오피니언 마이닝과 협업필터링을 이용한 웹툰 추천 시스템

심대수*, 박진수*, 박두순*
*순천향대학교 컴퓨터소프트웨어공학과
e-mail:tlaeotn123@naver.com

A Webtoon Recommendation System using Opinion Mining and Collaborate Filtering

Dae-Su Sim*, Jin-Soo Park*, Doo-Soon Park*
*Dept. of Computer Software Engineering, SoonChunHyang University

요 약

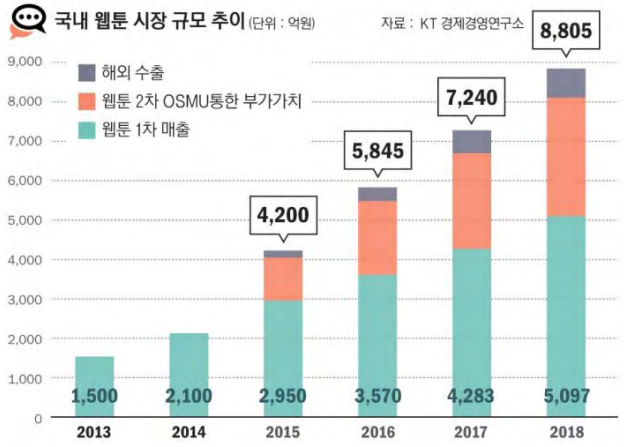
최근 다양한 웹툰 콘텐츠의 증가와 함께 스마트폰 보급률이 높아지면서, 사용자들의 실시간 웹툰 서비스의 이용이 증가하고 있다. 웹툰 콘텐츠의 가치가 갈수록 점점 높아지고 있으며, 각종 영화•애니메이션•게임 등 다양한 콘텐츠 사업에 많은 데이터가 사용되고 있다.

본 논문에서는 기존 웹툰의 리뷰를 오피니언 마이닝기법을 사용하여 각 웹툰의 선호도를 평가하며 나이, 성별, 선호 장르, 선호 웹툰 플랫폼 등과 같은 개인 성향을 통하여 사용자간의 유사도를 측정하는 협업 필터링 방법을 적용해 각각의 사용자들이 보고 싶어하는 웹툰을 자동적으로 추천해주는 웹툰 추천 시스템을 제안한다.

1. 서론

최근 스마트폰의 보급과 인터넷 환경의 성장에 따라 많은 스마트폰 사용자들이 장소와 시간에 구애받지 않으며 인터넷을 사용할 수 있게 되었다. 이에 따라 각 유명한 인터넷 포털의 콘텐츠중 하나인 '웹툰'이라는 콘텐츠 또한 자연스럽게 성장하게 되었으며, '웹툰'이라는 콘텐츠는 단순히 인터넷 만화 시장을 넘어서 드라마, 게임, 영화, 마케팅 시장 등으로 확산되고 있으며 어느새 대중들의 주목을 받는 하나의 트렌드로 자리 잡았다.

이와 같이 '웹툰' 콘텐츠에 나오는 여러 캐릭터들이나 시나리오 들이 하나의 트렌드가 되면서 인기를 얻고있으며 웹툰을 원작으로 만들어진 영화나 드라마가 크게 성공하게되며 웹툰의 파급력이 확대되어, 광고 플랫폼으로써 웹툰을 활용하는 사례마저 증가하고 있는 추세이다. (그림 1)은 국내 웹툰 시장 규모 추이를 나타낸 것이다[1]. 이러한 트렌드를 따라가기 위해 웹툰에 대한 많은 대중들의 관심이 높아지기 시작했다.



(그림 1) 국내 웹툰 시장 규모 추이 ('13~18)

기존의 웹툰 서비스는 대부분의 유통이 인터넷 포털 서비스 중 한 부분으로 제한적 이었으나, 최근 웹툰을 메인 콘텐츠로 하는 웹툰 전문 플랫폼이나 모바일 웹툰앱의 발전으로 규모가 확장되었다. 또한 플랫폼의 다양성을 추구하게 되어 보다 시장규모가 커지는 계기가 되었고, 이런 웹툰 전문 플랫폼은 대부분 유료 서비스를 제공하고 있으며, 이러한 유료 정책은 웹툰 시장의 매출과 작가의 수입성을 높여주게 되어 웹툰 시장의 전망은 갈수록 규모가 증가해가는 추세이다.

