사용자 성향 기반 적응형 추천시스템

김태환, 이승화, 오제환, 이은석
성균관대 컴퓨터공학과
e-mail : {th_kim, shlee, hide7674,eslee}ece.skku.ac.kr

An Adaptive Recommendation System based on User Propensity

Taehwan Kim, Seunghwa Lee, Jehwan Oh, and Eunseok lee
Dept of computer Engineering, Sungkyunkwan University

요 약

웹 상에 정보가 폭발적으로 증가함에 따라 각 사용자에게 맞는 정보를 선별하여 제공하는 개인화 서비스는 매우 중요한 이슈가 되었다. 기존 추천시스템들은 컨텐츠 기반 필터링과 협업 필터링 기법을 기반으로 한다. 그러나 이러한 방법들은 충분히 수집된 사용자 정보를 필요로 하기 때문에, 적절한 추천이 이루어지기 까지 다소 시간이 소요되는 문제를 가지고 있다. 또한 사용자의 성향이 지나치게 편중되는 경우, 사용자의 취향변화를 반영하여 새로운 상품을 추천하는 것은 어렵다. 실제로 사용자들은 웹 사이트의 방문 목적에 따라 개인화된 상품추천을 원하기도 하고, 많은 사용자들에게 인기 있는 상품을 원하기도 한다. 본 논문에서는 사용자의 행동분석을 기반으로, 협업 필터링을 기반으로 하는 개인화된 추천과 다수의 사용자들에게 공통적으로 인기 있는 상품의 추천비율을 동적으로 조합하여 최종 추천 상품들을 선별하는 새로운 적응형 추천시스템을 제안한다. 본 논문에서는 MovieLens 의 데이터 셋을 이용하여 기존 추천기법들과 추천결과에 대한 정확도를 비교 실험하였으며, 보다 높은 정확도를 보이는 실험결과를 통해 제안시스템의 유효성을 확인하였다.

1. 서론

인터넷 기술의 발전으로 웹 상에 너무 많은 정보가 흐르고 있다. 그러나 사용자가 실제로 필요로 하는 정보의 양은 극소수이기 때문에, 정보를 필터링하여 제공하는 추천시스템의 연구가 중요해지고 있다. 이는 특히 전자상거래 분야에서 사용자의 만족도를 향상시키기 위해 필수적으로 요구되고 있다.

전통적인 추천기법으로 컨텐츠 기반 추천과 협업추천방법이 있다. 컨텐츠 기반 추천기법은 사용자의 관심과 대상 컨텐츠의 유사도를 비교하여 적절한 컨텐츠를 선별하는 기법이다. 이는 추천대상이 문서인 경우 효과적으로 사용이 가능하지만, 멀티미디어 정보나 해당 아이템에 표현되어 있지 않은 비문자 정보는 무시된다는 단점이 있다.

이러한 약점을 보완하기 위해, 사용자의 선호도와 컨텐츠를 비교하는 대신에 여러 사용자들의 선호도를 비교하여, 유사한 사용자 간에 상호 추천을 하는 협업추천방법이 제안되었다. 이는 인간이 해당 컨텐츠를 보고 판단한 정보를 기반으로 하기 때문에, 비문자 정보로 이루어진 컨텐츠도 추천이 가능하다는 장점이 있다. 그러나 그룹 내에 사용자들이 평가하지 않은 새로운 아이템에 대해서는 추천이 이루어지기 어렵다는 문제가 있다.

