유사한 아이템을 도출하는 방식인 아이템 기반 협력 필터링은 유사한 사용자를 도출하는 방식인 사용자 기반 협력 필터링(User Based CF : IBCF)에 비하여 더 높은 정확도와 계산 속도를 보이는 것으로 알려져 있다[4]. 통상적으로 협력 필터링은 2단계로 이루어지는데 유사도 계산부와 추천을 위한 콘텐츠 리스트를 생성하는 부분으로 나뉜다[2]. 본 논문에서는 유사도를 계산하기 전 단계에 피드백 데이터를 전처리하는 부분을 고안하였고, 해당 데이터 전처리에 맞추어 추천 콘텐츠 리스트를 생성하는 단계에 변형을 가하였다. 데이터 처리 방식의 변형으로 계산량이 줄어드는 것을 확인할 수 있으며 실험을 통하여 정확도 향상을 검증한다.

II. 시스템 데이터 특성 분석

[그림 1]은 사용자들의 콘텐츠 소비 내역을 협력 필터링을 적용하기 위하여 사용자-아이템 행렬로 구성한 모습이다. 행렬의 요소인 $r(u,i)$는 u라는 사용자가 i라는 회별 프로그램을 시청했을 때 1의 값을, 시청하지 않았을 때 0의 값을 가진다. 이러한 행렬을 기반으로 아이템 유사도를 계산하게 되면 유사도 측정 지표가 이진값을 위한 지표들로 제한을 받게 된다. 또한 별점과 같이 사용자들로부터 다양한 스케일의 값을 피드백으로 받아들이게 되면 평가가 없는 부분은 값은 결측값(Missing Value)으로 처리하여 유사도에 영향을 미치지 않을 수 있으나 이진값인 경우에는 0은 부정적 평가로 처리될 수 밖에 없다. 따라서 [그림 1]과 같이 각 회별 프로그램에 연결되어 있는 시리즈 정보를 사용하여 아이템의 단위를 회별 프로그램에서 프로그램으로 변경하면 아이템 공간의 차원(De에서 Ds로 차원 변경)을 줄일 수 있을 뿐만 아니라 해당 프로그램에 대하여 변별력 있는 선호도를 추론하는 것이 가능할 수 있다. 이것은 서론에서 언급한 것과 같이 프로그램이 하위의 각 회별 프로그램에 대하여 가지는 일관성에 대한 가정에 기인한다.

가장 쉽게 프로그램에 대한 사용자의 선호도를 추론할 수 있는 방법은 사용자가 얼마나 자주 해당 프로그램의 회별 프로그램들을 봤는지를 이용하는 것이다. TV 프로그램에 대한 추천을 다루는 논문들에서 TV 프로그램의 시리즈 특성 때문에 이와 유사한 접근법을 취하는 연구들이 있었다. 사용자가 한 프로그램에 서비스된 회별 프로그램들 중 시청한 것의 비율을 해당 사용자의 전체 이력 기록 길이로 나누어서 프로그램 선호도를 추론하기도 하였고[5], 사용자가 해당 프로그램을 얼마나 자주 봤는지를 프로그램 선호에 대한 확신도(Confidence)를 계산하는 데 활용하기도 하였다[6].

그러나 실제 각 TV 프로그램들이 포함하고 있는 회별 프로그램들의 수는 아주 다양하다. 단발성 특집 프로그램부

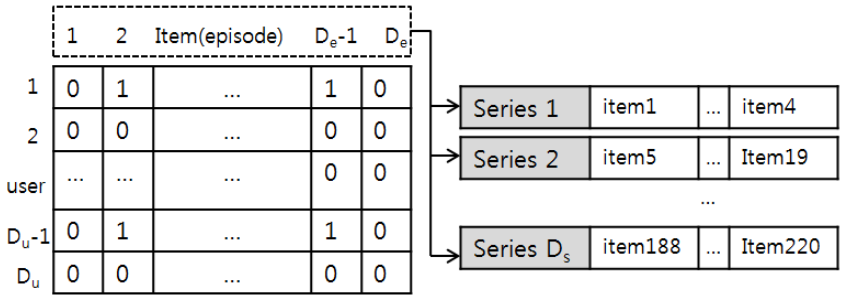

그림 1. 협력 필터링을 위하여 사용자-아이템 행렬로 표현된 소비 이력과 회별 프로그램의 시리즈 정보
Fig. 1. User-Item matrix for CF and Episode-Series mapping

