IPTV 시청자의 시청이력에 기반한 협력필터링 모델 분석

정하용 ○ 김문식

KT

hyj@kt.com, moonsix@kt.com

Collaborative Filtering Model Analysis based on IPTV Viewing Log

Ha-Yong Jung ○ MoonSik Kim

KT

요 약

협력 필터링(Collaborative Filtering)은 상품추천, 영화추천 등에 사용되는 대표적인 방법으로서, 사용자들의 사용이력에 기반해서 유사도가 높은 항목들을 찾아낸다. 본 연구에서는 상용 IPTV 서비스에 협력 필터링을 적용했을 때 만들어지는 모델을 분석하여 어떤 요소들이 협력 필터링 모델의 생성에 영향을 끼치는지 분석했다. 이를 통해 IPTV 영역에 협력 필터링을 적용했을 때 영향을 끼치는 요소들과 다른 영역과는 다르게 고려해야 할 사항들을 알 수 있었다.

1. 서 론

IPTV는 시청자들이 원하는 시점에 원하는 VOD를 스스로 선택해서 시청한다는 것이 UX(User Experience)적인 관점에서 기존의 TV와 가장 큰 차이점이다. 일반적으로 사용자 스스로 자신이 원하는 것을 선택하는 것은 사용자의 만족을 높일 수 있는 요소이지만 그것은 선택의 폭이 제한적일 경우의 이야기이다. 선택의 폭이 수 개에서 수십 개 내외가 아닌 수만 개에서 수십만 개가 된다면 사용자들은 만족이 아니라 선택의 고통에 빠지게 된다. 이것이 바로 기존의 TV 시청방식에 익숙해져 있는 일반 시청자들이 IPTV를 어렵게 느껴서, 결국은 사용하지 않게 만드는 가장 핵심적인 요소이다.

이런 관점에서 IPTV 시청자들에게 VOD 선택의 고통을 줄여주는 것은 매우 중요하다. 이 때 도움을 줄 수 있는 것이 바로 VOD 추천 서비스이다. 하지만 기존의 인기작 위주의 VOD 추천은 모든 사용자에게 똑같은 결과를 주기 때문에 쉽게 질리고 만족도가 떨어진다. 결국 사용자 맞춤형 VOD 추천 서비스가 필요한데, 이를 위해 사용될 수 있는 기술이 사용자 로그 기반의 협력 필터링(Collaborative Filtering, CF)이다.

CF는 상품추천, 영화추천 등 다양한 분야에서 사용되어왔지만 IPTV 영역에 CF를 적용하는 것은 기존의 CF를 적용하던 영역과는 다른 특징이 있다. 본 연구에서는 이를 확인하고자 상용 IPTV 서비스인 KT의 쿡TV 서비스 사용자들의 시청로그에 CF를 적용해서 CF 모델을 생성했으며, 생성된 CF모델을 분석해서 IPTV 영역에 협력 필터링을 적용할 때 고려해야 할 사항들을 파악하고자 한다.

2. 관련 연구

추천 시스템은 크게 추천의 기반정보를 어디서 얻느냐에 따라서 콘텐트 기반 방법(Content-based Approach)과 협력 필터링 방법(Collaborative Filtering Approach)로 나뉘어진다. 여기서 콘텐트 기반 방법은 추천의 기반정보를 콘텐트 아이템으로부터 얻는 방법이고, 협력 필터링 방법은 추천의 기반정보를 사용자의 사회적 환경으로부터 얻는 방법이다[1]. 이 중 실제 서비스에 주로 사용되는 방법은 협력 필터링 방법인데, 그것은 콘텐트 기반 방법은 콘텐트 내용 분석의 어려움 등으로 기반 데이터를 얻기 어려울뿐더러 데이터를 얻은 경우에도 그것을 통해서 사용자가 만족하는 결과를 얻기 어렵기 때문이다. 반면 협력 필터링 방법의 경우, 일반적으로 사용자의 행동 이력을 계속 기록해서 분석하면 되므로, 기반 데이터를 얻기가 쉬운 편이고, 그 메커니즘 자체도 실제 사회에서 사람이 추천하는 방식과 유사하기 때문에 사용자 만족도도 높은 편이다.