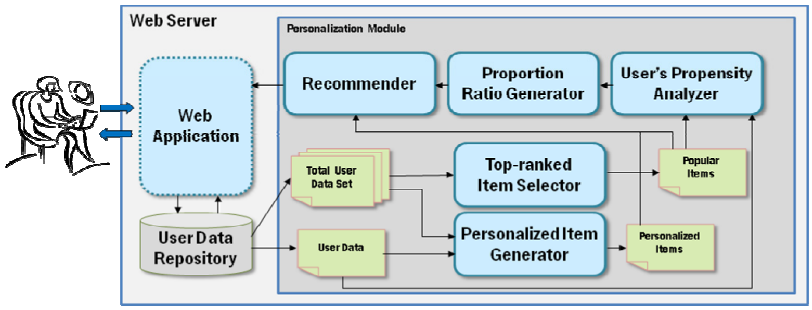
이러한 기존의 추천기법들은 충분히 수집된 사용자의 정보를 필요로 하기 때문에, 적절한 추천이 이루어지기까지 다소 시간이 소요되는 문제를 가지고 있다. 또한 사용자의 성향이 지나치게 편중되어 개인화 서비스가 이루어질 경우, 사용자의 취향변화를 반영하여 새로운 상품을 추천하는 것은 어렵다.

본 논문에서는 이와 같은 기존 추천시스템의 문제를 해결하기 위해, 사용자의 행동분석을 기반으로, 개인화된 추천 상품과 다수의 사용자들이 공통적으로 선호하는 상품의 추천비율을 동적으로 조합하여 최종 추천 상품들을 선별하는 적응형 추천시스템을 제안한다. 이를 통해 기존 추천기법의 약점이었던 Cold-start 문제를 개선하고, 사용자취향이 지나치게 편중되어 새로운 상품추천이 어려웠던 문제를 해결할 수 있다.

본 논문에서는 제안시스템의 평가를 위해 영화추천 사이트인 MovieLens 의 데이터 셋을 이용하여 기존 추천기법들과 추천결과에 대한 정확도 측면의 비교실험을 수행하였다. 다양한 상황에서의 정확도를 비교하기 위해, 사용자의 행동데이터 양을 변화시키며 실험을 진행하였으며, 기존 기법들에 비해 높은 정확도를 나타내는 실험결과를 통해 제안시스템의 유효성을 확인하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련연구를 소개하고, 3 장에서는 제안시스템의 전체적인 구조와 동작과정을 설명한다. 4 장에서는 시스템의 평가를 위한 실험과 그 결과를 소개하며, 5 장에서는 결론과 향후 과제를 논의한다.

본 연구는 지식경제부의 유비쿼터스컴퓨팅 및 네트워크원천기반기술개발사업, 교육과학기술부의 특정기초연구사업 R01-2006-000-10954-0 의 연구결과로 수행되었음



(그림 1) 시스템의 전체적인 구성

2. 관련연구

협업 필터링은 현재 업계에서 가장 많이 쓰이고 있는 방법으로 사용자 행동 데이터가 많을수록 더 정확한 추천이 이루어진다. 반면에 사용자 행동 데이터가 부족할 때 추천의 정확도가 떨어지는 문제가 있는데 이를 Cold-start 문제라고 한다. 이러한 문제를 해결하기 위해 사용자의 행동 데이터와 배타적인 항목의 성질에 관한 데이터 등을 동시에 이용하는 하이브리드 방법이 일반적으로 시도되고 있다. 예를 들어서, 사용자가 선호하는 영화의 장르를 이용해서 추천을 생성하는 방법과 함께 협업 필터링을 사용하는 등의 방식이 있다. 이런 방법은 Cold-start 문제를 개선하고 더 높은 정확도를 가지는 추천을 하는 경향을 보인다. 그러나 이는 사용자 또는 항목에 관해 더 많은 데이터를 수집하여 보유해야 하고, 두 가지 이상의 추천 알고리즘을 활용하기 때문에 시스템의 복잡도가 커지고 속도가 느려지는 단점이 있다.