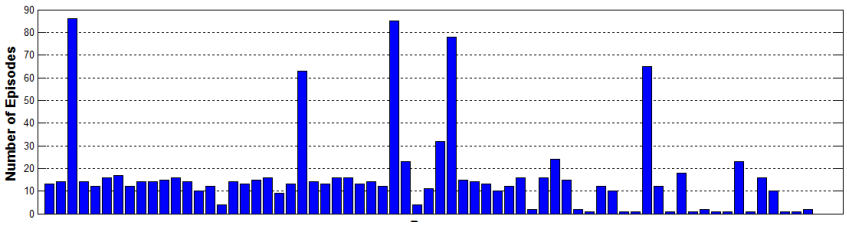

그림 2. 2개월간 콘팅 플랫폼에서 서비스된 프로그램들의 '회별 프로그램 수' 분포
Fig. 2. Distribution of 'number of episodes per a program' for 2 months in Conting platform

터 몇 년에 걸쳐서 방송되는 프로그램들이 존재한다. 물론 아이템 유사도를 계산하기 위하여 사용자 로그 분석 기간을 제한하여 유사도를 계산하기 때문에 로그 분석 기간 안에 서비스되는 콘텐츠들의 분포만 고려하게 된다. [그림 2]는 KBS 콘팅 서비스[2]에서 2010년 3월~4월의 2개월간 제공되는 VOD들의 프로그램 분포에 대한 것이다. VOD로 제공되는 콘텐츠들은 실시간으로 방송된 프로그램들 중 인기 있는 것들로 구성된다. 2개월간 총 67개의 프로그램에 대하여 1067개의 회별 프로그램이 제공되었으며 한 개의 프로그램 당 작게는 1개, 크게는 86개의 회별 프로그램이 서비스되었으며 평균은 16개이다.

위와 같은 TV 프로그램 분포로 볼 때, 시리즈 구조의 콘텐츠 시청 내역을 평가 점수로 변환하는 것에 고려해야 할 사항이 있다. 서비스 기간 내에 작은 수의 회별 프로그램 목록만 존재하는 프로그램의 경우 시청 내역으로 해당 프로그램에 대한 명시적 선호를 판단할 수 있는 여지가 부족하다는 것과 매우 많은 회별 프로그램 목록을 보유하고 있는 프로그램의 경우 뚜렷한 선호를 가지는 가지고 있는 사용자라고 할지라도 모든 회차를 거의 다 시청하기는 힘들다는 점이다. 따라서 이러한 점을 고려하여 비명시적 이진 시청 내역을 명시적 평가 점수와 같은 형태로 변환해야 한다.

III. VOD 서비스 플랫폼에서의 TV 콘텐츠 추천 시스템

1. 데이터 전처리 및 프로그램 유사도 처리

Ⅱ에서 분석한 데이터 특성을 기반으로 사용자 피드백 데이터를 모델링한다. [그림 1]의 사용자의 회별 프로그램 시청 기록인 r(u,i)를 사용자 프로그램 선호점수인 rp(u,i)로 변환하는 것이 목적이다. [그림 1]과 같이 사용자, 회별 프로그램, 프로그램에 대한 차원이 각각 Du, De, Ds일 때 변환에 필요한 값들을 정리하면 아래와 같다.

사용자가 해당 프로그램의 회별 프로그램들을 얼마나 지속적으로 봤는지를 반영하면 식 (1)과 같아진다.

$$r_p(u,i) = \frac{n_p(u,i)}{N_p(i)} = \frac{\sum_{j \in p}^{De} r(u,j)}{N_p(i)} \quad (1)$$

표 1. 사용자의 프로그램 선호 점수 도출을 위한 시청이력과 프로그램 정보
Table 1. User's consumption history data & program information to draw the user's preference for program

• 사용자 u의 회별 프로그램 i에 대한 시청 여부	$r(u,i)$	$1 \le u \le D_u$ $1 \le i \le D_e$
• 사용자 u가 시청한 모든 회별 프로그램 수	$L(u)$	$1 \le u \le D_u$ $1 \le i \le D_s$
• 프로그램 i가 포함하는 회별 프로그램 수	$N_p(i)$	
• 사용자 u가 특정 프로그램 i의 회별 프로그램을 시청한 수	$n_p(u,i)$	
• 사용자 u의 프로그램 i에 대한 선호 점수	$r_p(u,i)$	