협력 필터링 방법의 이와 같은 장점 때문에 실제 상용 서비스에서 그것을 응용하는 예도 많이 있다. 가장 대표적인 서비스는 Amazon.com의 도서 추천 서비스이다[2]. 이 서비스는 상용 서비스에서 좋은 평가를 받고 있으며, 실제 추천해 주는 도서의 연관성이 높아서 사용자 만족도가 높은 것으로 알려져 있다[3]. 이외에도 Digg.com[4], IMDB(Internet Movie Database)[5], Last.fm[6], Netflix[7], Tivo[8] 등 많은 상용 서비스들에서 협력 필터링 기반의 추천 서비스를 제공하고 있다.

3. 접근 방법

3.1. 데이터

분석에 사용된 시청이력 데이터는 KT의 상용 IPTV 서비스인 쿡TV 서비스 사용자들이 2009년 12월 1일부터 2009년 12월 26일까지 약 1개월 동안 시청한 시청이력 데이터로서 건수로는 약 9천만 건에 해당한다. 이외에도 협력필터링 모델 생성에 영향을 끼치는 요소를 찾기 위해 쿡TV 서비스의 VOD 컨텐츠 메타데이터 정보가 사용됐다.

3.2. 알고리즘

사용자의 시청이력 정보를 이용한 VOD 추천 서비스 모델을 만들기 위해 사용한 알고리즘은 협력 필터링(Collaborative Filtering)이다. 전형적으로 많이 사용되는 User-Item 방식의 CF가 아닌 Item-Item CF 방식을 사용했다. 주어진 사용자와 구매패턴이 유사한 사용자들을 찾아서 그 사용자들은 구매했지만 주어진 사용자는 구매하지 않은 아이템을 추천해 주는 User-Item CF와 다르게 Item-Item CF는 주어진 아이템을 구매한 사용자들이 비율적으로 가장 유사하게 구매한 다른 아이템을 찾는 방식이다. 최근의 연구들에서 Item-Item CF의 성능이 User-Item CF보다 일반적으로 더 좋다는 것이 확인되고 있다. 아이템들간의 유사도 측정방법은 코사인 유사도(Cosine Similarity)를 사용했다.

3.3. 분석 방법

데이터 분석은 3.1의 데이터를 3.2의 방법으로 모델링해서 만들어진 CF 모델에 어떤 규칙들이 포함되어 있는지를 살펴보는 방식으로 진행했다. CF 모델 생성에 영향을 끼치는 요소들을 분석하기 위해 VOD 컨텐츠의 특성 별로 모델에 포함된 규칙들에 어떤 것들이 있는지를 살펴봤으며, 분석 대상이 된 VOD 컨텐츠의 특성은 시리즈 물, 시청등급, 카테고리, 가격, 시청빈도, 주연배우, 감독 등이다.

4. 협력 필터링 (Collaborative Filtering) 모델 분석

4.1. 시리즈 물

기존의 CF는 대부분 상품 추천이나 영화 추천의 영역에서 이루어졌다. 하지만 IPTV의 VOD 추천 서비스와 영화 추천 서비스는 여러 차이점이 존재하는데, 가장 큰 차이점은 IPTV VOD 서비스는 영화뿐만이 아니라, TV 방송을 서비스한다는 점이다.

이것은 협력 필터링에도 큰 영향을 끼치게 되는데, 가장 크게 영향을 끼치는 요소는 TV 방송들은 대부분 시리즈로 구성된다는 점이다. 영화는 하나하나의 영화가 대부분 다른 영화와 큰 연관관계 없이 독립적인 것과 다르게 TV 방송은 여러 개의 VOD가 하나의 프로그램을 구성하고 이것을 수개월에서 수년의 기간 동안 나눠서 방송한다. 따라서 이와 같은 시리즈 물 들은 기존에 시청했던 시청자들이 다시 시청할 가능성이 크고 동일 시리즈 물 내 VOD들 간의 연관관계가 강하게 될 가능성이 크다.

실제 CF 모델에도 이와 같은 시리즈 물의 특징적인 연관관계가 드러나는지 확인하기 위해 특정 시리즈 물인 VOD 중 임의로 몇 개의 VOD를 선택해서 그 VOD를 시청했을 경우 CF 모델을 통해서 받게 되는 추천결과 리스트를 살펴본 결과는 다음 표1과 같았다.

