Movielens 데이터를 이용한 영화 추천 시스템 개발

박윤기*, 정현철*, 홍지원*, 김상욱*
*한양대학교 컴퓨터공학부
e-mail : {yunkipark, skyjhcblue}@hanyang.ac.kr, {nowiz, wook}@agape.hanyang.ac.kr

A Movie Recommendation System Using Movielens Data

Yoon-Ki Park*, Hyun-Cheol Jung*, Ji-Won Hong*, Sang-Wook Kim*
*Department of Computer Science and Engineering, Hanyang University

요 약

본 논문에서는 영화를 즐기는 이용자들에게 영화를 추천하는 추천 시스템 개발에 대해 논한다. 본 논문에서는 유저 기반 협업 필터링 기술을 적용한 영화 추천 웹 서비스를 개발하였고, 웹 크롤링 기술을 이용하여 추천된 영화의 상세한 정보를 제공할 수 있도록 하였다. 유저 스터디를 수행한 결과 이 영화 추천 시스템을 이용한 사용자들의 만족도는 대체로 높았다.

1. 서론

주 5 일 근무제가 보편화 되고 최근 여가시간에 대한 중요성이 대두되면서 여가시간을 알차게 보내려는 사람들이 많아졌다. 이러한 여가시간에 사람들이 가장 많이 하는 것 중 하나가 바로 영화관람이다. 세계적으로 수많은 영화가 시장에 쏟아져 나오고 있고 사람들은 자신이 좋아할 만한 영화를 찾기 위해 홈페이지를 방문하고 블로그를 보거나 잡지를 구독하는 등 영화를 선택하는 데에 많은 시간과 노력을 들이고 있다. 본 논문에서는 이러한 문제점을 해결하고 사람들이 더욱 다양한 영화를 접할 수 있도록 하기 위해 타깃유저(추천을 제공받을 유저)와 많은 다른 유저들이 영화마다 입력한 평가점수 정보를 이용하여 타깃유저가 아직 보지 않은 영화들의 평가점수를 예측하고 그 중에서 높은 평가점수를 받은 영화를 추천하는 시스템을 제작하게 되었다.

2. 데이터 정보

데이터집합은 미네소타대학의 GroupLens Research Project[1]에서 수집된 MovieLens Data Set 로 943 명의 사용자들이, 18 개 장르, 1682 개 영화에 대해 100,000 건의 평점을 매긴 자료를 바탕으로 한다. 각 영화는 미국 영화데이터베이스인 IMDB (Internet Movie Database, http://www.us.imdb.com)의 장르 기준에 따른 것으로 최소 1 개 장르에서 최대 5 개 장르까지 속한다.

3. 핵심 기술

본 구현에 사용한 핵심 기술로는 유저기반 협업 필터링 알고리즘과 웹 크롤링이 있다.

협업 필터링(Collaboration Filtering)[2, 3]이란 타깃유저가 보지 않은 영화에 줄 평가점수를 다른 유저들이 매긴 평가점수의 데이터베이스를 이용하여 예측하는 것이다. 유저기반 협업 필터링은 메모리기반 협업 필터링으로서 유저간의 유사도를 이용하여 weighted sum 을 수행한다. 이렇게 다른 유저들의 해당 영화에 대한 선호도를 반영하여 예상점수에 영향을 주게 된다. 타깃유저 a아이템 j에 줄 예상평가점수를 $p_{a,j}$라 하면 다음과 같이 다른 유저들의 평가점수의 weighted sum 으로 구할 수 있다.

$$p_{a,j} = \bar{v}_a + \kappa \sum_{i=1}^{n} w(a,i)(v_{i,j} - \bar{v}_i)$$

여기에서 $\bar{v}_a$는 타깃유저 a가 매긴 평가점수들의 평균 점수를 말하고, $w(a,i)$는 타깃유저 a와 유저 i의 유사도를 의미하며 가중치로 사용된다. 또한, $v_{i,j}$는 유저 i가 아이템 j에 매긴 점수이다. κ는 normalizing factor 로서 각 유사도의 절대값의 합의 역수를 대입한다. n은 0 이 아닌 유사도를 가진 데이터베이스의 모든 유저를 의미한다.

