하둡에서 개인 성향을 이용한 웹툰 추천 시스템

이건호*, 윤원탁*, 황동현*, 박두순*
*순천향대학교 컴퓨터소프트웨어공학과
e-mail:cjdtjcjdtj@naver.com

A Webtoon Recommendation System Using Personal Propensity in Hadoop

Keon-Ho Lee*, Won-Tak Yoon*, Dong-Hyun Hwang*, Doo-Soon Park*
*Dept. of Computer Software Engineering, SoonChunHyang University

요 약

최근 국내의 콘텐츠 생산률이 증가함에 따라, 많은 사람들이 즐길 수 있는 콘텐츠들이 많아 졌다. 하지만 사람들은 많아진 콘텐츠로 인해, 오히려 원하는 정보를 빠른 시간에 얻는 것이 힘들어졌다. 이러한 문제를 해결하기 위해 다양한 방식의 새로운 서비스들이 제공 되고 있다.

추천 시스템 중에서 웹툰을 추천해주는 알고리즘으로 협업필터링 방법이 가장 많이 사용되고 있다. 협업필터링 방법에는 희박성과 확장성, 투명성의 문제점들을 가지고 있다. 따라서 본 논문에서는 협업필터링 방법의 희박성 문제를 보완하고자 개인의 성향을 반영하여 효율이 좋은 웹툰 추천 시스템을 제안하고, 하둡 시스템에서 구현한다.

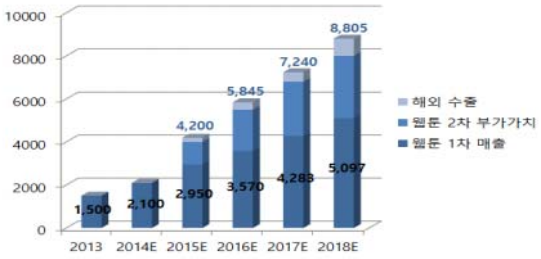
1. 서론

최근 국내의 콘텐츠 생산률이 증가함에 따라, 많은 사람들이 즐길 수 있는 콘텐츠들이 많아 졌다. 사람들은 스마트폰을 사용하여 간단하게 콘텐츠를 접하게 되었고, 그 중 친숙하게 다가갈 수 있는 많은 웹툰들이 이목을 끌게 되었다. 웹툰은 인터넷을 뜻하는 '웹(Web)'과 만화를 뜻하는 '카툰(Cartoon)'의 합성어로서 인터넷 및 모바일 환경에 게제 목적을 가지고 제작된 만화 장르중 하나를 뜻한다[1].

이렇게 사람들의 많은 이목이 집중되자 대형 기획사들은 제작 초기부터 캐릭터, 영화, 드라마 등 다양한 콘텐츠 산업 연계가 가능한 웹툰을 개발하기 시작하였으며, 웹툰 데이터의 가치와 신뢰성도 점점 높아지고 있어, 다양한 사업 분야에 많은 데이터가 사용되고 있다.

이와 같이 웹툰은 단순 만화가 아닌 엔터테인먼트로서 플랫폼 및 시장이 계속 성장하고 있다. (그림 1)은 최근 3년간 우리나라 만화산업의 매출액 규모이고, (그림 2)는 KT경제경영연구소에서 예측한 웹툰 시장 규모 추이이다[1].

(그림 1) 최근 3년간 우리나라 만화산업의 매출액 규모



(그림 2) 웹툰 시장 규모 추이 (2013~2018)

본 논문에서는 추천 시스템에 주로 사용되는 협업 필터링 방법을 사용한다. 협업 필터링 방법에서 주로 나타나는 문제점인 희박성 문제를 해결하기 위하여 개인의 성향을 이용하는 방법 중 인구통계학적 데이터에서 개인의 성향을 파악하고 이를 협업필터링의 최근접 이웃을 구성하는 입력 데이터로 사용한다[2].