표 1 CF 모델에서 유사도가 높은 결과의 예 (시리즈 물)

시청한 VOD 제목	MBC_125회_거침없이 하이킥	4화 뽀롱뽀롱 뽀로로(13~16)
결과리스트 항목	거침없이 하이킥 126회	3화 뽀롱뽀롱 뽀로로(9~12)
	거침없이 하이킥 124회	6화 뽀롱뽀롱 뽀로로(21~24)
	거침없이 하이킥 110회	7화 뽀롱뽀롱 뽀로로(25~28)
	거침없이 하이킥 106회	2화 뽀롱뽀롱 뽀로로(5~8)
	거침없이 하이킥 115회	8화 뽀롱뽀롱 뽀로로(29~32)
	거침없이 하이킥 119회	9화 뽀롱뽀롱 뽀로로(33~36)
	거침없이 하이킥 117회	10화 뽀롱뽀롱 뽀로로 (37~40)
	거침없이 하이킥 122회	1화 뽀롱뽀롱 뽀로로(1~4)
	거침없이 하이킥 127회	11화 뽀롱뽀롱 뽀로로 (41~44)
	거침없이 하이킥 133회	12화 뽀롱뽀롱 뽀로로 (45~48)
시리즈 비율 (10위)	10 / 10 = 100%	10 / 10 = 100%
시리즈 비율 (전체)	99 / 99 = 100%	53 / 85 = 62%

표1에서 확인할 수 있는 바와 같이 시리즈 물 내 VOD들 간의 연관관계가 매우 강한 것을 확인할 수 있었다. CF 모델 상에서 "MBC_125회_거침없이 하이킥"과 유사도가 높은 VOD들은 상위 10위 전부가 동일 시리즈 물 내의 VOD였고, 상위 99위까지도 전부 동일 시리즈 물 내의 VOD였다. 마찬가지로 "4화 뽀롱뽀롱 뽀로로(13~16)"과 유사도가 높은 VOD들은 상위 10위 전부가 동일 시리즈 물 내의 VOD였고, 상위 85위 중 62%인 53개 VOD가 동일 시리즈 물 내의 VOD였다.

시리즈 물은 특히 드라마, 어린이, 애니메이션, 교육 등의 카테고리에서 많이 나타났으며, 시리즈 물 내 VOD의 개수가 충분히 많을 경우 대부분의 시리즈 물에서 유사도 상위 100위 중 90% 이상이 동일 시리즈 물이었다.

이와 같은 현상이 일반적인 현상인지 확인하기 위해서 CF 모델 전체에서 시청한 VOD가 특정

시리즈였을 때 그와 유사도가 높은 VOD 역시 동일 시리즈인 비율을 살펴본 결과는 다음 표2와 같았다.

표 2 CF 모델 전체에서 시리즈 물의 비율

	조건 VOD가 시리즈 물	결과 VOD가 시리즈 물	결과 VOD가 동일 시리즈 물
규칙 개수	2,828,476	2,769,417	1,500,335
비율	100%	98%	53%

표2에서 확인할 수 있는 바와 같이 시리즈 물 내 VOD들 간의 연관관계가 매우 강한 것은 특정 예에서만 드러나는 특수한 현상이 아니라 CF 모델의 규칙 전반에서 나타나는 일반적인 현상임을 확인할 수 있었다.

4.2. 시청 등급

시청 등급은 시청자의 연령에 따라 시청자들의 VOD 시청을 직접적으로 제한하는 요소이기 때문에 시청자들의 VOD 시청패턴에 큰 영향을 줄 수 있는 요소이다. 특히 시청등급이 크게 나누어지는 연령은 19세인데, 쿡TV 서비스의 경우 시청등급이 19세 이상인 VOD의 경우 비밀번호를 입력해야 시청이 가능하다. 따라서 시청등급이 19세 이상인 VOD는 성인들만 시청할 가능성이 높으며, 시청등급이 19세 이상인 VOD들 간의 연관관계가 강하게 될 가능성이 크다.

실제 CF 모델에도 이와 같은 시청 등급의 특징적인 연관관계가 드러나는지 확인하기 위해 시리즈 물과 마찬가지로 시청 등급이 19세 이상인 VOD 중 임의로 몇 개의 VOD를 선택해서 그 VOD를 시청했을 경우 CF 모델을 통해서 받게 되는 추천결과 리스트를 살펴본 결과는 다음 표3과 같았다.