또 다른 협업 필터링의 문제로 신규 사용자 문제가 있다. 추천 시스템이 설치가 되고 시간이 지날수록 시스템이 저장하고 있는 사용자 행동 데이터는 많아지고 그에 따라 협업 필터링은 더 정확한 추천을 하게 된다. 하지만 추천시스템을 처음 이용하는 사용자의 경우 그 사용자가 이전에 했던 행동에 대한 데이터가 전혀 없거나 적은 상태이기 때문에 협업 필터링은 부정확한 추천을 하게 되는 문제이다. 신규 사용자 문제를 해결하기 위해 명시적으로 사용자 취향이나 선호도를 판단해서 추천을 한다거나, 사용자 이전 행동 데이터와 관련이 없는 컨텐츠에 기반한 자료를 이용해서 추천하는 방법들이 시도되고 있지만 개인화 된 추천을 하지 못하고, 또한 저장해야 되는 데이터의 양이 많아지고 시스템이 복잡해지는 특성이 있다.

위의 방법들 외에도 다양한 방법들이 협업 필터링의 문제를 해결하고 성능을 향상시키기 위해서 도입되어 왔다. 가장 일반적으로 사용되는 방법이 하이브리드 추천 방식으로 2 가지 이상의 추천 방법을 이용해서 추천을 생성하는 방식이다. 하이브리드 방식은 각 각의 추천방법의 단점을 완화하고 추천정확도가 향상되는 경향이 있지만, 시스템의 복잡도가 커지고 추천을 생성하는 시간이 오래 걸리는 문제를 피할 수가 없다.

이처럼, 협업 필터링의 문제를 개선하고 성능을 향상시키는 대부분의 추천방법들은 공통적으로 시스템의 복잡도가 높아지고 추천을 생성하는 시간이 길어지는 단점이 있다.

3. 제안시스템

제안시스템은 사용자 행동데이터를 기반으로 계산된 사용자 성향을 이용해서 동적으로 추천전략의 비율을 조정하고 보다 균형이 있는 추천정보를 생성하는 것을 목표로 한다. 이와 동시에 시스템이 저장해야 되는 데이터의 양을 유지하고, 추천결정을 위해 소요되는 시간을 길게 하지 않으면서 협업필터링의 Cold-start 문제와 신규 사용자 문제를 해결한다.

3.1 제안시스템 구성

제안시스템은 추천시스템 서버에 위치하며, 사용자의 행동정보를 기반으로 성향을 분석하고, 추천전략의 비율을 동적으로 결정한다. 본 시스템의 전체적인 구성은 (그림 1)과 같으며, 각 구성요소의 기능은 다음과 같다.

- User Data Repository: 모든 사용자의 행동 데이터를 저장하고 있는 데이터베이스이다. 저장되는 행동 데이터는 각 사용자가 어떤 영화에 몇 점의 평가를 부여했는지의 데이터가 저장되고 이것을 이용해서 협업 필터링의 추천 목록과 일반적으로 선호되는 항목의 추천목록을 생성한다.
- Top-ranked Item Generator: 일반적으로 선호도가 가장 높은 항목을 추천목록으로 생성한다. 실제로, 일반적으로 선호도가 높은 목록이라는 것은 베스트 셀러를 지칭하는 것으로 전체 항목 중에서 가장 많은 사용자가 이용한 항목을 의미한다.
- Personalized Item Generator: 사용자가 이전 행동 데이터를 이용해 협업 필터링 기법을 적용하여 개인화된 추천을 생성한다.
- User's Propensity Analyzer: 사용자의 성향을 분석한다. 사용자가 개인화 된 추천(Personalized Items)과 대중적인 선호도에 기반한 추천(Popular Items) 사이에서 어떤 쪽에 가까운 성향을 나타내는지

표시하는 값으로 사용자 행동 데이터 저장소에 있는 데이터를 이용한다.