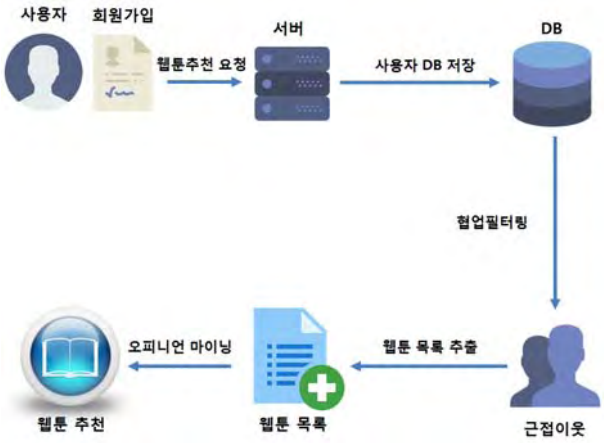
본 논문에서는 기존의 사용자 기반 협업필터링의 문제중 Cold Start 문제에 대해서 조금 더 나은 방안을 제시해

※ 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업의 연구결과로 수행되었음 (IITP-2016-H8601-16-1009)

보고자 한다. 협업 필터링에 대한 데이터가 적을 시 발생하는 Cold Start에서는 사용자의 유사도를 나이, 직업, 성별, 선호 장르, 선호 웹툰 플랫폼 등과 같은 개인 성향을 Jaccard Similarity로 계산하고, 여러 가중치를 두어 결과를 비교해본 뒤, 사용자들이 웹툰 서비스를 이용한 후에 평가한 리뷰를 오피니언 마이닝을 통하여 사용자들의 선호도를 구하여 기존의 시스템보다 정확도 높게 개인에게 웹툰을 추천해주는 시스템을 제안한다.

2. 웹툰 추천 시스템의 구성

본 논문에서 구현하게 될 추천 시스템 시나리오는 (그림 2)와 같다.



(그림 2) 추천 시스템 시나리오

(그림 2)와 같은 추천시나리오처럼 사용자에게 웹툰을 추천하기위해서는 사용자의 평가 목록을 통해 사용자간의 유사도를 측정해 유사한 근접이웃을 구성하거나, 웹툰간의 유사도를 측정하여 유사한 웹툰을 추천하는 방식이 있을 수 있는데, 본 논문에서는 웹툰과 웹툰 사이의 유사도를 측정하는 것이 아닌 사용자간의 유사도 계산하여 근접이웃을 구성하는 방식을 사용한다.

첫 번째로 신규 회원이나 휴먼 회원의 경우 근접이웃을 구성하기 위한 사용자의 개인화 요소를 회원가입을 통해 받게 되는데 이때 개인화 요소는 나이, 직업, 성별, 선호 장르, 선호 웹툰 플랫폼을 입력받는다. 입력받은 사용자의 DB는 서버 데이터베이스에 저장한다.

두 번째로 신규 회원 혹은 휴먼 회원에게 추천할 때 회원가입 시 입력받은 개인화 요소에 협업 필터링을 이용하여 최 근접 이웃을 구성한다. 이때 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 추천하는 방식이다 [2].

세 번째로 최 근접 이웃의 웹툰 데이터를 추출한 뒤 각 웹툰의 긍정/부정 점수가 높은 3편의 웹툰을 추출한다. 이때 긍정/부정 점수는 각각의 웹툰의 리뷰를 크롤링 한 후, SO-PMI를 이용해 각각 계산한다.

네 번째로 긍정 점수가 가장 높은 3편의 웹툰을 추출한 뒤 사용자에게 추천한다.

본 논문에서는 사용되는 개인화 요소를 나이, 직업, 성별, 선호장르, 선호 웹툰 플랫폼으로 분류하였으며 나이는 10살 단위로 연령을 분류하였고, 직업은 통계청의 자료를 토대로 대분류 항목을 기준으로 구성하였다.
개인화 요소의 분류 결과는 (표 1)과 같다.

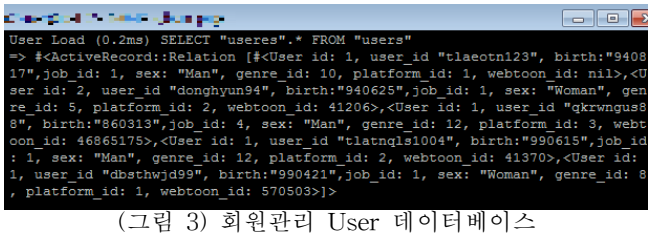
(표 1) 사용자 개인화 요소의 분류[3]

Index	Age	Job	Sex	Genre	Platform
1	10대	학생	남성	액션	네이버
2	20대	관리자	여성	스릴러	다음
3	30대	전문가		공포	카카오 페이지
4	40대	사무직		범죄	레진코믹스
5	50대	서비스업		판타지	탑툰
6	기타	판매직		전쟁	
7		농랑어업 종사자		코미디	
8		기능직		멜로	
9		기계조작 및 조립 종사자		미스터리	
10		노무직		SF	
11		군인		성인물	
12				기타	