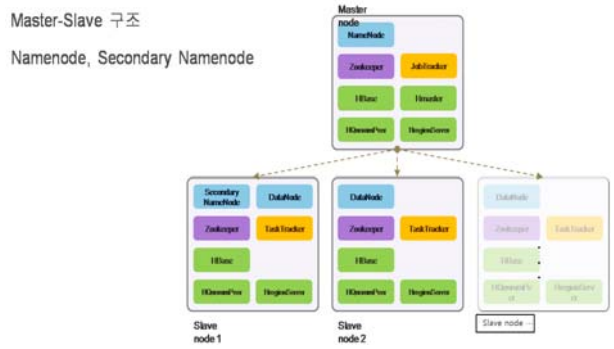
본 논문에서는 데이터 처리에서 가장 주목받고 있는 병렬처리 시스템인 하둡 시스템에 구현한다.

2. 웹툰 추천 시스템의 구성

하둡을 실행한다는 것은 네트워크상의 서로 다른 서버에서 여러 개의 상주 프로그램들을 실행한다는 것을 뜻한다. 하둡에서 상주 프로그램들은 특별한 역할들을 가지고 있는데, 가장 중요한 상주 프로그램은 NameNode, DataNode, JobTraker, TaskTraker이다.

하둡은 분산 저장과 분산연산에 대해 master/slave 구조를 가지고 있다. 이러한 분산 저장 시스템을 하둡 파일 시스템(HDFS)이라고 한다. (그림 3)은 HDFS(Hadoop

Distributed File System)의 전체적인 구조이다.

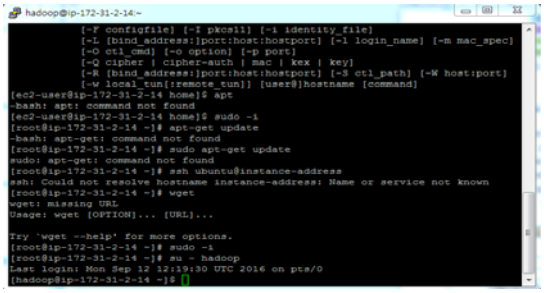


(그림 3) 하둡 파일 시스템(HDFS)의 master/slave 구조

하둡에서 데이터마이닝 알고리즘을 맵 리듀스로 작동하도록 하기 위해서 오픈소스인 머하웃을 많이 사용한다. 본 논문에서는 클라우드 컴퓨팅 시스템인 AWS(Amazon Web Service)의 EMR(Elastic MapReduce) 서비스와 오픈소스인 머하웃을 이용하여 구현하였다. (표 1)은 하둡 시스템의 구축 환경이고, (그림 4)는 AWS(Amazon Web Service) EC2에서 하둡을 설치하여 로그인한 화면이다.

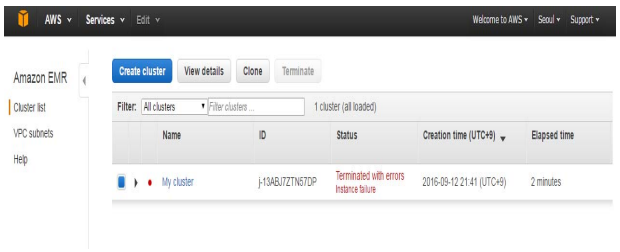
[표 1] 하둡 시스템 구성 환경

시스템 구성요소	시스템 세부 내용
운영체제	AWS EC2 Linux Instance
Java	1.8
Hadoop	1.0.3
MapReduce	Amazon EMR



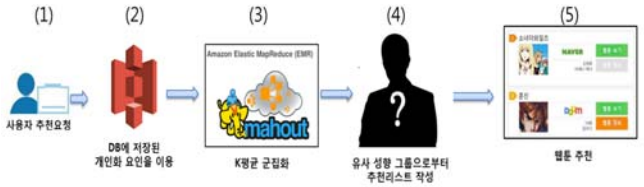
(그림 4) 하둡 시스템 로그인 화면

하둡상에서 데이터마이닝 알고리즘을 맵 리듀스로 작동하도록 제공하는 AWS(Amazon Web Service)의 기능인 AWS EMR(Elastic MapReduce)을 사용하여 좀 더 편리하게 사용할 수 있다. (그림 5)는 AWS EMR(Elastic MapReduce)의 콘솔 환경 설정 화면이다.