2) http://conting.conpia.com

Ⅱ장에서 분석한 내용 중 $N_p(i)$가 아주 작은 경우에는 식 (1)로 선호를 판별할 수 없는 경향을 고려한 가중치 $w_1(N_p(i))$과 매우 많은 회별 프로그램 목록을 보유하고 있는 프로그램의 경우에는 사용자 시청 리스트의 길이 $L(u)$ 또한 고려해야 하는 점을 반영한 가중치 $w_2(N_p(i),L(u))$를 식 (1)에 사용하여 해당 요소들의 영향을 보정할 수 있다.

$$r_p(u,i) = \frac{n_p(u,i)}{N_p(i)} w_1(N_p(i))\, w_2(N_p(i), L(u)) = \frac{n_p(u,i)}{N_p(i)} w(N_p(i), L(u)) \quad (2)$$

$w_1(N_p(i))$는 오직 해당 프로그램이 보유하는 회별 프로그램 수에만 의존하는 값으로 회별 프로그램 수가 어느 정도의 값을 넘으면 프로그램의 회별 시청 비율만으로 프로그램 선호도를 유추할 수 있다는 가정을 반영한다. $w_2(N_p(i),L(u))$는 사용자 시청 리스트의 길이가 해당 프로그램의 전체 회별 프로그램 수보다 작을 때는 사용자 시청 리스트에서 해당 프로그램이 차지하는 비율을 주로 고려하고, 반대의 경우에는 해당 프로그램의의 얼마나 많은 회별 프로그램을 시청했는지 만으로 프로그램 선호도를 유추할 수 있다는 가정을 반영한다. 수식으로 표현하면 (식) 2와 같다.

$$w_1(N_p(i)) = 1 - e^{-\lambda N_p(i)},$$
$$w_2(N_p(i), L(u)) = \begin{cases} \frac{N_p(i)}{L(u)}, & \text{if } L(u) < N_p(i) \\ 1, & \text{if } L(u) \geq N_p(i) \end{cases} \quad (3)$$

가중치 $w(N_p(i), L(u))$ 값을 해당 프로그램의 보유 회별 프로그램 수 $N_p(i)$과 사용자 시청 리스트 길이 $L(u)$에 대하여 도시해 보면 아래 [그림 2]와 같은 그래프를 얻을 수 있다. (λ는 적당히 설정할 수 있는데, [그림 3]에서는 2부작의 경우 해당 프로그램의 2회를 모두 다 봤다고 해도 선호도를 판단할 수 있는 확신도가 반 정도라는 가정으로 $\lambda = \frac{\log 2}{2}$로 설정했다.)

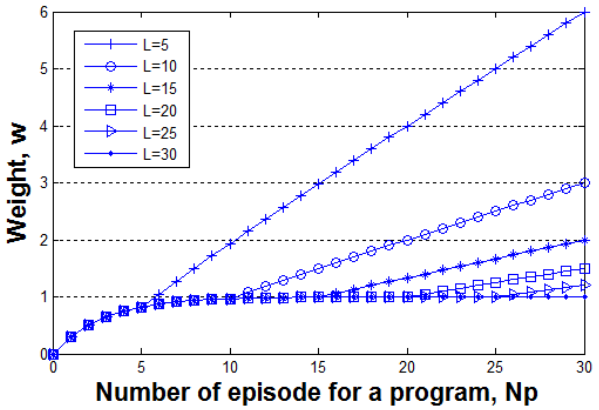

그림 3. 사용자 프로그램 선호점수 변환을 위한 가중치
Fig. 3. Weight for conversion to program rating

이제는 프로그램에 대한 명시적인 사용자 선호 점수를 얻게 되었으므로 Du × Ds 의 차원을 가지는 (사용자 - 아이템) 행렬 Rp에 협력 필터링을 적용할 수 있다. 행렬 Rp를 구성하는 값들은 0과 1 사이의 임의의 값이므로 0을 비선호로 고정하여 이진값에만 적용할 수 있었던 타니모토(Tanimoto) 계수나 코사인 유사도와 같은 유사도 척도[7]의 제한에서 벗어날 수 있다. 협력 필터링에서 쓰이는 일반적이고 다양한 유사도 함수[2]들이나 새로운 사용자에 대한 초기 추천 문제(Cold Start Problem) 같은 특별한 목적을 가진 유사도 함수[8]들을 적용할 수 있다. 아이템 기반의 협력 필터링을 적용하므로 각각 다른 프로그램들 사이의 유사도를 구한다[4]. 이 논문에서는 이 중 기본적인 코사인 유사도(Cosine Similarity : COS)와 피어슨 유사도(Pearson Correlation Coefficient : PCC)를 적용해 본다. 각 유사도 지표는 식(4)와 식(5)와 같다. 코사인 유사도나 피어슨 유사도를 계산할 때 0은 결측값으로 처리하였다.