유저 a와 i의 유사도는 다음과 같이 구할 수 있다.

$$\mathrm{w(a,i)} = \frac{\sum_j (v_{a,j} - \bar{v}_a)(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_j (v_{a,j} - \bar{v}_a)^2 \sum_j (v_{i,j} - \bar{v}_i)^2}}$$

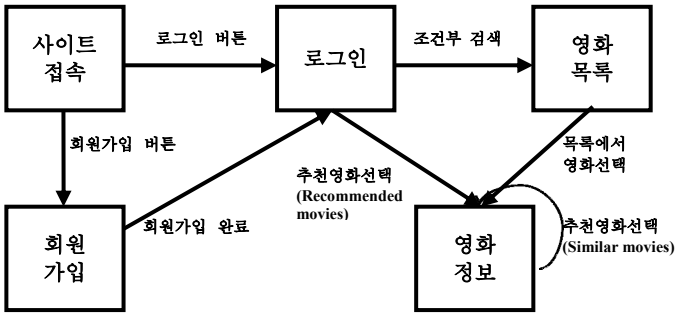
여기에서 j는 유저 a와 i가 공통으로 점수를 준 아이템을 의미한다.

웹 크롤링은 자동으로 웹 문서에서 원하는 정보를 수집하는 작업이다. 웹 크롤링의 기본 원리는 최초 시작 URL 주소리스트를 시작점으로 해서 HTTP 프로토콜을 이용해 웹 문서를 요청하고 그 문서의 내용에서 링크 URL 을 추출하여 다시 추가적인 문서를 수집해 나가는 것이다. 이 과정을 계속 반복하여 시작 URL 의 모든 페이지 소스를 가져온다.

물론 실제 사용되는 크롤링에서는 효율성과 사이클링 방지를 위해 중복 문서를 제거하거나 웹 호스팅 서버에 과부하가 걸리는 것을 막기 위해 수집 속도를 조절하는 등의 여러 가지 정책이 사용되지만 본 프로그램에서는 웹 페이지의 정보를 추출하기 위해 해당 페이지의 모든 정보를 가져온다. 페이지에서 원하는 정보를 추출하기 위해서 정규식(Regular Expression)을 사용한다. 정규식은 일정한 패턴을 인식하는 데에 아주 유용하게 사용되는 표현이다. 웹 소스에서 원하는 정보(포스터, 감독, 배우, 작가 등)를 얻는데 사용하였다.

4. 구현

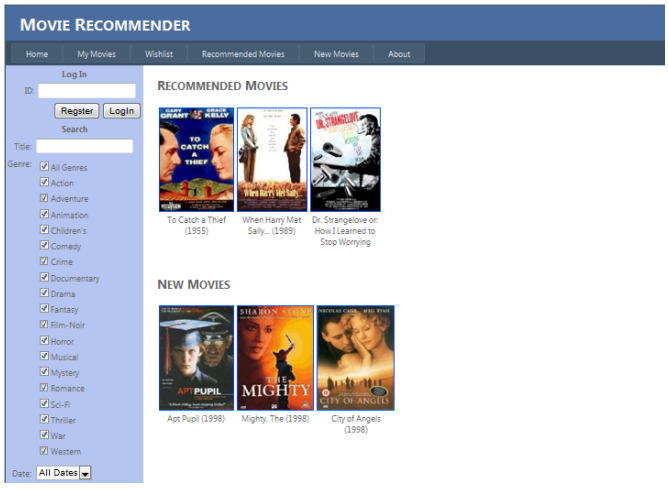
시스템 제작을 위해 C#언어를 기반으로 한 ASP.NET 을 사용하였다. ASP.NET 은 웹 폼 기반의 개발환경으로 개발할 때 직관적이고 디자인과 코드의 분리를 통해 코드의 관리가 용이한 장점이 있다. 데이터베이스는 Microsoft 사에서 제공하는 MSSQL 을 사용하여 구성하였다.