- Proportion Ratio Generator: 시스템이 보유하고 있는 사용자 행동 데이터의 양에 따라 사용자의 성향에 따른 최적의 추천 조합 비율을 다르게 나타난다. 예를 들어서 시스템이 보유하고 있는 데이터가 적다면 사용자의 데이터가 이 사용자는 협업 필터링의 결과를 더 선호하는 사용자라고 표시한다고 하더라도 일반적으로 선호되는 품목을 추천했을 때 정확도가 더 높은 경향을 보인다. 그러므로 데이터의 양에 따라 최적의 추천 비율을 결정한다.
- Recommender: 결정 된 최적화 된 추천 비율과 사용자의 성향에 따라 사용자에게 협업 필터링의 결과 몇 개와 일반적으로 선호 되는 추천목록 몇 개를 조합하여 최종 추천목록을 생성한다.
- Total User Data Set: 시스템이 저장하고 있는 사용자의 이전 행동 데이터를 의미한다.
- User Data: 시스템에 접속하고 있는 사용자의 이전 행동 데이터를 의미한다.

3.2 제안시스템 동작과정

제안시스템의 동작과정을 사용자에게 영화를 추천하는 상황을 가정하여 설명한다.

사용자는 추천시스템을 이용하는 비디오 대여 사이트에 접속을 한다. 이 때 추천시스템이 활성화가 된다. Personalized Item Generator 는 데이터 저장소에 있는 Total User Data Set 과 User Data Set 을 불러들여 협업 필터링 방법을 이용해 추천목록 20 개(Personalized Items)를 생성한다. Top-ranked Item Generator 는 Total User Data Set 을 이용해 가장 많은 사용자가 이용한 영화목록 20 개(Popular Items)를 생성한다. User's Propensity Analyzer 는 Popular Items 와 User Data 를 불러들여 사용자의 성향을 결정한다. Optimized Proportion Ratio Generator 는 시스템이 보유하고 있는 데이터의 양에 따라서 가장 최적화 된 추천비율을 결정한다. 마지막으로 Recommender 는 이전에 생성 된 Personalized Items 와 Popular Items, 사용자의 성향, 그리고 최적의 추천비율을 이용해서 개인화 된 추천목록 20 개를 생성한다.

하이브리드 된 제안시스템을 구성하고 동작하기 위해서 협업 필터링에서 사용되는 데이터 이 외에 추가적인 데이터가 필요하지 않은 것은 명확하지만 일반선호도 추천목록을 계산하기 위한 부하는 피할 수 없다. 이를 해결하기 위해서 Popular Items 를 생성하는 과정과 최적의 추천비율을 결정하는 과정을 시스템의 사용률이 적은 새벽 시간대에 배치한다고 가정한다

제안 시스템의 핵심기능인 추천전략의 동적인 비율 결정과정은 3.3 절에서 보다 자세히 설명한다.

3.3 추천 비율 결정 알고리즘

우리는 제안시스템을 설계하기 이전에, 협업 필터링을 이용해 개인화 된 추천을 생성하는 전략과 다수의 사용자가 선호하는 항목을 추천하는 전략의 적절한 비율 결정에 대한 알고리즘을 결정하기 위해 다음과 같은 실험을 하였다.

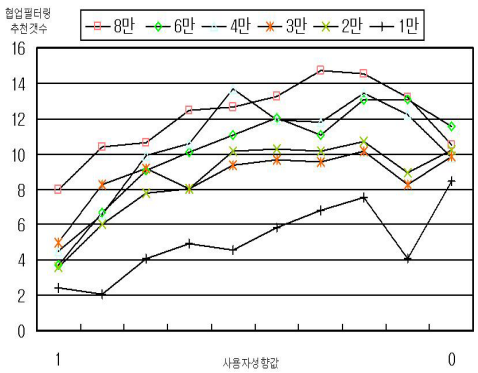
사용자가 개인화 된 추천과 일반적으로 선호되는 추천 중에 어느 정도 개인화 된 추천을 원하는 지를 암시적으로 알아보기 위해 사용자의 이전 행동 데이터를 이용한다.