$$COS(\vec{i}, \vec{j}) = \frac{\vec{i} \bullet \vec{j}}{\| \vec{i} \|_2 * \| \vec{j} \|_2} \quad (4)$$

$$PCC(i,j) = \frac{\sum_{u \in U} (R_p(u,i) - \overline{R_p(u,i)})(R_p(u,j) - \overline{R_p(u,j)})}{\sqrt{\sum_{u \in U} (R_p(u,i) - \overline{R_p(u,i)})^2} \sqrt{\sum_{u \in U} (R_p(u,j) - \overline{R_p(u,j)})^2}} \quad (5)$$

행렬 Rp는 아이템의 차원이 De에서 Ds로 줄어들었기 때문에 계산량이 많은 유사도 처리의 계산 복잡도(Computational Complexity)는 $O(D_e^2)$에서 $O(D_s^2)$로 줄어든다.

2. 추천 콘텐츠 리스트 생성

Ⅲ.1에서 계산한 아이템 유사도 행렬 S의 요소인 $S(i,j)$는 i번째 프로그램과 j번째 프로그램의 유사도이다. 유사도 행렬과 사용자들의 선호평가 점수로 구성된 R_p를 이용하여 목표 사용자의 개인화된 추천 리스트를 만들어 낼 수 있다. 다양한 기법들이 사용될 수 있지만 아이템 기반의 협력 필터링인 경우 단순 가중치 평균을 사용하여 사용자 u가 시청하지 않은 아이템 i에 대한 예상 점수 $P(u,i)$를 도출할 수 있다[4]. 식 (6)에서의 N은 사용자 u가 시청한 적이 있는 프로그램들이다.

$$P(u,i) = \frac{\sum_{n \in N} R_p(u,n) S(i,n)}{\sum_{n \in N} |S(i,n)|}, \quad \text{if } R_s(u,i) = 0 \qquad (6)$$

그런데 다른 추천 도메인과 달리 TV 프로그램은 시리즈 특성 때문에 사용자가 소비한 적이 없는 신규 아이템에 대해서만 예상 점수를 도출해서 추천하는 것은 적당하지 않다. 예전에 아주 만족도를 가지고 시청한 프로그램에 새로운 회별 프로그램이 제공되면 사용자가 만족스럽게 시청할 것은 예상 가능한 일로 과거 사용자가 시청한 적이 있는 프로그램의 신규 회별 프로그램 또한 추천의 대상이 되어야 하기 때문이다. 추천은 새로운 것에 대한 발견도 있지만 개인화된 콘텐츠 배치를 제공한다는 점에서 사용자가 콘텐츠를 선택할 때 손쉬운 지름길로 활용되는 것이 의미가 있을 수 있다. 게다가 KBS VOD 서비스 플랫폼들에서는 [그림 4]의 콘팅에서와 같이 프로그램 단위가 아닌 회별 프로그램 단위로 콘텐츠를 배열해서 서비스하기도 한다.

그림 4. 콘팅 플랫폼에서 회별 프로그램 단위의 콘텐츠 배치
Fig. 4. Contents arrangement by episode unit in Conting

이를 위해서는 식 (6)에 이미 사용자가 시청한 프로그램에 대한 선호 평가가 추가되어야 하는데 이것은 이미 계산해 둔 R_s를 그대로 활용할 수 있기 때문에 식 (6)에 식 (7)을 추가하면 된다. 이렇게 구성할 수 있는 이유는 식(6)이 해당 사용자의 평점들의 가중치 합으로 구성되어 식(6)과 식(7)의 스케일이 동일하기 때문이다.