(그림 5) AWS EMR 콘솔 환경 설정 화면

하둡에서 개인화 웹툰 추천 시스템의 구성도는 (그림 6)과 같으며 데이터 마이닝 연산은 하둡에서 병렬로 처리되는 데이터마이닝 오픈소스인 머하웃을 이용하여 연산을 처리한다[2].



(그림 6) 웹툰 추천 시스템의 구성도

사용자가 웹툰 추천 요청을 하였을 때, 개인화 요인들을 최적의 데이터로 K-평균 군집화를 사용하여 최근접 이웃을 구성하고, 추천 웹툰 리스트를 작성하여 사용자가 원하는 웹툰을 추천한다. (그림 6)을 좀 더 설명하면 다음과 같다.

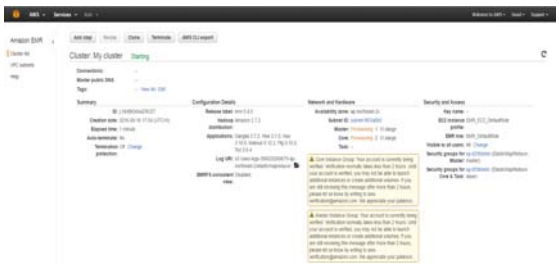
(1) 협업필터링을 사용하기 위한 데이터를 회원가입시 사용자로부터 개인화 요인을 입력 받고, AWS 클라우드 서버 데이터베이스에 저장한다. (2) 사용자가 추천요청을 하게 되면, DB에 저장된 개인화 요인을 적용하여 협업필터링 방법을 적용한다, (3) 개인화 요인들을 데이터로 하여 K-평균 군집화를 사용하여 최근접 이웃을 구성하고, 유사한 사용자들을 추려내어 유사 성향 그룹을 만들어낸다. (4) 유사 성향 그룹으로부터 웹툰 추천 리스트를 작성하여, (5)와 같이 웹툰을 추천하게 된다.

3. 웹툰 추천 시스템의 구현

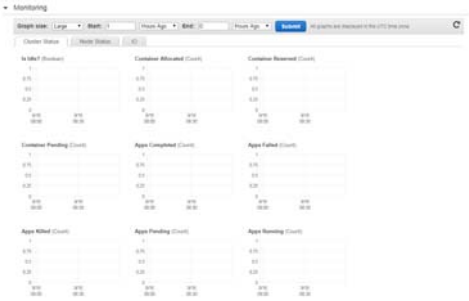
하둡을 이용하여 추천시스템을 구현하기 위해서는 하둡 구축 환경을 구성하여야 한다. 하둡은 윈도우와 Linux 모두 설치가 가능하지만, 윈도우는 호환이 좋지 않기 때문에 본 논문에서는 Linux 환경으로 구성하였다.

Linux 환경을 구축하기 위해서는 직접 Linux 운영체제를 설치하여 구성하는 방법과 클라우드 컴퓨팅 시스템을 이용하여 서버를 구축할 수 있다. 본 논문에서는 최근 개발자들이 많이 이용하고 있는 AWS(Amazon Web Service) EC2 인스턴스를 사용하여 개발 환경을 구축하였고, 연계되는 AWS EMR(Elastic MapReduce) 기능과 오픈소스인 머하웃을 이용하여 본 프로그램을 구현하였다.

(그림 7)은 AWS EMR 클러스터를 생성하여 환경을 구축한 콘솔 환경 설정 화면이고, (그림 8)은 클러스터가 만들어져 3개의 인스턴스를 추가적으로 할당받아 모니터링 할 수 있는 화면이다.