3. 웹툰 추천 시스템의 구현

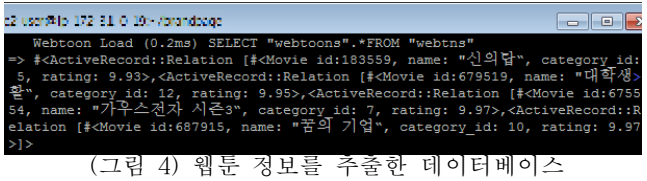
우선적으로 추천 시스템을 구현하기 위해서는 사용자간의 유사도를 판별할 수 있어야 하며, 이를 위해서 회원가입 시 사용자의 개인화 요인을 받아야 하는데 본 논문에서는 개인화 요인을 나이, 직업, 성별, 선호 장르, 선호 웹툰 플랫폼으로 구성하였다. 별개로 각 웹툰의 긍정/부정 점수는 웹툰의 상위 리뷰를 크롤링하여 긍정/부정 점수를 계산하는 방식을 사용한다.

(그림 3)은 5개의 개인화 요소를 회원가입시 사용자에게 받아 DB화 하여 AWS 클라우드 서버에 저장한 모습이다. 저장된 순서는 ID, 나이, 직업, 성별, 선호장르, 선호 웹툰 플랫폼 이다.



(그림 3) 회원관리 User 데이터베이스

웹툰을 추천하기 위해서는 (그림 3)의 회원들의 개인화 요소가 들어있는 DB 뿐만이 아니라, 각 웹툰의 정보에 대한 DB를 가지고 있어야 한다. 본 논문에서는 웹 페이지의 정보를 가져오는 크롤링 방법을 통하여 각각의 포털의 웹툰의 제목, 장르들을 추출해 (그림 4)와 같이 데이터베이스를 구성하였다.



(그림 4) 웹툰 정보를 추출한 데이터베이스

(그림 3)의 사용자 데이터베이스를 이용하여 근접 이웃을 구성할 수 있는데 개인화 요소, 가중치에 따라 모두 근접 이웃의 구성이 달라진다. (표 2)는 사용자 유사도를 나이, 직업, 성별, 장르, 플랫폼 다섯 가지로 Jaccard Similarity를 이용해 근접 이웃을 구성한 사용자 데이터베이스의 일부이다.

(표 2) User_1에 대한 유사도와 근접이웃 구성

ID	Union	Intersection	Similarity	Recommand
user_1	x	x	x	
user_20	6.00	4.00	0.6667	덴마
user_14	7.00	3.00	0.4286	신의 탑
user_18	7.00	3.00	0.4286	기기괴괴
user_28	7.00	3.00	0.4286	조우
user_29	7.00	3.00	0.4286	블랙수트
user_2	8.00	2.00	0.2500	체이서
user_4	8.00	2.00	0.2500	20세 보고서
user_5	8.00	2.00	0.2500	연애혁명
user_7	8.00	2.00	0.2500	신도림
user_8	8.00	2.00	0.2500	여중생A

개인화 요소를 나이, 직업, 성별의 세 가지를 사용했을 때와 나이, 성별, 직업, 장르의 네 가지를 사용했을 때의 근접 이웃의 추천 웹툰은 (표 3), (표 4)와 같다. (표 2)와 (표 3)과 (표 4)를 비교해보면 추천되는 웹툰이 변화하는 것을 알 수 있다.

(표 3) User_1에 대한 유사도와 근접이웃 구성

ID	Union	Intersection	Similarity	Recommand
user_1	x	x	x	
user_28	3.00	3.00	1.0000	조우
user_2	4.00	2.00	0.5000	체이서
user_4	4.00	2.00	0.5000	20세 보고서
user_9	4.00	2.00	0.5000	버스트
user_11	4.00	2.00	0.5000	캠핑은 빡세
user_14	4.00	2.00	0.5000	신의 탑
user_16	4.00	2.00	0.5000	하울링
user_18	4.00	2.00	0.5000	기기괴괴
user_20	4.00	2.00	0.5000	덴마

(표 4) User_1에 대한 유사도와 근접이웃 구성

ID	Union	Intersection	Similarity	Recommand
user_1	x	x	x	
user_20	5.00	3.00	0.6000	덴마
user_28	5.00	3.00	0.6000	조우
user_2	6.00	2.00	0.3333	체이서
user_4	6.00	2.00	0.3333	20세 보고서
user_9	6.00	2.00	0.3333	버스트
user_11	6.00	2.00	0.3333	캠핑은 빡세
user_14	6.00	2.00	0.3333	신의 탑
user_16	6.00	2.00	0.3333	하울링
user_18	6.00	2.00	0.3333	기기괴괴