표 3 CF 모델에서 유사도가 높은 결과의 예 (시청등급 19세)

시청한 VOD 제목	바람피기 좋은날	색즉시공
결과리스트 항목	연애의 목적	바람피기 좋은 날
	연애	연애, 그 참을수 없는 가벼움
	내 여자의 남자친구	연애
	누구나 비밀은 있다	내 여자의 남자친구
	색다른 동거	연애의 목적
	색즉시공	청춘
	밀애(상)	스물넷
	스물넷	누구나 비밀은 있다
	음란서생	밀애(상)
	청춘	색, 계
19금 비율 (10위)	10 / 10 = 100%	10 / 10 = 100%
19금 비율 (전체)	30 / 38 = 79%	27 / 31 = 87%

표3에서 확인할 수 있는 바와 같이 시청 등급이 19세 이상인 VOD들 간의 연관관계가 매우 강한 것을 확인할 수 있었다.

하지만 시청등급이 19세 이상인 VOD는 시청자의 연관 시청패턴과 CF모델이 서로 매우 다른 모습을 보였다. 즉, 시청자들의 연관시청패턴을 살펴봤을 때, 시청등급이 19세 이상인 VOD를 시청하더라도 그 VOD와 함께 가장 많이 시청한 VOD는 일반적으로 인기가 있는 시리즈 물이나 유아 프로그램 등의 VOD인 경우가 대부분이었다. 이것은 IPTV의 시청 로그가 IPTV 셋탑박스 단위로 기록되기 때문에 생기는 현상으로 가족이 하나의 셋탑박스를 통해서 IPTV를 시청하기 때문에 벌어지는 현상이다. 하지만 CF모델은 단순히 특정 VOD와 함께 가장 많이 시청한 VOD가 아닌 시청패턴이 비율적으로 특정VOD와 가장 유사한 VOD를 찾기 때문에 시청등급이 19세 이상인 VOD간에 강한 연관관계를 찾을 수 있었다.

시리즈 물과 마찬가지로 이와 같은 현상이 일반적인 현상인지 확인하기 위해서 CF 모델 전체에서 시청한 VOD의 시청등급이 19세 이상이었을 때 그와 유사도가 높은 VOD의 시청등급 역시 19세 이상인 비율을 살펴본 결과는 다음 표4와 같았다.

표 4 CF 모델 전체에서 시청등급 19세 이상의 비율

	조건 VOD가 19세 시청가	결과 VOD가 19세 시청가
규칙 개수	116,437	66,680
비율	100%	57%

4.3. 카테고리

카테고리는 특정 VOD 내용의 유형을 대표적으로 나타내는 요소이기 때문에 시청자들의 선호 카테고리나 장르를 반영해서 시청자들의 VOD 시청패턴에 큰 영향을 줄 수 있는 요소이다. 카테고리는 VOD의 내용을 반영하는 요소이기 때문에 그 VOD의 시청자들의 내용적인 선호 성향을 나타낼 가능성이 크다. 따라서 카테고리가 동일한 VOD들 간의 연관관계가 강하게 될 가능성 역시 크다. 특히 IPTV 서비스는 영화뿐만이 아니라 드라마, 뉴스, 연예오락, 다큐멘터리, 시사, 애니메이션, 뮤직비디오 등 다양한 방송 VOD를 포함하기 때문에 카테고리가 시청패턴에 영향을 끼칠 가능성은 더욱 크다.

실제 CF 모델에도 이와 같은 카테고리의 특징적인 연관관계가 드러나는지 확인하기 위해 앞선 분석들과 마찬가지로 카테고리가 영화인 VOD들 중 임의로 몇 개의 VOD를 선택해서 그 VOD를 시청했을 경우 CF 모델을 통해서 받게 되는 추천결과 리스트를 살펴본 결과는 다음 표5와 같았다.

표 5 CF 모델에서 유사도가 높은 결과의 예 (카테고리: 영화)

시청한 VOD 제목	추격자	웰컴 투 동막골
결과리스트 항목	예스터데이	원스 어폰 어 타임
	연애	언니가 간다
	색다른 동거	황산벌
	연애, 그 참을수 없는 가벼움	가문의 부활
	해바라기	사생결단
	원스 어폰 어 타임	해바라기
	바람피기 좋은 날	80일간의 세계일주
	색즉시공	도레미파솔라시도
	수	내생애 가장 아름다운 일주일
	내 여자의 남자친구	광복절 특사
영화 비율 (10위)	10 / 10 = 100%	10 / 10 = 100%
영화 비율 (전체)	28 / 28 = 100%	41 / 41 = 100%

표5에서 확인할 수 있는 바와 같이 카테고리가 영화인 VOD들 간의 연관관계가 매우 강한 것을 확인할 수 있었다.