$$P(u,i) = R_s(u,i), \quad \text{if } R_s(u,i) \neq 0 \qquad (7)$$

식 (6),(7)의 결과로 모든 프로그램에 대한 사용자의 선호 점수가 도출되면 선호 점수를 내림차순으로 정렬하여 점수가 높은 프로그램으로 추천 리스트를 구성할 수 있다. 추천의 단위가 회별 프로그램인 경우에는 프로그램을 회별 프로그램으로 매핑시켜주는 과정이 필요하게 된다. 추천 매핑 과정은 TV 시리즈의 순차적 특성 상 사용자가 본인이 가장 마지막으로 본 것을 찾아보길 원한다는 점과 한 시리즈 내에서 너무 많은 추천이 일어나면 추천의 의미가 없어질 수 있다는 점을 고려해서 구성해야 한다. 추천 프로그램을 회별 프로그램 단위로 매핑하는 과정은 [그림 5]와 같다.

[그림 5]에서 우측 상단은 프로그램 예상 점수로 정렬된 프로그램 리스트이다. 시청한 적 있는 프로그램과 한 번도 시청한 적이 없는 프로그램이 섞여 있다. 한계값 C는 더 다양한 프로그램을 추천하는 데 중점을 줄 것인지, 다양하지는 않더라도 예상 점수가 높을 것 같은 프로그램의 다른 회차들을 더 추천해 줄 지를 결정하는 설정값이다.

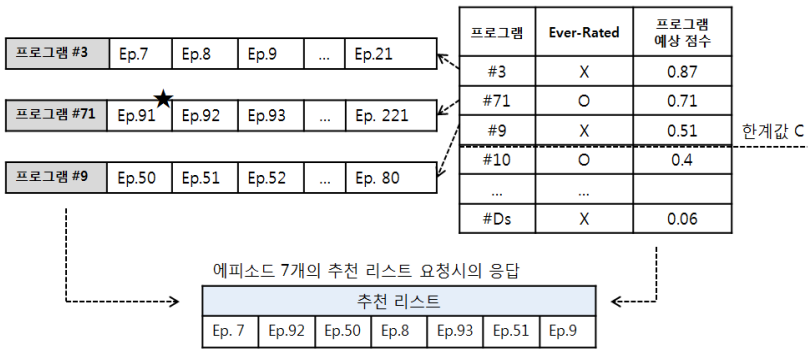

그림 5. 라운드 로빈 기반의 프로그램- 회별 프로그램로의 매핑
Fig. 5. Mapping from program to program episode based round-robin rule

C에 의해서 프로그램 목록이 결정되면 라운드 로빈 방식으로 해당 프로그램을 각 회별 프로그램으로 매핑한다. 각 프로그램에 대해서는 포함되어 있는 회별 프로그램들에 대한 정보가 있는데 사용자가 한 번이라도 본 적이 있는 프로그램은 마지막으로 본 회차의 다음 회차를, 한 번도 본 적이 없는 프로그램 시리즈는 제일 처음 회차를 리스트로 제공해 주며 프로그램 목록의 끝에 도착하면 처음 프로그램으로 다시 되돌아가서 그 다음 회차를 선택하게 된다. [그림 5]의 ★표시는 해당 사용자가 그 프로그램에 대하여 마지막으로 본 회별 프로그램을 표시하고 있으며 실제 서비스에서 이런 식으로 추천 시스템이 빠르게 동작하기 위해서는 모든 사용자에 대하여 마지막으로 본 회별 프로그램 목록을 색인해 두어야 한다.

추천 리스트 생성의 계산 복잡도(Computational Complexity)는 회별 프로그램 단위로 처리할 때의 $O(D_e^2)$에서 $O(D_s(D_e+D_s))$로 줄어든다.

Ⅳ. 모의실험

1. 실험 환경

Ⅲ장에서 제안된 내용의 효과를 알아보기 위하여 실험을 수행하였다. 실험에서 확인하고자 하는 첫 번째 는 예측 정확도에 대한 것으로 제안된 알고리즘을 구성하는 요소들을 변경하며 실험하여 프로그램 단위별 처리, 데이터 전처리 시의 가중치 추가가 예측 정확도에 미치는 영향을 알아본다. 두 번째는 제안 알고리즘의 처리 속도 개선 효과이다.