(그림 1) 프로그램 구성

4.1. 최근 개봉 영화

처음 프로그램이 실행 되었을 때 영화 정보가 저장된 데이터베이스 테이블을 영화의 개봉일 순으로 정렬하여 가장 최근 개봉한 영화 Top-3 를 뽑아 포스터와 함께 보여준다. 또한 로그인 후에도 추천영화 Top-3 와 함께 보여준다.



(그림 2) 로그인 후 화면

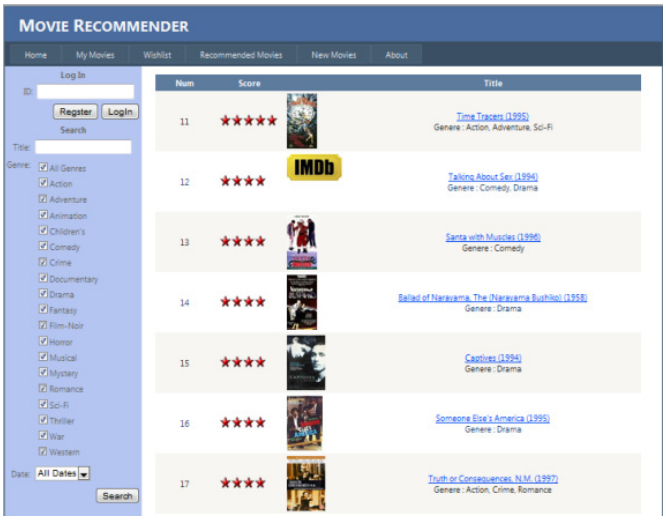
4.2. 로그인 & 추천 영화

유저 로그인시 타깃유저의 평가점수 기록과 다른 유저들의 평가점수 기록을 사용하여 유저기반 협업 필터링으로 타깃유저가 점수를 주지 않은 영화들의 점수를 예측한다. 점수가 높은 순으로 영화를 정렬하여 로그인 후 화면에는 로그인 전에 보여주었던 최신 영화 Top-3 와 함께 예측점수가 Top-3 인 영화를 보여준다. (그림 2)

4.3. 조건부 검색

기본적으로 영화 제목을 이용하여 해당 문자열을 포함한 영화를 검색할 수 있다. 검색된 영화의 점수를 각각 예측하여 점수의 내림차순으로 정렬되게 하였다.

또한 검색된 모든 영화 중 사용자가 선택한 장르와 선택한 연도를 모두 만족하는 영화를 필터링 하여 조건부 검색을 수행할 수 있다. 장르는 한번에 여러 가지를 동시에 선택할 수 있도록 체크박스로 처리하였고 연도는 콤보박스로 만들어 한번에 10 년 단위로 선택할 수 있도록 하였다.



(그림 3) 조건부 검색

4.4. 회원가입 창

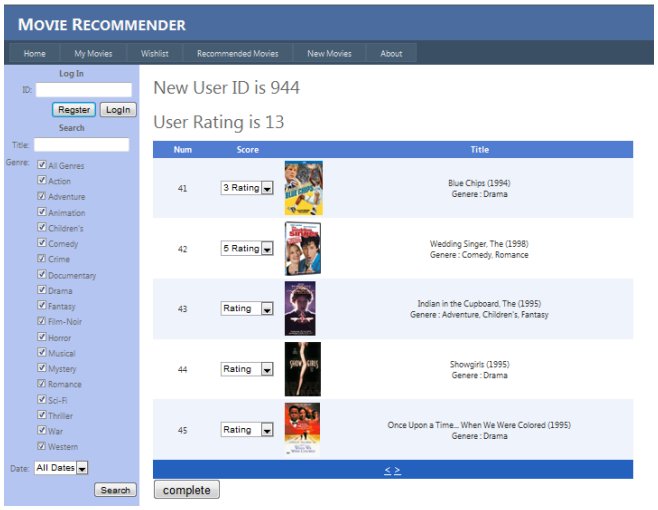
회원 가입버튼을 클릭했을 때 유저아이디를 자동으로 할당하여 회원으로 등록하고 전체영화목록을 무작위로 리스트에 나타낸다. 이 리스트에서 회원가입하는 유저는 자신이 본 영화를 찾아 점수를 줄 수 있다. 추천 서비스를 제공하기 위해서는 타깃유저가 어느 정도의 평가점수를 준 상태이어야 하기 때문에 기본적으로 회원가입을 했을 때에는 최소 15 개의 영화에 점수를 주도록 하였다. (그림 4)