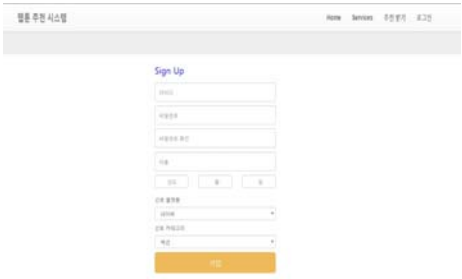
(그림 7) AWS EMR 클러스터 구축 환경



(그림 8) AWS EMR 데이터 모니터링 환경

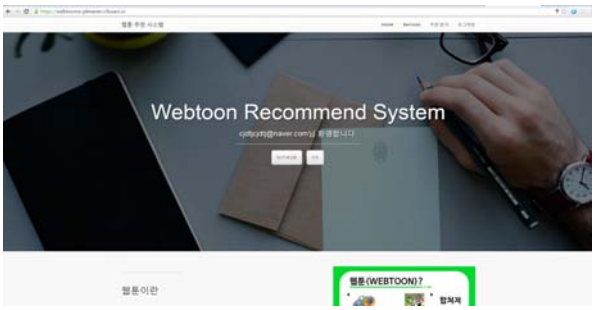
AWS EC2 인스턴스는 Linux 콘솔 환경으로 구성되어 있어, 클라우드 환경에 가장 빠르게 접목시킬 수 있는 Ruby on Rails 프레임워크를 사용하여 개발하였다.

웹툰 추천 시스템을 이용하기 위해서는 사용자의 회원가입을 필수로 하고, 회원가입에서 입력 받은 데이터를 저장하여 협업필터링 개인화요인으로 사용한다. 이 때 생년월일, 성별, 선호 카테고리, 선호 플랫폼을 필수적으로 입력받아 회원가입이 이루어진다. 회원가입 양식은 (그림 9)와 같다.



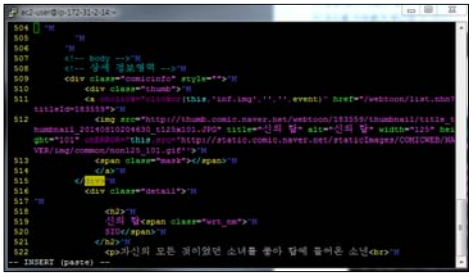
(그림 9) 웹툰 추천 시스템 회원가입 양식

하둡 환경으로 구성된 AWS 클라우드 서버에 사용자들이 회원가입을 통해서 얻은 정보를 저장하게 된다. 데이터베이스에 저장되는 순서는 ID, 이름, 성별, 생년월일, 선호 웹툰 플랫폼, 선호 카테고리 순이다. 회원가입 후 웹툰 추천 시스템의 로그인한 홈 화면은 (그림 10)과 같다.



(그림 10) 사용자가 로그인한 홈 화면

사용자들에게 웹툰을 추천하기 위해서는 웹툰 데이터베이스가 필요하다. 본 논문에서는 Linux 환경인 AWS 클라우드 서버에서 wget 명령어를 사용하여 웹페이지의 정보를 가져왔다. (그림 11)은 웹툰 웹페이지에서 가져온 정보를 나타낸 것이고, (그림 12)은 (그림 11)에서 추출한 웹툰 데이터를 바탕으로 데이터베이스를 생성하였다[3].



(그림 11) 웹페이지에서 가져온 웹툰 정보

```
 Webtn Load (0.2ms)  SELECT "webtns".* FROM "webtns"
 => #<ActiveRecord::Relation [#<Webtn id: 1, name: "프리드로우", writer: "전선욱", plat_id: 1, cate_id: 7>, #<Webtn id: 2, name: "은밀하게 위대하게", writer: "최종훈", plat_id: 2, cate_id: 3>, #<Webtn id: 3, name: "흑색논란", writer: "최보람,손규호", plat_id: 5, cate_id: 7>, #<Webtn id: 4, name: "최강왕따", writer: "노도한", plat_id: 3, cate_id: 4>, #<Webtn id: 5, name: "남친 있음", writer: "백혜경", plat_id: 6, cate_id: 9>]>
```

(그림 12) 웹툰 정보를 바탕으로 추출한 데이터베이스

웹툰 정보를 바탕으로 추출한 데이터베이스를 하둡에서 AWS EMR과 함께 사용하려면, AWS S3 데이터 스토리지에 저장하여야 한다. 저장된 AWS S3 데이터 베이스는 (그림 13)과 같다.