추천 점수 예상을 위한 트레이닝 셋(Training Set)과 추천 시스템의 성능을 평가하기 위한 테스트 셋(Test Set)은 KBS 콘팅 사이트의 실제 VOD 시청 로그 데이터를 사용하였다. KBS 콘팅 사이트에서 2010년 3월에서 4월까지 수집된 이용 로그를 트레이닝 셋으로, 5월 한 달 간 수집된 이용 로그를 테스트 셋으로 사용하였다. 데이터 상세 내역은 [표 2]와 같다. Ⅲ.1에서 설명한 데이터 변환에 대한 예시는 [표 3]과 같다. 사용자 로그 수와 회별 프로그램 보유 수의 상관

표 2. 데이터셋 상세내용
Table 2. Dataset specification

	아이템 단위	사용자 수	아이템 수	사용자 평가 건수	데이터 기간
트레이닝 셋	회별 프로그램 (에피소드)	61545	1067	312298	2010년 3~4월 (2달)
	프로그램 (시리즈)	61545	67	89408	
테스트 셋	회별 프로그램 (에피소드)	10000	1379	33060	2010년 5월 (1달)

표 3. 특정 사용자 '가'를 위한 데이터 변환 예시
Table 3. Data conversion example for a user '가'

사용자 ID	프로그램별 시청기록	프로그램 시리즈 정보	'가'의 해당 프로그램에 대한 선호점수 도출
'가' (총 9개의 회별 프로그램 시청)	A-1회, A-2회, A-3회, A-4회	A : (1회~5회)	$\frac{4}{5}*(1-e^{\frac{\log 2}{2}*5})*1$ = 0.66
	B-1회, B-2회, B-3회, B-4회	B : (1회~40회)	$\frac{4}{40}*(1-e^{\frac{\log 2}{2}*40})*\frac{40}{9}$ = 0.44
	C-2회	C : (1회~2회)	$\frac{1}{2}*(1-e^{\frac{\log 2}{2}*2})*1$ = 0.25

관계를 고려하여 개인화된 프로그램별 선호점수가 계산되며 선호를 판별할 수 없는 짧은 시리즈의 불확실성도 반영되었다.

사용자 선호 평가 점수 자체가 실제 데이터가 아니기 때문에 결과 평가를 위하여 MAE(Mean Absolute Error)와 같은 예측 정확도를 사용할 수는 없으므로 사용자의 시청 여부에 대한 분류 정확도 기준(Classification Accuracy Metrics) 중 F1 점수를 사용한다. F1 점수는 정확도(Precision)와 재현도(Recall)의 조화평균(Harmonic Mean)이며 정확도는 추천 시스템이 추천한 리스트의 수 중 사용자가 실제 시청한 것으로 나타난 것의 비율이이며, 재현도는 사용자의 소비 리스트 중 추천이 된 것의 비율이다.

2. 실험 결과

[그림 6]은 테스트 셋에 존재하는 10000명의 사람들에게 추천을 제공했을 때의 결과 평가 지표에 대한 것이다. F1 점수는 추천 개수에 따라 결과가 달라지므로 추천 개수에 따라 그래프를 도시하였다.

회별 프로그램에 대한 사용자 소비 이력을 프로그램 사용자 선호평가 점수로 변환하여 프로그램 유사도를 계산하고 프로그램 유사도를 기반으로 다시 회별 프로그램들을 추천하는 방법을 RoSE (Recommendation on Series-Episode)라고 표기하였다. 각 실험에서 사용한 유사도 지표(PCC, COS)을 함께 표시하였다. 회별 프로그램을 협력 필터링 처리 단위로 하는 아이템 기반 IBCF(Item-Based Collaborative Filtering)를 RoSE에 대한 베이스라인으로 비교하였다. 또한 RoSE를 사용하되, 프로그램 단위의 평가 점수에 식 (2), 식(3)의 가중치를 적용했을 때의 결과를 도시하였다.

