R 시스템에서 협업 필터링과 K-NN 을 이용한 미국 드라마 추천 시스템

주완수*, 이한형*, 일홈존*, 박두순*
*순천향대학교 컴퓨터소프트웨어공학과
e-mail : snaku95@sch.ac.kr

American Drama Recommendation System using Collaborative Filtering and K-NN in R System

Wan-Su Joo*, Han-hyung Lee*,
Ilkhomjon*, Doo-Soon Park*
*Dept. of Computer Software Engineering, SoonchunHyang University

요 약

스마트 폰과 태블릿 PC를 이용하여 실시간 영상 재생 서비스(OTT: Over The Top)를 이용하는 사람들이 폭발적으로 증가하고 있다. 그에 따라 실시간 영상 재생 서비스를 즐길 수 있는 수많은 콘텐츠들이 증가하고 있다. 이에 따라 사용자는 자신의 취향에 맞는 드라마가 어떤 드라마인지 찾기가 어렵다. 따라서 본 논문에서는 사용자 스타일에 가장 적합한 미국 드라마 추천 시스템을 제안하기 위하여 선호 장르 2개, 연령대, 성별, 미국인 여부를 이용하여 유클리드 방법으로 유사도를 계산하고 협업 필터링 방법을 적용하여 드라마를 추천하는 시스템을 R을 이용하여 구현하였다.

1. 서론

최근 TV, PC, 스마트폰, 태블릿 등 콘텐츠를 소비할 수 있는 디바이스가 확대되면서 이용자들은 시간과 장소에 제약받지 않고 언제 어디서나 자신이 원하는 콘텐츠를 즐길 수 있는 환경이 형성되었다. 이에 따라 언제 어디서나 어떠한 단말기로도 콘텐츠를 이용할 수 있는 환경이 구성되었고 이에 따라 인터넷을 기반으로 한 실시간 영상 재생 서비스가 빠르게 정착하였다. 실시간 영상 재생 서비스 산업의 성장과 함께 코드 커팅 추세가 확산되면서 미국의 넷플릭스 온라인 스트리밍 서비스가 기존 유료방송의 가입자를 추월하는 등 실시간 영상 재생서비스 서비스 동영상 산업의 시장 지배력은 더욱 커질 것으로 예상된다[1].

또 한 국내 실시간 영상 재생 서비스 이용자 중에서 주 5일 이상 실시간 영상 재생 서비스를 이용하는 비율은 2016년에는 24.1%, 2017년에는 30.8%, 2018년에는 36.0%까지 증가하였다[2]. 또한, 국내 실시간 영상 재생서비스에 대한 시장 규모 추이는 (그림 1)과 같다.

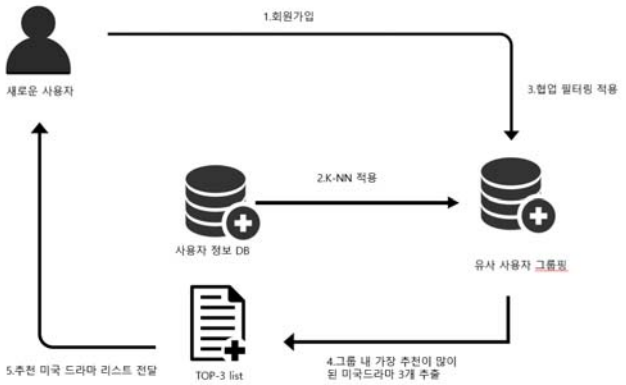
(그림 1) 국내 OTT 시장 규모 추이[3].

이러한 실시간 영상 재생 서비스 산업의 발전과 더불어 미국 TV 드라마 콘텐츠의 증가로 인해 사용자에게 폭 넓은 선택지가 생겼지만 사용자들은 자신의 취향에 맞는 드라마가 어떤 드라마인지 잘 모르는 경우가 많다. 이러한 이용자들의 편의를 위하여 K-NN을 사용하여 그룹을 나누어 사용자들이 원하는 드라마 추천 시스템을 보다 효율적으로 구현하고자 한다.

* corresponding author : 박두순
- 이 논문은 한국연구재단의 지원을 받아 수행되었음(No. NRF-2017R1A2B1008421)

2. 미국드라마 추천시스템의 구성

본 논문에서 구현하게 될 미국드라마 추천시스템의 시나리오는 (그림 3)과 같다.