앞선 분석들과 마찬가지로 이와 같은 현상이 일반적인 현상인지 확인하고자 했으나 확보한 VOD 메타데이터에서 카테고리 정보가 완전하지 않은 관계로 앞선 분석들처럼 CF 모델 전체에서 카테고리가 동일한 규칙들의 비율을 확인할 수는 없었다. 하지만 표5에서 확인한 예 뿐만 아니라 다양한 카테고리의 많은 예들을 통해서 살펴본 결과 이와 같이 카테고리가 동일한 VOD들 간의 연관관계가 매우 강한 것은 대부분의 카테고리에서 동일하게 나타나는 일반적인 현상이었다.

4.4. 가격

쿡TV 서비스는 대부분의 VOD를 무료로 시청할 수 있지만 일부 최신 VOD들이나 성인 VOD들의 경우 별도의 가격을 지불해야 시청할 수 있다. 가격은 일반적으로 모든 상품의 구매에 있어서 가장 큰 영향을 끼치는 요소이며, 이것은 IPTV 서비스에서도 마찬가지이다. 특히 쿡TV 서비스의 경우 대부분의 VOD는 무료로 시청이 가능하기 때문에 가격에 의해 시청패턴이 가장 뚜렷하게 나누어지는 기준은 VOD 가격이 얼마인가가 아니라 VOD가 유료 여부이다. 따라서 유료 VOD를 시청한 시청자는 또 다른 유료 VOD를 시청할 가능성이 높으며, 유료 VOD들 간의 연관관계가 강하게 될 가능성이 있다.

실제 CF 모델에도 이와 같은 가격의 특징적인 연관관계가 드러나는지 확인하기 위해 앞선 분석들과 마찬가지로 유료 VOD 중 임의로 몇 개의 VOD를 선택해서 그 VOD를 시청했을 경우 CF 모델을 통해서 받게 되는 추천결과 리스트를 살펴본 결과는 다음 표6과 같았다.

표 6 CF 모델에서 유사도가 높은 결과의 예 (가격: 유료)

시청한 VOD 제목	추격자 (1800원)	겟 썸 (1400원)
결과리스트 항목	예스터데이 (무료)	람보 4 : 라스트 블러드 (무료)
	연애 (무료)	더블 팀 (무료)
	색다른 동거 (무료)	13구역 (무료)
	연애, 그 참을수 없는 가벼움 (무료)	블랙 호크 다운 (무료)
	해바라기 (무료)	80일간의 세계일주 (무료)
	원스 어폰 어 타임 (무료)	러브 인 맨하탄 (무료)
	바람피기 좋은 날 (무료)	욕망 (무료)
	색즉시공 (무료)	3:10 투 유마(2007) (무료)
	수 (무료)	레지던트 이블 3 (무료)
	내 여자의 남자친구 (무료)	사생결단 (무료)
유료 비율 (10위)	0 / 10 = 0%	0 / 10 = 0%
유료 비율 (전체)	1 / 28 = 4%	2 / 31 = 6%

표6에서 확인할 수 있는 바와 같이 예상과는 다르게 유료 VOD들 간의 연관관계는 거의 없는 것을 확인할 수 있었다. 반대로 무료 VOD들 간의 연관관계는 무척 강했는데, 이것은 특별히 무료 VOD들 간의 연관관계가 강하다기 보다는 유료 VOD든 무료 VOD든 상관없이 무료 VOD와의 연관관계가 강한 것으로 파악된다. 이와 같은 결과는 IPTV 시청자들이 추가적인 부담 없이 볼 수 있는 무료 VOD를 선호하기 때문에 무료 VOD의 시청빈도가 클 뿐만 아니라, VOD의 절대적인 개수 자체도 무료 VOD가 유료 VOD보다 훨씬 많기 때문에 나타나는 현상으로 보인다.

앞선 분석들과 마찬가지로 이와 같은 현상이 일반적인 현상인지 확인하기 위해서 CF 모델 전체에서 시청한 VOD가 유료였을 때 그와 유사도가 높은 VOD 역시 유료인 비율을 살펴본 결과는 다음 표7과 같았다.