실험 결과는 RoSE 기법을 적용하여 추천을 처리한 경우들이 회별 프로그램들을 독립적인 아이템으로 처리한 베이스라인보다 높은 추천 성능을 보여줌을 알 수 있다. 동일한 COS 유사도 지표 선택시 베이스라인보다 RoSE의 정확도가 약 3.65배 높다. 동일하게 RoSE를 적용한 경우 유사도 지표를 COS에서 PCC로 변경하자 정확도가 더 높아지는 것은 적당한 유사도를 채택함에 따라 정확도는 더 높아질 수 있음을 보여준다. 또한 RoSE와 PCC가 적용된 경우에 프로그램 단위의 평가 점수의 특성을 고려한 세밀한 가중치 적용을 더하자 추천 정확도가 약 10.6% 향상되는 것을 확인할 수 있다. 본 실험을 통하여 각각의 회별 프로그램 시청 기록만으로 알 수 없었던 사용자의 미래 선호 예측이

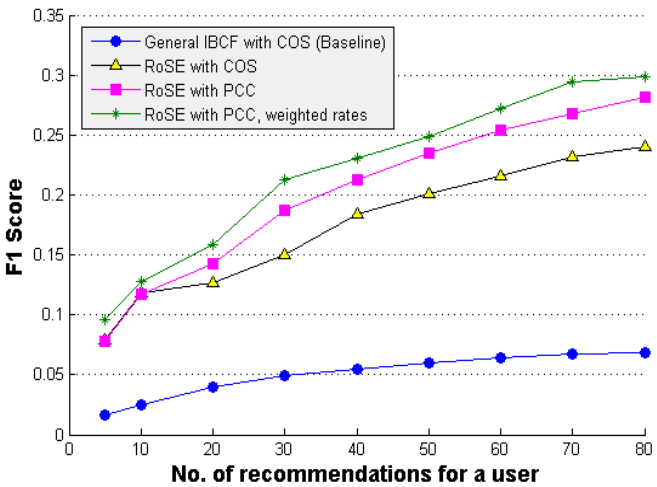

그림 6. 추천 성능 평가 지표 결과
Fig. 6. Evaluation result for recommendation

협력 필터링 처리 단위를 프로그램 단위로 변환하여 계산한 뒤 다시 회별 프로그램으로 변환하여 추천함으로써 정확해진 것을 알 수 있다. 또한 세밀한 가중치 부여를 통하여 추가적인 정확도 개선을 얻을 수 있음을 알 수 있었다.

[그림 7]은 동일 유사도 지표 PCC를 사용하여 테스트 셋의 10000명의 사람들에게 추천 결과를 생성해 내는 데 걸리는 시간을 보여준다. RoSE의 처리 시간이 회별 프로그램들을 독립적인 아이템으로 처리(베이스라인)했을 때의 약 4.4%만 소요되었다.

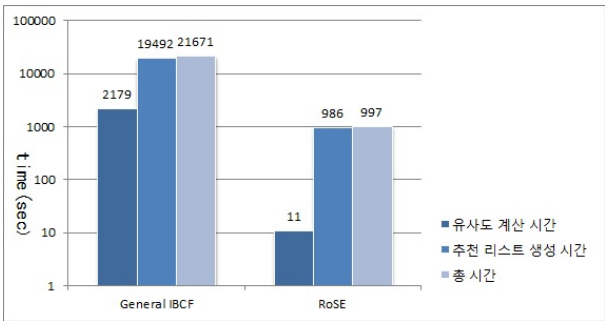

그림 7. 추천 결과 생성 소요 시간
Fig. 7. Run-time for generating recommendation

V. 결 론

본 논문에서는 TV 프로그램 환경에서 실용적으로 협업 필터링을 사용하여 개인화된 추천을 적용할 수 있는 방법을 연구하였다. 실제 TV 프로그램이 소비되고 추천되는 단위 아이템은 회차별 프로그램이지만 회차별 프로그램 단위로 협력 필터링을 적용하면 사용자 선호에 대한 불확실함으로 추천 정확도가 다소 낮으며 계산량도 많다. 따라서 협력 필터링 처리 단위를 프로그램 단위로 변경하게 되면 지속적인 프로그램 시청 누적 회수를 해당 프로그램에 대한 선호점수로 판단할 수 있게 되어 차원 축소 효과와 함께 선호를 판별할 수 있는 의미 있는 축(latent)을 얻을 수 있게 된다. 반면 프로그램 단위로 협력 필터링을 적용하게 되면 사용자가 한 번 시청한 프로그램은 더 이상 신규 프로그램으로 처리되지 않으며 한 프로그램 내에서 여러 회차를 순차적으로 시청하는 사용자의 소비 패턴과도 맞지 않게 되므로 추천 프로그램 매핑 방법으로 이를 해결하였다.