$$User's\,Propensity = \frac{A \cap B}{A} \qquad (1)$$

User's Propensity = 사용자 성향 값
A = 사용자가 이전에 점수 매긴 항목의 집합
B = 일반적으로 선호되는 항목의 집합

제안되고 있는 사용자의 성향은 이 전에 사용자가 점수 매긴 항목이 일반적으로 선호되는 항목과 교집합이 클수록 큰 값을 가지게 되고 반대의 경우 작은 값을 가지게 된다. 그리고 계산 된 사용자의 성향 값이 작을수록 협업 필터링과 일반적으로 선호되는 항목을 조합해서 추천할 때 협업 필터링의 결과를 더 많이 포함시킨다면 추천이 향상된 정확도를 보일 것이라고 가정하고 검증하였다. 검증은 각 사용자에게 20 개의 두 방법이 조합된 추천을 하고 최적의 추천 조합법을 찾은 다음 사용자 성향과의 상관관계를 그래프로 나타내었다.

<표 1> 다양한 크기의 사용자 이전 행동 데이터에서의 사용자 성향에 따른 최적의 CF 결과의 비율



다양한 크기의 데이터에서 검증을 해본 결과 모든 상황에서 제안하고 있는 사용자의 성향 값에 따른 최적의 협업 필터링 결과의 조합 비율은 가정한 대로의 패턴을 보이므로 타당하다고 할 수 있다. 그리고 사용자의 성향을 결정하기 위해 협업 필터링에서 사용되는 사용자 이전 행동 데이터 외에 다른 데이터가 필요하지 않으므로 시스템이 수집해야 할 데이터의 양은 유지되고 있다.

<표 1>에서 볼 수 있듯이 시스템이 보유하고 있는 데이터의 양이 많아 질수록 하이브리드 된 추천목록에서 협업 필터링의 추천목록의 비율이 높아지므로 시스템이 보유하고 있는 데이터의 양에 따라 추천

전략의 비율을 조절해야 할 필요가 있다. 이를 위해 <표 1>에서 나타난 그래프를 이용해 데이터의 양이 많아 질수록 협업 필터링의 추천목록의 비율이 높아지는 정도를, 사용자 성향에 따라 계산하여 최적화된 추천비율을 결정하였다. 이를 위해, 사용자 성향 값에 따라 기준이 되는 추천의 조합비율을 30,000 데이터 셋에서 추출하고, 데이터 셋의 크기에 따라 협업 필터링의 추천 개수를 더하고 빼는 식의 접근을 하였다

.

$$Ratio\ of\ Recommendation = result(p) \pm \alpha \quad (2)$$

Ratio of Recommendations = 최적의 추천조합비율을 의미함

여기서 *result* 는 30,000 데이터 셋에서 추출된 사용자 성향에 따른 최적의 협업 필터링 조합비율을 저장하고 있는 자료이고 p 는 사용자 성향 값, 그리고 α는 데이터 셋의 크기에 따른 변수를 뜻한다.

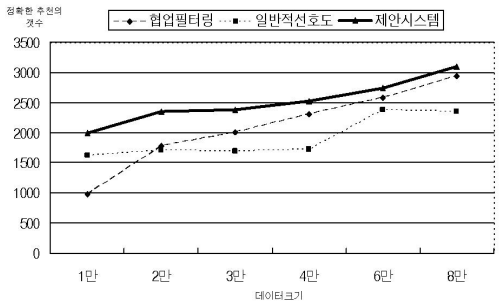
4. 실험 및 평가

우리는 제안시스템의 평가를 위해 미네소타 대학의 GruopLens 연구 프로젝트에서 수집한 무비렌즈 데이터 셋을 이용하여 제안시스템이 개인화되고 균형 잡힌 추천을 생성하면서 협업 필터링의 Cold-start 문제와 신규 사용자문제를 해결하면서 동시에 제안시스템이 추천을 생성하는 시간이 더 길어지지 않고 수집해야 하는 데이터가 더 많아지지 않는지에 대한 실험을 하였다.

본 데이터 셋은 1682 개의 영화에 대해서 943 명의 사용자가 매긴 100,000 개의 점수를 저장하고 있다. 점수는 1-5 점의 범위를 가지고 있고, 각 사용자는 영화들에 최소 20 개의 점수를 매겼다.