표 7 CF 모델 전체에서 유료 VOD의 비율

	조건 VOD가 유료	결과 VOD가 유료
규칙 개수	139,265	76,210
비율	100%	55%

CF 모델 전체에서 유료 VOD들 간의 연관관계를 살펴본 결과 표6의 예에서 확인할 수 있었던 것과 다르게 CF 모델 전체에서는 유료 VOD들 간의 연관관계가 매우 강한 것을 확인할 수 있었다. 하지만 데이터를 좀 더 자세히 분석한 결과 이것은 유료 VOD들 간의 연관관계가 강한 것이 아님을 알 수 있었다. 통계적으로는 유료 VOD들 간의 연관관계가 강한 것처럼 보였지만 유료 VOD들의 대부분은 시리즈

물과 성인물이 차지하고 있었다. 즉, 유료 VOD들 간의 연관관계가 강한 것이 아니라 유료 VOD들이 대부분 시리즈 물과 성인물로 이루어져 있어서 통계적으로 봤을 때 시리즈 물과 성인물의 강한 연관관계가 유료 VOD들 간의 강한 연관관계인 것처럼 드러난 것이다. 유료 VOD들 중에서 시리즈 물과 성인물을 제외하고 유료 VOD의 비율은 살펴본 결과는 다음 표8과 같았다.

표 8 CF 모델 전체에서 유료 VOD의 비율 (시리즈 물과 성인물 제외)

	조건 VOD가 유료	결과 VOD가 유료
규칙 개수	6,264	436
비율	100%	7%

시리즈 물과 성인물을 제외하고 나자 유료 VOD들 간의 연관관계가 약한 것은 특정 예에서만 드러나는 특수한 현상이 아니라 CF 모델의 규칙 전반에서 나타나는 일반적인 현상임을 확인할 수 있었다.

4.5. 시청 빈도

상품의 인기도(판매개수)는 많은 사람들이 그 상품에 대해 만족한다는 것을 간접적으로 나타내기 때문에 대부분의 상품의 구매에 있어서 큰 영향을 끼치는 또 하나의 요소이다. IPTV 서비스에서 VOD의 인기도는 해당 VOD의 시청 빈도를 통해서 파악할 수 있다.

시청빈도는 VOD에 따라서 매우 변동 폭이 크다. 수회밖에 시청되지 않은 VOD가 있는 반면에 수십만 회 이상 시청된 VOD도 있다. 인기 VOD들 간의 연관관계를 확인하기 위해서는 시청빈도가 높은 인기 VOD를 구분하는 것이 필요한데 임의로 10,000회 이상 시청된 VOD를 인기 VOD로 간주하여 실험을 진행했다. 시청빈도가 10,000회 이상인 VOD 중 임의로 몇 개의 VOD를 선택해서 그 VOD를 시청했을 경우 CF 모델을 통해서 받게 되는 추천결과 리스트를 살펴본 결과 시청빈도가 높은 VOD들 간의 연관관계가 매우 강한 것을 확인할 수 있었다. 이것은 시청빈도가 낮은 VOD들 간에도 마찬가지여서 시청빈도가 낮은 VOD들 간의 연관관계 또한 매우 강한 것을 확인할 수 있었다. (시청빈도와 관련된 자료들은 KT 내부자료와 연관될 수 있기 때문에 부득이 생략했다.)

앞선 분석들과 마찬가지로 이와 같은 현상이 일반적인 현상인지 확인하기 위해서 CF 모델 전체에서 시청한 VOD의 시청빈도가 10,000회 이상이었을 때 그와 유사도가 높은 VOD 역시 시청빈도가 10,000회 이상인 비율을 살펴본 결과는 다음 표9와 같았다.

표 9 CF 모델 전체에서 시청빈도가 높은 VOD의 비율

	조건 VOD가 인기VOD	결과 VOD가 인기VOD
비율	100%	70%

이와 같은 결과는 시청빈도가 낮은 VOD들 간에도 마찬가지여서, 시청빈도가 낮은 VOD들 간의 연관관계가 강하게 나타났다.