이번에는 나이, 성별, 직업, 장르, 플랫폼에 가중치를 두어 추천한다. 나이(20%), 성별(10%), 직업(10%), 장르(20%), 플랫폼(40%) 와 나이(40%), 성별(10%), 직업(20%), 장르(10%), 플랫폼(20%) 그리고 나이(20%), 성별(20%), 직업(40%), 장르(10%), 플랫폼(10%)의 가중치를 주었을 때 추천된 웹툰들은 (표 5), (표 6), (표 7)과 같다. (표 5)은 가중치 없이 5개의 개인화 요소를 Jaccard Similarity로 유사도를 구하였을 때보다 추천된 웹툰이 많이 줄어든 것을 볼 수 있고 이는 플랫폼에 민감함을 알 수 있다. (표 6)은 가중치를 주지 않았을 때와 가장 유사한 추천을 하고 있다. 이는 나이에는 민감하지 않다는 것이다. (표 7)은 약간의 차이는 보이지만 나이보다는 민감하지만 일반적인 경향임을 알 수 있다.

(표 5) User_1에 대한 유사도와 근접이웃 구성

ID	Union	Intersection	Similarity	Recommand
user_1	x	x	x	
user_20	1.10	0.90	0.8182	덴마
user_14	1.30	0.70	0.5385	신의 탑
user_18	1.30	0.70	0.5385	기기괴괴
user_15	1.40	0.60	0.4286	노블래스
user_29	1.40	0.60	0.4286	블랙수트
user_5	1.50	0.50	0.3333	연애혁명
user_7	1.50	0.50	0.3333	신도림
user_8	1.50	0.50	0.3333	여중생A
user_13	1.50	0.50	0.3333	기기괴괴

(표 6) User_1에 대한 유사도와 근접이웃 구성

ID	Union	Intersection	Similarity	Recommand
user_1	x	x	x	
user_14	1.20	0.80	0.6667	신의 탑
user_20	1.20	0.80	0.6667	덴마
user_28	1.30	0.70	0.5385	조우
user_18	1.30	0.70	0.5385	기기괴괴
user_2	1.40	0.60	0.4286	체이서
user_9	1.40	0.60	0.4286	버스트
user_29	1.50	0.50	0.3333	블랙수트
user_5	1.60	0.40	0.2500	연애혁명
user_8	1.60	0.40	0.2500	여중생A
user_26	1.60	0.40	0.2500	너란 남자

(표 7) User_1에 대한 유사도와 근접이웃 구성

ID	Union	Intersection	Similarity	Recommand
user_1	x	x	x	
user_28	1.20	0.80	0.6667	조우
user_29	1.30	0.70	0.5385	블랙수트
user_14	1.30	0.70	0.5385	신의 탑
user_4	1.40	0.60	0.4286	20세 보고서
user_11	1.40	0.60	0.4286	캠핑은 빡세
user_16	1.40	0.60	0.4286	하울링
user_24	1.40	0.60	0.4286	레바툰
user_2	1.40	0.60	0.4286	체이서
user_9	1.40	0.60	0.4286	버스트

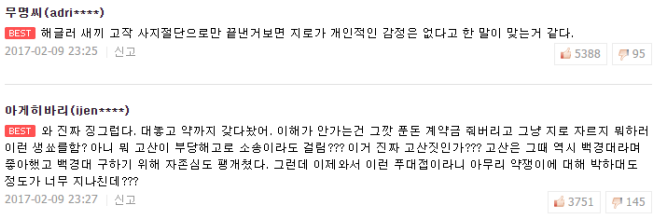
본 논문에서는 여러 가중치를 실험한 후 개인화 요인을 나이, 성별, 직업, 장르, 플랫폼 5가지의 개인화 요소를 사용하였으며, 각각의 가중치는 나이(20%), 성별(15%), 직업(15%), 선호 장르(25%), 플랫폼(25%) 으로 가중치를 두

어 근접 사용자를 추출한다. 또한 본 논문에서는 추천을 하기 전 마지막 세 명의 사용자가 추천한 웹툰에 대해 오피니언 마이닝기법을 사용하여 긍정/부정 점수를 측정하여 추천한다. (표 8)는 각 웹툰의 긍정 부정점수의 일부이다.