(그림 13) AWS S3 스토리지 데이터베이스

사용자 정보와 웹툰의 데이터베이스가 준비되었으니, 이제 본 논문에서 사용자 생년월일(20%), 성별(20%), 선호플랫폼(20%), 선호 카테고리(40%)에 점수를 부여하고, 그것을 협업 필터링 방법에 적용하여 웹툰을 추천하게 된다. (그림 14)는 개인화 요인을 사용하지 않은 기존의 웹툰 추천 결과를 통한 3가지의 웹툰이며, 가중치를 부여한 개인화 요인을 바탕으로한 추천 결과는 (그림 15)와 같다.

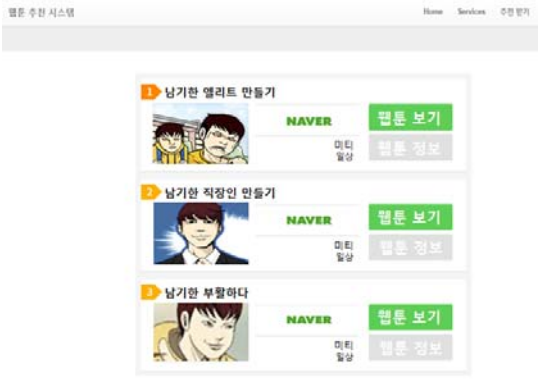
개인화 요인을 사용하지 않고 추천된 웹툰					
user_id	성명	성별	플랫폼	추천웹툰	작가
kunhoo94111	이건호	남	네이버	신의 탑	SIU
sjunhoo	이준호	남	레진코믹스	마도사의 탑	성상영
kimso93	김소연	여	다음	트레이스	네스티캣
dkssud13	이도연	남	카카오페이지	아도니스	팀 아도니스
minji1133	하민지	여	탑툰	총수	정기영

(그림 14) 개인화 요인을 사용하지 않고 추천된 웹툰

개인화 요인에 가중치를 부여해 추천된 웹툰					
user_id	성명	성별	플랫폼	선호 카테고리	추천 웹툰
bat123456	박주현	남	네이버	일상	남기한엘리트만들기
ygunghe1	윤경희	여	네이버	일상	컨트롤제트
cshoo9403	최송희	여	네이버	일상	남기한직장인만들기
tlaeotn123	심대수	남	네이버	일상	마음의소리
wkdxogns13	장태훈	남	다음	일상	궁상가족

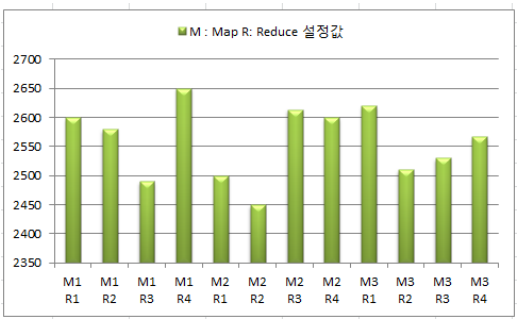
(그림 15) 하둡에서 개인화 요인에 가중치를 부여해 추천된 웹툰

(그림 14)와 (그림 15)를 비교하였을 때, 개인화 요인을 사용하지 않고 추천된 웹툰과 개인화 요인에 가중치를 부여하여 추천된 웹툰이 가중치 유무에 따라 다르게 결과가 나오는 것을 볼 수 있다. (그림 14)에서는 플랫폼과 선호 카테고리에 관계없이 웹툰이 추천되었고, (그림 15)에서는 선호 플랫폼과 선호 카테고리에 부여한 가중치에 따라, '일상' 카테고리와 '네이버' 플랫폼 위주로 추천되었다. 이와 같이 웹툰 추천에서는 사용자들이 선호하는 선호 플랫폼과 선호 카테고리의 기준을 바탕으로 추천의 결과가 완전히 다르게 나올 수 있다. (그림 16)은 "일상" 카테고리와 "네이버" 선호 플랫폼에 따른 결과 화면이다.