TV 프로그램 서비스 플랫폼에서 수집된 데이터를 사용한 실험 결과는 본 논문에서 제시한 방식이 정확도와 계산 복잡도 면에 있어서 더 향상된 결과를 가져옴을 알 수 있다. 정확도 향상은 아이템 단위 기준 변경, 평가 데이터 변환 기법 등에 의한 것이며, 계산 복잡도 감소는 유사도 계산 대상의 차원 축소에 의한 것으로 처리 속도를 단축시켜 준다. 제안한 방식은 추천 분야에서 광범위하게 쓰이고 연구되는 협력 필터링의 여러 가지 기법의 조합에 유연하게 적용될 수 있는데, 그 이유는 해당 방법이 사용자 소비 이력 데이터 전처리와 추천 결과 후처리에 초점을 두고 있기 때문이다. 본 논문에서는 제시한 방법의 효과만을 알아보기 위하여 가장 일반적인 방식의 협업 필터링 기법을 적용하였지만 실제 시스템 구현시 다양한 협업 필터링 기법의 조합 실험과 설정값 튜닝이 수반될 것이다. 해당 기법은 시리즈 특성을 가진 아이템을 추천하는 다른 도메인들에도 사용될 수 있으나 시리즈가 각 아이템에 대한 대표성이 약하거나 시리즈나 단편의 비율이 비슷할 때는 본 제안 방법의 효과가 줄어들 수 있다. 따라서 그러한 경우에는 콘텐츠 메타데이터를 이용하는 하이브리드 방식을 도입하는 것이 유용할 수 있다.

참 고 문 헌

[1] Soo-Young Oh, Yeonhee Oh, Sunghee Han, Hee Jung Kim, “Broadcast Content Recommender System based on User’s Viewing History”, JBE, Vol. 17, No. 1, pp130~140, Jan, 2012
[2] Xiaoyuan Su and Taghi M. Khoshgoftaar. "A Survey of Collaborative Filtering Techniques," Advances in Artificial Intelligence Vol. 2009, Article No. 4, 2009.
[3] István Pilászy and Domonkos Tikk. 2009. "Recommending new movies: Even a few ratings are more valuable than metadata," in Proc. Recsys 2009, ACM, New York, 2009.
[4] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. "Item based collaborative filtering recommendation algorithms," in Proc. 10th International Conference on WWW, ACM, NewYork, 2001.
[5] Eunhui Kim, Shinjee Pyo, Eunkyung Park, and Munchrul Kim. "An automatic recommendation scheme of TV program contents for IPTV

personalization," IEEE Transactions on Broadcasting, Vol. 57, No. 3, 2011.

[6] Yifan Hu, Yehuda Koren and Chris Volinsky. "Collaborative filtering for implicit feedback datasets," in Proc. 8th IEEE International Conference on Data Mining, pp. 263-272, 2008.

[7] Manzhao Bu, Shijian Luo, and Ji he. "A fast collaborative filtering algorithm for implicit binary data," IEEE 10th International Conference on Computer-Aided Industrial Design & Conceptual Design, pp. 973 - 976, 2009.

[8] Hyung Jun Ahn. "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," Information Sciences 178, ScienceDirect, pp. 37-51, 2008.

저 자 소 개

한 성 희

- 2001년 2월 : 고려대학교 전기전자전파공학부 학사
- 2003년 2월 : 고려대학교 전기공학과 석사
- 2003년 1월 ~ 2006년 5월 : 삼성전자 무선사업부 선임연구원
- 2007년 3월 ~ 현재 : KBS 기술연구소 연구원
- 주관심분야 : 콘텐츠 추천 시스템, 방송 자막 활용, 하이브리드 방송 플랫폼

오 연 희

- 2000년 2월 : 서울대학교 컴퓨터공학과 학사
- 2002년 9월 : University College London, MSc in DCNDS 석사
- 2008년 11월 ~ 2009년 10월 : NHK 기술연구소 객원연구원
- 2003년 1월 ~ 현재 : KBS 기술연구소 선임연구원
- 주관심분야 : 콘텐츠 추천/검색, 메타데이터, 정보 추출, 멀티미디어 콘텐츠 서비스

김 희 정

- 1985년 2월 : 이화여자대학교 전자계산학과 학사
- 1988년 2월 : KAIST 전산학과 석사
- 1988년 ~ 현재 : KBS 기술연구소 방송기술연구부장
- 주관심분야 : 콘텐츠 추천, 동영상 편집전송, 컴퓨터 그래픽스