(표 8) 각 웹툰의 긍정/부정 점수의 일부

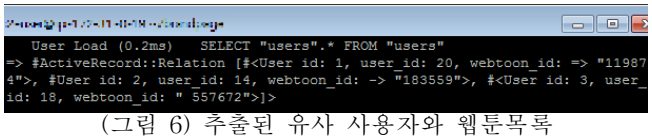
ID	Union	Intersection	Similarity	Recommand	Score
user_1	x	x	x		
user_20	1.15	0.85	0.7391	덴마	-6
user_14	1.40	0.60	0.4286	신의 탑	-2
user_18	1.40	0.60	0.4286	기기괴괴	-3
user_29	1.45	0.55	0.3793	블랙수트	-3
user_28	1.50	0.50	0.3333	조우	0
user_15	1.50	0.50	0.3333	노블래스	-2
user_5	1.60	0.40	0.2500	연애혁명	1
user_8	1.60	0.40	0.2500	여중생A	0

위와 같이 긍정/부정 점수의 대부분이 음수값인 부정에 가까웠는데 그 이유는 웹툰의 리뷰에 있어서 대부분 긍정적인 단어가 신조어로 표현하는 경우가 많았으며, 또한 대부분의 리뷰가 웹툰의 장르에 따라 긍정/부정이 크게 바뀌게 된다. (그림 5)는 추천 웹툰의 리뷰의 일부 이다.

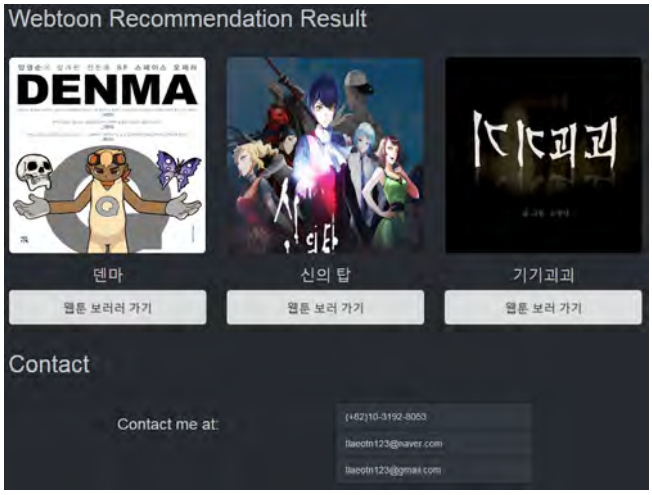


(그림 5) 추천 웹툰 리뷰의 일부

최종적으로 본 논문에서는 나이, 성별, 직업, 장르, 플랫폼 5가지의 개인화 요소를 사용하였으며, 각각의 가중치는 나이(20%), 성별(15%), 직업(15%), 선호 장르(25%), 플랫폼(25%) 으로 가중치를 두고 근접 사용자를 추출한 뒤 각 웹툰의 긍정/부정점수를 측정해 가장 높은 긍정점수를 가진 웹툰을 강조 추천하였다. (그림 6)은 추출된 유사 사용자와 웹툰 목록이며 (그림 7)은 20대, 남자, 학생, SF, 네이버 라는 개인화 요소를 두었을 때 추천되는 웹툰을 나타낸 것이다.



(그림 6) 추출된 유사 사용자와 웹툰목록



(그림 7) 사용자에게 추천된 웹툰 결과

(그림 7)에서 웹툰 보러가기 버튼을 클릭하면 해당 웹툰으로 이동된다.

4. 결론

본 논문에서는 수많은 웹툰 정보들 사이에서 사용자 개인에게 적합한 웹툰을 추천하기 위하여 협업 필터링 방식을 사용하며 새로운 사용자와 휴먼 사용자에게 추천을 할 때 발생하는 Cold Start를 보안하기 위해 개인화 요소를 이용한 협업 필터링과 오피니언 마이닝을 이용한 긍정/부정 점수를 이용해 추천 시스템을 구현 하였다.

개인화 요소를 나이, 성별, 직업의 세 가지를 사용했을 때와 나이, 성별, 직업, 장르의 네 가지를 사용했을 때를 비교해 보면 추천된 웹툰이 변화하는 것을 알 수 있다.

다음으로는 개인화 요소에 가중치를 준 경우와 가중치를 부여하지 않은 경우에도 추천이 달라짐을 볼 수 있다.

참고문헌

[1] 국내 웹툰 시장 규모 추이, kt경제경영연구소, 2015.

[2] 김영아, 박두순, “협업 필터링 기반 드라마 추천 시스템”, 한국정보처리학회 추계학술대회 발표 논문집, 제주한라대학교, pp. 1137-1138, 2013.11

[3] “한국표준 직업분류 표_대분류”, 통계청, 2015.