4.6. 주연배우 / 감독

모든 상품은 구입하기 전에는 자신이 원하는 상품인지 완전하게 확인할 수 없기 때문에 상품에 대한 정보를 통해 그 상품을 선택할 것인지를 판단한다. 특히 VOD와 같은 무형 컨텐츠 상품의 경우에는 그 컨텐츠의 메타데이터가 그 컨텐츠를 선택하기 전 그 컨텐츠를 시청할 것인지에 대해 판단할 수 있는 정보이다. 특히 가격이나 시청등급 등을 제외하고 VOD의 내용과 관련해서 사람들이 VOD를 선택하는 데 많이 사용하는 메타데이터는 카테고리, 출연배우, 감독 등이다. 많은 사람들은 자신이 선호하는 배우가 출연한 작품 혹은 자신이 좋아하는 작품을 만든 감독의 신작 등에 호감을 가지고 시청하는 경우가 많다. 따라서 동일한 주연배우 혹은 동일한 감독의 VOD는 일반적으로 연관관계가 강할 것이라고 예상하는 대표적인 요소이다.

실제 CF 모델에도 이와 같은 주연배우와 감독의 특징적인 연관관계가 드러나는지 확인하기 위해 앞선 분석들과 마찬가지로 특정 주연배우의 VOD를 임의로 몇 개 선택해서 그 VOD를 시청했을 경우 CF 모델을 통해서 받게 되는 추천결과 리스트를 살펴본 결과는 다음 표10과 같았다.

표 10 CF 모델에서 주연배우가 같은 결과의 예

시청한 VOD 제목	복수는 나의 것	송강호	역도산	설경구
결과리스트 항목	생활의 발견	김상경	090509[토]1273회 연예가 중계	한석준
	구타유발자들	한석규	와일드 카드	정진영
	여자는 남자의 미래다	유지태	생방송 TV연예_151화	서경석
	아름답다	차수연	090124[토]1258회 연예가 중계	한석준
	오프 로드	조한철	신라의 달밤	차승원
	박하사탕	설경구	15강_[입문] 일본어뱅크 회화 35	
	쥬얼리		국내영화_바람의 전설	이성재
	경계	서정	국내영화_나비	김민종
	초록물고기(상)	한석규	국내영화_전투의 매너HD	서유정
	바보같은 사랑 19회	이재룡	외국영화_가타카	에단 호크
동일주연 비율 (10위)	0 / 10 = 0%		0 / 10 = 0%	
동일주연 비율 (전체)	1 / 81 = 1%		4 / 55 = 7%	

표10에서 확인할 수 있는 바와 같이 예상과는 다르게

주연배우가 같은 VOD들 간의 연관관계는 거의 없는 것을 확인할 수 있었다. 마찬가지로 감독이 같은 VOD들 간의 연관관계도 거의 없었는데, 이것은 동일 주연배우나 감독의 VOD들 간에 연관관계가 없다기보다는 앞서 살펴본 다른 요소들과 비교해 동일 주연배우나 감독의 연관관계가 상대적으로 약하기 때문인 것으로 보여진다. 왜냐하면 상위 10위권 내에는 동일 주연배우 혹은 동일 감독의 작품이 거의 없었지만, 상위 100위권 내에는 동일 주연배우 혹은 동일 감독의 VOD가 포함되는 경우가 많았기 때문이다. 상위 10위권 내의 결과는 대부분 카테고리가 같거나 시청등급이 같은 VOD가 많았다. 특히 쿡TV 서비스에서 주연배우가 같은 VOD 혹은 감독이 같은 VOD는 시리즈 물을 제외하면 많아야 5~10개 내외인 경우가 대부분이라는 것을 고려한다면 이와 같은 결과를 보고 동일 주연배우 혹은 동일 감독의 VOD들 간에 연관관계가 없다고 판단하기에는 무리가 있다. 게다가 많은 사람들이 아직 정보가 전혀 없는 신작의 경우에는 아무런 정보가 없기 때문에 주연배우나 감독을 보고 VOD를 선택하지만, 구작의 경우에는 주연배우나 감독 보다 다른 사람들의 그 VOD에 대한 평가에 강한 영향을 받는 경향이 있다는 점도 고려해야 할 것이다.

앞선 분석들과 마찬가지로 이와 같은 현상이 일반적인 현상인지 확인하기 위해서 CF 모델 전체에서 시청한 VOD가 특정 배우가 주연인 VOD였을 때 그와 유사도가 높은 VOD 역시 특정 배우가 주연인 비율을 살펴본 결과는 다음 표11과 같았다. (시리즈 물로 인한 왜곡을 없애기 위해 시리즈 물은 제외했다.)