(그림 3) 추천시스템 시나리오

(그림 3)의 추천시스템 시나리오의 설명은 다음과 같다. 사용자에게 미국드라마를 추천하기에 앞서 사용자에게 입력받은 개인정보를 통해 사용자와 유사한 사용자를 추출하는 방법과 미국드라마 간의 근접 정도를 측정하여 유사한 미국드라마를 추천하는 방법이 있다. 그중 본 논문에서는 사용자 간의 유사 정도와 미국드라마 간 유사 정도를 최근접 이웃 분류방식으로 계산하여 유사한 사용자 리스트를 추출하는 방법을 선택했다.

(1) 추천시스템을 사용하기에 앞서 사용자는 회원가입을 한다. 회원가입 시 개인정보로 사용하게 될 선호 장르 1, 선호 장르 2, 연령대, 성별, 미국인 여부를 입력한다. 연령대 분류와 장르 분류는 미국드라마 순위 사이트인 IMDb의 분류 방법을 따라 하여 연령대는 0~18, 18~29, 30~44, 45세 이상으로 구분하였고, 장르는 액션 어드벤처, 애니메이션 등 26가지의 장르로 구분하였다[4]. 개인화 요소를 입력받고 사용자에게 선호하는 미국드라마를 입력받는다.

(2) 그 후 최근접 이웃 알고리즘을 통해 분류되어 있지 않은 레코드들이 분류된 레코드 중 가장 비슷한 속성을 가진 레코드로 할당한다. 이때, 레코드 분류 기준은 유클리드 거리 계산 공식을 사용한다.

(3) 협업 필터링이란, 사용자들의 선호도와 관심 표현을 바탕으로 선호도, 관심도가 비슷한 사용자들을 식별해 내는 방법으로 과거에 이용한 콘텐츠가 비슷하다면 사용자 간에 유사한 성향을 가지고 있다고 판단하고 그 근거를 토대로 이루어진다.

(4) 추출된 유사 사용자가 추천하는 미국드라마 중에서 가장 많이 추천된 미국드라마를 추천한다. 중복되는 드라마가 없다면 유사도가 가장 높은 사용자가 추천한 미국드라마를 추천한다[5].

(5) 최종적으로 사용자에게 미국드라마를 추천해준다.

협업 필터링이란, 사용자들의 선호도와 관심 표현을 바탕으로 선호도, 관심도가 비슷한 사용자들을 식별하는 방법으로 과거에 이용한 콘텐츠가 비슷하다면 사용자 간에 유사한 성향을 가지고 있다고 판단하여 그 근거를 토대로 이루어진다.

이렇게 판별된 데이터를 바탕으로 추천받을 사용자와 가장 유사한 사용자 3명을 추출하며 추출된 유사 사용자가 추천하는 미국드라마 중에서 가장 많이 추천된 미국드라마를 추천한다.

3. 미국드라마 추천시스템의 구현

본 논문에서 구현한 미국드라마 추천시스템은 협업 필터링의 희소성 문제를 해결하기 위해 회원가입이 필수로 되어야 이용할 수 있다. 회원가입을 통해 사용자는 개인화 요인을 입력하게 되는데 요소는 선호 장르 1, 선호 장르 2, 연령대, 성별, 미국인 여부, 선호하는 미국드라마로 이루어져 있다. 회원가입의 양식은 (그림 4)와 같다.

(그림 4) 회원가입 양식

다음과 같은 양식에 따라서 사용자가 입력한 정보는 회원 관리 테이블에 저장되고, 회원 관리 테이블은 ID, 선호 장르 1, 선호 장르 2, 연령대, 성별, 미국인 여부, 선호하는 미국드라마로 구성되어있다. 선호 데이터 처리는 R에서 시행하였다. 회원 정보를 수치화한 데이터베이스는 1,000명의 트레이닝용 더미 데이터를 사용하였다. 트레이닝용 더미데이터의 일부는 (그림 5)와 같다.