(그림 16) 웹툰 추천 결과 화면

성능평가로는 하둡 환경 AWS EMR 설정에서 맵과 리듀스에 사용되는 CPU의 코어수를 변경하는 것으로 본 시스템의 최대 개수인 4개를 기준으로 평가하였다. (그림 17)은 맵과 리듀스에서 사용되는 코어 수 변경에 따른 시간을 10회 측정하여 그 평균치들을 그래프로 나타낸 그래프이다.



(그림 17) 맵 리듀스 코어 수 변경에 따른 시간 측정

성능평가를 위해 약 200여가지의 임의의 웹툰 데이터를 이용하여 작업 노드가 1대인 경우부터 12대인 경우까지 각각 10회씩 WordCount 수행하여 수행 시간의 평균 값 및 표준편차에 대해서 평가하였다.

(그림 17)의 그래프를 보면 맵과 리듀스의 코어 수에 따라 시간이 다르게 측정되는 것을 볼 수 있다. AWS EMR을 통하여 노드가 추가됨에 따른 MapReduce의 응용 수행의 응답 시간을 보인 것이다. MapReduce는 맵(Map) 단계와 리듀스(Reduce) 단계로 처리 과정을 나누어 작업하게 되는데, Map은 흩어져 있는 데이터를 Key, Value의 형태로 연관성 있는 데이터 분류로 묶는 작업이고, Reduce는 Map화한 작업 중 중복 데이터를 제거하고 원하는 데이터를 추출하는 작업이다. (그림 17)의 그래프에서 Y축은 이러한 MapReduce 노드를 추가함에 따른 시간 측청을 나타낸 수치이고, X축은 본 시스템의 최대 개수인 4개까지 순차적으로 성능을 테스트한 결과이다. 결과적으로 Map2와 Reduce2 노드의 추가 응답 시간이 가장 빠른 것을 볼 수 있었다. 하지만 이것에 대해 일반적인 식을 이끌어내지는 못하였다. 그것은 테스트한 데이터가 하둡에 잘 맞지 않은 경우이거나 하둡에서 맵과 리듀스의 코어수가 훨씬 큰 경우에 결과가 잘 나타날 수 있다.

4. 결론

본 논문에서는 현재 연재되고 있는 많은 웹툰 중 사용자와 가장 잘 맞는 웹툰을 추천해주는 시스템을 구현하였다. 기존에 있던 웹툰 추천을 해주는 시스템과는 다르게 개인화 요인을 이용한 협업필터링 방법을 사용하였고, 이 시스템을 하둡 환경에서 구현 및 성능 평가를 진행한 것이 기존 시스템과의 차별성이다. 성능평가를 통해 일반적인 식을 구하지는 못하였는데 이것은 테스트를 진행한 데이터가 하둡에 잘 맞지 않는 경우이거나 하둡에서 맵과 리듀스의 코어수가 훨씬 큰 경우에 결과가 잘 나타날 수 있다. 향후에는 이러한 면을 고려해서 연구가 진행되어야 할 것이다.

참고문헌

[1] 고정민, 양지훈, 고창만, 박지혜, 백경지, 만화 유통 환경 개선 방안 - 웹툰 산업을 중심으로, 한국 콘텐츠 진흥원 연구 보고서, pp.1-154, 2016. 9

[2] 김선호, 김세준, 모하영, 김채린, 박규태, 박두순,"하둡에서 개인 성향을 이용한 영화 추천 시스템," 한국정보처리학회 춘계학술발표대회 논문집, 제21권 제1호, 아주대학교, pp. 642-644, 2014. 4

[3] 이건호, 박두순,"클라우드 컴퓨팅 시스템에서 구현한 웹툰 추천 시스템," 한국정보처리학회 춘계학술발표대회 논문집, 동국대학교, pp. 451-454, 2016. 4