표 11 CF 모델 전체에서 주연배우가 동일한 VOD의 비율

	조건 VOD가 특정 주연	결과 VOD가 동일 주연
규칙 개수	56,436	891
비율	100%	2%

위의 결과는 동일 감독의 경우에도 4%로 마찬가지였다. 주연배우가 동일하거나 감독이 동일한 VOD들 간의 연관관계가 상대적으로 약한 것은 특정 예에서만 드러나는 특수한 현상이 아니라 CF 모델의 규칙 전반에서 나타나는 일반적인 현상임을 확인할 수 있었다. 하지만 CF 모델이 상위 100위까지의 결과를 가지고 있는 반면 동일 주연배우나 동일 감독의 VOD는 많아야 5~10개 내외인 것을 고려하면 주연배우가 동일하거나 감독이 동일한 VOD들 간의 연관관계가 없다고 보기는 어렵다.

5. 결론

지금까지 우리는 IPTV 영역에서 시청자의 시청로그에 기반하는 협력필터링(Collaborative Filtering)을 적용함에 있어서 크게 영향을 끼치는 요소가 무엇인지 살펴봤다.

무엇보다 영향력이 컸던 요소는 시리즈 물로서 VOD의 절대적인 개수 자체도 많았을 뿐만 아니라 시리즈 내 VOD들 간의 연관관계가 매우 강했다. 특히 시리즈 물은 기존에 CF가 많이 적용되던 영화나 상품 영역에는 거의 존재하지 않는 요소로서 TV 방송을 아우르는 IPTV 영역에 CF를 적용하기 위해서는 가장 중요하게 고려해야 할 요소라는 것을 알 수 있었다.

한편 시청등급, 카테고리, 시청빈도 등은 예상대로 연관관계가 강하게 드러났지만, 반대로 가격, 주연, 감독 등은 예상과는 다르게 연관관계가 약한 것으로 드러났다. 특히 데이터를 구분하지 않고 CF 모델을 만들었음에도 불구하고 동일한 시청등급, 카테고리의 VOD들 간에는 강한 연관관계가 존재하는 것을 확인함으로써 시청등급 별 혹은 카테고리 별로 CF 모델을 따로 생성하는 것이 불필요하다는 것을 알 수 있었다. 또한 주연배우와 감독의 예 등에서 드러났듯이 VOD들 간의 연관관계에 영향을 끼칠 수 있는 요소가 동시에 여러 개가 존재하면 좀더 영향력이 강한 요소에 영향을 받는 VOD들이 더 높은 유사도를 가지게 됨을 알 수 있었다.

이와 같은 결과를 통해 협력필터링(Collaborative Filtering)은 별도의 인위적인 방법을 적용하지 않더라도 카테고리 등의 VOD 컨텐츠 메타데이터와 연령 등의 시청자 프로파일 정보를 상당부분 반영하고 있음을 알 수 있다. 이것은 시청자들이 IPTV를 시청할 때 자신의 연령이나 VOD 컨텐츠의 카테고리 등을 통해 VOD를 선택하는 시청패턴을 가지기 때문인 것으로 보인다. 특히 이와 같은 결과는 VOD 컨텐츠 메타데이터나 시청자의 프로파일 정보를 구할 수 없는 환경에서도 시청자의 시청로그만 가지고 있으면 해당 요소들이 반영된 CF 모델을 만들 수 있음을 의미한다고 할 수 있다.

6. 참고문헌

1. Recommender system. “ Wikipedia, the free encyclopedia. ” [온라인] http://en.wikipedia.org/wiki/Recommendation_system.
2. “Amazon. ”[온라인] http://www.amazon.com/.
3. *Amazon.com Recommendations: Item-to-Item Collaborative Filtering.* Greg Linden, Brent Smith and Jeremy York. s.l. : IEEE Computer Society, 2003.
4. “Digg.com. ”[온라인] http://digg.com/.
5. “ Internet Movie Database. ” [온라인] http://www.imdb.com/.
6. “Last.fm. ”[온라인] http://www.last.fm/.
7. “Netflix. ”[온라인] http://www.netflix.com/.
8. “Tivo. ”[온라인] http://www.tivo.com/.