Filter

	NUM	ID	Age	Sex	American	Genre 1	Genre 2	Favorite
1	1	20189511	53	1	2	5	13	Luchshe, chem lyudi
2	2	20148155	59	2	1	15	19	BH90210
3	3	20166510	61	2	1	17	25	It's Always Sunny in Philadelphia
4	4	20120669	28	2	1	22	14	The Terror
5	5	20167183	26	2	2	15	8	Big Little Lies
6	6	20152449	20	1	1	4	20	Friends
7	7	20114313	46	2	1	19	19	Brassic
8	8	20186446	28	2	2	3	13	Agents of S.H.I.E.L.D.
9	9	20158403	60	2	2	25	17	Deep Water
10	10	20123260	57	2	2	12	3	House M.D.
11	11	20140263	59	2	2	9	24	The Expanse
12	12	20149697	37	1	1	2	11	Fear the Walking Dead
13	13	20190512	46	2	1	20	11	True Detective
14	14	20146804	45	2	2	19	9	The Blacklist
15	15	20110020	28	2	2	14	5	Westworld
16	16	20164918	42	2	2	24	20	Wu Assassins
17	17	20106085	16	1	1	11	3	The Mandalorian

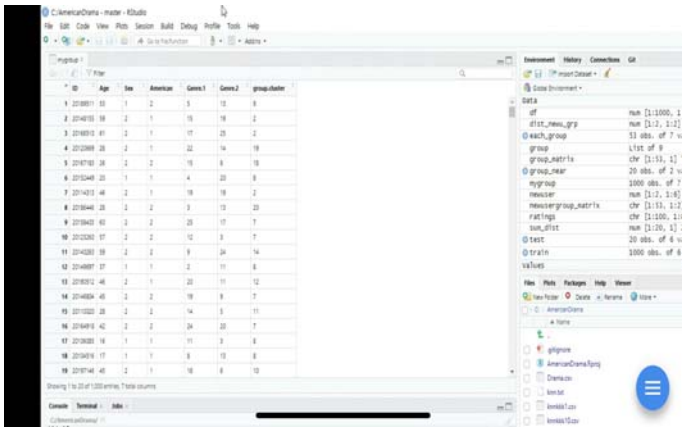
(그림 5) 회원 정보 데이터베이스의 일부

본 논문에서는 사용자가 회원가입 할 때 입력한 개인화 요인을 각각 선호 장르 1, 선호 장르 2, 연령대, 성별, 미국인 여부로 유사도를 측정하였다. 이러한 데이터를 토대로 사용자와 다른 사용자 간의 유사도를 유클리디안 거리 점수 기반 유사도를 통해 구한다. 유클리디안 거리 점수 기반 유사도를 이용한 수식은 아래 (수식 1)과 같다.

$$\sqrt{(p1-q1)^2+(p2-q2)^2+(p3-q3)^2+(pn-qn)^2} = \sqrt{\sum_{i=1}^{n}(pi-qi)^2}$$

(수식 1) 유클리디안 거리 점수 기반 유사도 공식

이렇게 계산된 유사도는 사용자와 비슷한 성향의 사용자가 있는 그룹 안에서 비교를 하게 되는데, 이때 사용하는 것이 최근접 이웃 알고리즘(K-NN)이다. 최근접 이웃 알고리즘은 얼마나 거리가 가까운 데이터와 비교를 할 것인가를 판단하고 그 거리 안에서 그룹핑하여 유사도를 비교한다. R에서 그룹핑을 할 때 1000명의 유저, 10개의 그룹으로 설정하였다. 그 이유는 10개의 그룹일 때 가장 좋은 정확성을 보였다. 그룹핑을 한 결과는 (그림 6)과 같다.



(그림 6) 그룹핑 결과

그룹핑을 하여 유사도를 비교하고 그 거리를 k라고 한다. k를 정하기 전에 선행되어야 하는 작업이 있다. 바로 표준화이다. K-NN 에서 가깝다는 개념은 유클리드 거리(Euclidean Distance)로 정의한 데, 유클리드 거리를 계산할 때는 단위가 매우 중요하다 [5]. 유클리드 거리 계산에 사용되는 공식은 (수식 2)와 같다.

$$\sqrt{(Ax - Bx)^2 + (Ay - By)^2}$$

(수식 2) 유클리드 거리 계산 공식

이러한 유클리디안 거리 점수 기반 유사도 공식을 이용하여 개인화 요소를 나이가 23살이고, 성별은 남자, 국적이 미국인이 아니고 선호하는 장르는 판타지, 코미디를 가진 사용자의 정보는 (그림 7)과 같다.