시스템에 사용자 행동 데이터가 쌓여 갈수록 협업 필터링, 일반적인 선호도, 제안하는 시스템의 추천 정확도가 어떻게 변하는지 알아보기 위해 10,000, 20,000, 30,000, 40,000, 60,000, 80,000 개의 트레이닝 셋에서 각각 실험 하였고 20,000 개의 테스트 셋과 비교를 하였다.

<표 2> 다양한 크기의 사용자 행동 데이터에서의 협업 필터링, 일반선호도, 제안시스템의 추천의 정확도 비교



<표 2>에서 협업 필터링은 10,000 개의 트레이닝 셋에서 가장 낮은 정확도를 보이다가 데이터의 양이 많아질수록 정확도가 급격하게 높아지고 있다. 10,000 개의 트레이닝 셋에서 가장 낮은 정확도를 보이는 것으로 Cold-start 문제를 확인 할 수 있다.

일반적선호도는 많은 사람이 선호한 항목을 추천하는 방식으로, 통계적 특성을 가지고 있기 때문에 트레이닝 셋의 크기에 크게 영향을 받지 않는 것으로 나타나고 있다.

협업 필터링의 결과와 일반적으로 선호하는 항목, 사용자 이전 행동 데이터로부터 판단 된 사용자 성향, 최적의 추천전략 비율을 결정하는 알고리즘을 이용해서 하이브리드 된 추천방법을 사용하는 제안시스템의 경우, 먼저 모든 크기의 트레이닝 셋에서 가장 좋은 결과를 보이고 있다. 그리고 10,000 개의 트레이닝 셋에서의 결과가 협업 필터링보다 2 배정도 정확도가 향상되고 있으므로 Cold-start 문제를 해결하고 있다고 볼 수 있다. 또한, 시스템의 구성과 동작 과정에서 만약 사용자가 신규 사용자라서 이전 행동 데이터가 없다면 일반적으로 선호하는 항목을 추천하는 방식을 통해서 합리적인 정확도를 가지는 추천을 하고 있다.

5. 결론 및 향후 과제

본 논문에서는 개인화된 상품의 추천과 일반선호 상품의 추천비율을 사용자의 성향에 기반하여 조합하는 적응형 추천시스템을 제안하였다. 이를 통해 사용자 취향이 변하더라도 적절한 추천을 할 수 있고, Cold-start 문제, 신규 사용자 문제를 해결할 수 있다. 또한 저장해야 되는 데이터의 양을 늘리지 않고, 오프라인에서의 계산을 병행하여 수행시간 역시 비슷하게 유지하고 있다. 향후 과제로 영화 이외의 도메인에 적용하여 제안시스템이 가정하고 있는 사용자 성향 분석 및 추천법이 적용될 수 있는지 확인하고 발전시킬 계획이다.

참고문헌

[1] Bergasa-Suso J. Sanders D.A., and Tewkesbury G.E., "Intelligent Browser-based Systems to Assist Internet Users", IEEE Transactions on Education, vol.48, no.4, pp.580-585, Nov.2005.

[2] Zanker M., Jannach D., Gordea S., and Jessenitschnig M., "Comparing Recommendation Strategies in a Commercial Context", IEEE Intelligent Systems, vol.22, no.3, pp.69-73, May.2007.

[3] Hyung jun Ahn, "A new similarity measure for collaborative filtering to alleviate the Cold-starting problem", Information Sciences, vol.178, no.1, pp.37-51, Jan.2008.

[4] Felix Hernandez del Olmo, and Elena Gaudioso, "Evaluation of recommender systems: A new approach", Expert Systems with Applications, vol.35, no.3, pp.790-804, Oct.2008.

[5] Mohammad Yahya H., Al-Shamri, Kamal K., and Bharadwaj, "Fuzzy-genetic approach to recommender systems based on a novel hybrid user model", Expert Systems with Application, vol.35, no.3, pp.1386-1399, Oct.2008.