4.5. 영화 정보 창

영화를 클릭한 후 데이터베이스에 저장된 영화 URL 을 통해 웹 크롤링으로 해당 페이지 소스를 가져온다. 정규식을 이용하여 원하는 정보(포스터, 감독, 작가, 배우, 스토리)를 추출하여 영화 정보 창을 완성

하게 된다.

선택한 영화를 기준으로 아이템기반 유사도[4]를 계산하여 가장 비슷한 성격의 영화 Top-3 를 보여준다. 또한 이 아이템기반 유사도와 미리 계산해 놓은 예측 평가점수를 곱하여 선택한 영화와 관련된 영화 중에서 타깃유저가 선호할 만한 영화를 추천한다. (그림 5)



(그림 4) 회원가입 창

(그림 5) 영화 정보 창

5. 유저 스터디

프로그램을 사용하는 실제 사용자들의 만족도를 평가하기 위해서 유저 스터디를 진행하였다. 유저 스터디 참가자들은 회원가입을 하여 자신의 아이디를 만들고 회원가입 후에 자신이 본 영화 15 편을 찾아 점수를 입력하였다. 점수 입력 후 Recommended Movie 를 선택하여 참가자들이 자신에게 추천된 영화를 확인하도록 하였다. 다음 <표 1>은 추천 리스트에서 Top-10 영화에 대해 참가자들의 만족도를 조사한 결과이다. 총 50 명의 참가자에 대해 0 점에서 5 점 사이 1 점 단위로 점수를 줄 수 있도록 하여 만족도를 조사하였다.

만족도	사람 수
★★★★★ (5점)	11
★★★★☆ (4점)	23
★★★☆☆ (3점)	9
★★☆☆☆ (2점)	6
★☆☆☆☆ (1점)	1
☆☆☆☆☆ (0점)	0
평 균	3.74 / 5.0

<표 1> 사용자 만족도

조사를 마친 결과 만족도 점수의 평균은 3.74 점이었다. 참가자들은 추천 결과를 대체로 만족스럽게 생각하였다. 오래된 영화가 많다는 점에서는 낮은 점수를 주었던 사람들이 있었다. 영화를 좋아하는 성향이 한쪽에 편중되어 있을 수록 비슷한 선호도를 가진 유저들에게 영향을 많이 받아 추천의 정확도가 높게 나오기 때문에 선호하는 장르가 뚜렷하지 않은 유저들도 비교적 추천의 만족도가 낮았다.

6. 결론

본 논문에서는 Movielens 데이터를 이용한 영화 추천 시스템 개발에 대해 다루었다. 이를 위해 데이터의 수집하고 웹 서비스를 구축하였으며, 협업 필터링을 사용한 영화 추천 시스템의 사용자 평가를 수행하여 결과가 좋음을 보였다.

감사의 글

본 연구는 지식경제부 및 정보통신산업진흥원의 IT 융합 고급인력과정 지원사업의 연구결과로 수행되었음(NIPA-2011-C6150-1101-0001).

참고문헌

[1] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl, “1994. GroupLens: an open architecture for collaborative filtering of netnews,” In Proc. of ACM CSCW, 1994.

[2] Xiaoyuan Su and Taghi M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques,” Advances in Artificial Intelligence, 2009.

[3] John S. Breese, David Heckerman, and Carl Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering,” Microsoft Research, 1998.

[4] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, “Item-Based Collaborative Filtering Recommendation Algorithms,” Proc. of Int’l Conf. on WWW, 2001.