ID	Age	Sex	American	Genre 1	Genre 2	Favorite
20154641	23	1	1	10	5	Glee

(그림 7) 사용자의 정보

사용자의 유사도를 측정한 결과를 토대로 사용자에 대한 그룹 내 Top-3 리스트를 나타내면 (그림 8)과 같다.

Drama	Count
Doctor Who	11
The Naked Director	7
It's Always Sunny in Philadelphia	6

(그림 8) 사용자의 Top-3 List

이 사용자의 개인화 요소 중에 나이를 30살로 변경한 사용자의 정보는 (그림 9)와 같다.

ID	Age	Sex	American	Genre 1	Genre 2	Favorite
20154641	30	1	1	10	5	Glee

(그림 9) 나이 정보를 변경한 사용자의 정보

나이 정보를 변경 후 그룹 내 Top-3 리스트를 나타내면 (그림 10)과 같다.

Drama	Count
The Flash	13
On Becoming a God in Central Flori	10
Schitt's Creek	6

(그림 10) 나이 정보 변경 후 Top-3 List

마지막으로 이 사용자의 개인화 요소 중에 성별을 여자로 변경한 사용자의 정보는 (그림 11)과 같다.

ID	Age	Sex	American	Genre 1	Genre 2	Favorite
20154641	30	2	1	10	5	Glee

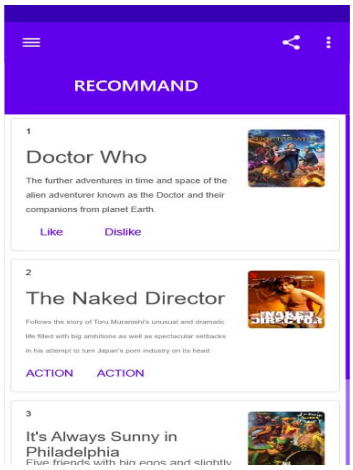
(그림 11) 성별을 변경한 사용자의 정보

성별을 변경 후 그룹 내 Top-3 리스트를 나타내면 (그림 12)와 같다.

Drama	Count
Jack Ryan	8
Homeland	7
The Flash	4

(그림 12) 성별 정보 변경 후 Top-3 List

이렇게 추천된 미국드라마는 사용자에게 미국드라마의 설명과 사진을 통해 제공되며 미국드라마를 추천한 사용자들의 수와 실제 사용했던 사용자들이 부여한 평점을 볼 수 있도록 구현되었다. 나이가 23살이고, 성별은 남자, 국적이 미국인이 아니고 선호하는 장르는 판타지, 코미디를 가진 사용자에게 추천된 미국드라마는 (그림 13)과 같다.



(그림 13) 추천 결과 화면

4. 결론

본 논문에서는 수많은 미국드라마 중에서 사용자의 취향에 맞는 미국드라마를 보다 효율적으로 추천하기 위해 사용자를 그룹핑하기 위한 K-NN 알고리즘과 협업 필터링을 기반으로 미국드라마를 추천해 주는 프로그램을 구현하였다. 그리고 기존의 사용자 기반 협업 필터링의 문제 중 희소성 문제에 대해서 보다 적합한 추천을 위해 사용자 간에 유사도 측정에 있어 유클리드 거리 기반 유사도 공식을 이용하였다.

향후 연구 과제는 협업 필터링의 문제점으로 거론되고 있는 Cold Start를 효율적으로 극복하기 위한 방안을 갈구해야 할 것이다. 본 논문에서는 그 방법으로 K-NN 알고리즘을 채택하였으나 조금 더 효율적인 방법에 대해 연구할 필요가 있다.

참고문헌

[1] 이해미, “국내외 OTT 서비스 현황 및 콘텐츠 확보 전략 분석”, 정보통신산업진흥원, 이슈리포트 2018-제34호 p1~p16, 2018.08

[2] 이선희, “온라인 동영상 제공 서비스(OTT) 이용 행태 분석, 정보통신정책연구원”, p1~p13, 2019.05.15.

[3] https://brunch.co.kr/@yamju/326

[4]IMDb-https://www.imdb.com/chart/tvmeter?sort=rk,asc&mode=simple&page=1

[5] 신해란, 주완수, 박두순, "협업 필터링과 K-nn을 이용한 모바일 게임 추천 시스템", 2019년 춘계학술발표대회 논문집 제25권 제1호, p283~p286, 2019, 05