제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

그래프 기반 협동적 여과를 이용한 음악 추천 시스템

김형일, 이진석, 이정현, 조진관, 김경섭, 김준태
동국대학교 컴퓨터공학과
{hikim,jslee,jhlee,chinkwan,kskim,jkim}@dongguk.edu

A Music Recommendation System by Using Graph-based Collaborative Filtering

Hyungil Kim, Jinseok Lee, Jeonghyun Lee,
Chinkwan Cho, Kyoungsup Kim, Juntae Kim
Dept. of Computer Engineering, Dongguk University

요 약

본 논문에서는 각 사용자들의 취향에 맞는 음악을 추천하는 개인화된 음악 추천 시스템을 소개한다. 추천 시스템이란 사용자의 선호도를 분석하고 아이템들에 대한 사용자의 선호도를 예측하여 영화, 음악, 기사, 책, 웹 페이지 등과 같은 아이템들을 추천하는 시스템을 말한다. 추천 시스템들에서 가장 많이 사용하고 있는 협동적 추천 방식은 선호도 데이터를 기반으로 유사한 사용자들을 찾고, 유사 사용자들의 선호도를 기반으로 예측을 수행하는 것으로서, 여러 장점들이 있으나 희소성(sparsity) 문제와 확장성(scalability) 문제에 대해 취약점을 가지고 있다. 아이템들의 전체 수에 비해 매우 적은 수의 아이템 선호도 데이터만 존재한다면 사용자들의 유사도를 계산하기가 어려우며, 또한 사용자의 수가 늘어날수록 유사도 계산에 걸리는 시간이 급격하게 늘어남으로써 수백만 사용자가 있는 웹 사이트 등에서 실시간으로 추천을 수행하기 어렵다. 본 논문에서 소개하는 음악 추천 시스템은 이러한 문제점들을 해결하기 위해 그래프 기반 협동적 여과 기법을 사용한다. 그래프 기반 협동적 여과 기법은 기존의 협동적 여과 기법들과 달리 아이템들 사이의 연관관계를 그래프 모델로 표현하고 저장함으로써 묵시적인 선호도 정보들을 누적하여 희소성 문제를 해결하고, 추천 아이템을 선정하는데 필요한 계산 시간을 크게 단축하여 대규모 데이터에서 실시간 추천을 가능하게 한다는 장점이 있다.

1. 서론

추천 시스템은 사용자의 선호도를 분석하고 아이템들에 대한 사용자의 선호도를 예측하여 각 사용자의 취향에 맞는 영화, 음악, 기사, 책, 웹 페이지 등과 같은 아이템들을 추천하는 시스템이다. 다양한 추천 기법 중에 협동적 여과(Collaborative Filtering)는 상용화된 시스템에 성공적인 적용이 이루어진 기법이다.

협동적 여과 기반 추천 시스템은 아이템들에 대한 각 사용자들의 평가 정보를 이용한다[2][5][7]. 가장 일반적인 계산 방식은 사용자들 사이의 평가 정보를 비교하여 유사 사용자를 추출하고, 아이템들에 대한 유사 사용자의 선호도를 기반으로 특정 아이템에 대한 사용자의 선호도를 예측하는 것이다. 협동적 여과 기법은 아이템의 내용 정보를 필요로 하지 않기 때문에 내용을 분석하기 어려운 음악이나 영화 같은 아이템을 추천할 수 있다.

협동적 여과는 많은 장점이 있으며, 다양한 응용 시스템에 성공적으로 적용된 기법이지만 데이터의 희소성(sparsity) 문제와 확장성(scalability) 문제에 취약하다. 어떤 사용자에 대해 매우 적은 선호도 정보만 존재한다면 많은 유사 사용자를 찾기 어려우며, 따라서 추천 정확도는 낮아지게 된다. 또한 사용자의 수가 늘어날수록 유사도 계산에 걸리는 시간이 급격하게 늘어남으로써 수백만 사용자가 있는 웹 사이트 등에서 실시간으로 추천을 수행하기 어렵다.

본 논문에서는 협동적 여과 방식의 단점인 확장성 문제와 희소성 문제를 해결할 수 있는 그래프 기반 협동적 여과 기법을 사용하는 음악 추천 시스템을 소개한다. 그래프 기반 추천 방식은 아이템 추천을 계산하는 데에 소모되는 시간이 사용자의 수에 비례하여 늘어나지 않는 장점이 있으며, 사용자들의 명시적인 평가 선호도가 아닌 구매내력이나 정보의 접근횟수 등과 같은 묵시적인 선호 정보로부터 아이템들 사이의 연관관계를 그래프로 표현하여 관리함으로서 특정 사용자의 평가 정보가 희소하더라도 추천이 가능하다.

2. 관련 연구

협동적 여과 기반 추천 시스템은 많은 사용자로부터의 아이템에 대한 평가를 이용하는 기법이다. 사용자 기반(user-based) 협동적 여과는 사용자들 사이의 평가 정보를 비교하여 유사 사용자들을 추출하고, 유사 사용자들의 선호도를 기반으로 사용자의 아이템 선호도를 예측한다. 사용자 유사도는 피어슨(Pearson) 상관관계나 벡터 유사도와 같은 통계적 기법에 의해 계산된다. 유사도가 계산되고 나면 식 (1)에서와 같이 아이템에 대해 유사한 사용자들의 평가에 대한 가중 평균을 계산하여 사용자에 대한 아이템 선호도를 예측한다. 식 (1)에서 $P_{a,i}$ 는 아이템 i 에 대한 사용자 a 의 선호도 예측 값이고, $r_{u,i}$ 는 사용자 u 의 아이템 i 에 대한 평가 값이다. $S_{a,u}$ 는 사용자 a 와 사용자 u 사이의 유사도이고, n 은 유사 사용자들의 총 수이다.

$$P_{a,i} = \frac{\sum_{u=1}^{n}(s_{a,u} \cdot r_{u,i})}{\sum_{u=1}^{n} s_{a,u}} \quad (1)$$

GroupLens[7]와 같은 협동적 여과 시스템에서는 사용자의 선호도를 예측하는데 상관관계(correlation) 기반 기법이 사용되었고, 다양한 변형된 기법들이 추천 시스템의 정확도 향상을 위해 제안되었다. Breese 등[2]은 각 아이템에 얼마나 많은 사용자가 평가를 수행했느냐에 따라 서로 다른 가중치를 적용하는 방법 등을 제안하였고, Herlocker 등[5]은 다양한 유사도 계산 방식과 유사도 가중치 방법에 대한 실험을 수행한 바 있다. Billsus와 Pazzani[1], Sarwar 등[9]은 충분한 정보가 없을 때 사용자의 선호도 예측을 위해 속성 추출 기술을 적용하는 방법과, 사용자-아이템 선호도 행렬의 차원을 줄이기 위하여 SVD(Singular Value Decomposition)을 사용하는 방법을 제안하였다.

최근에는 Sarwar[10], Deshpande 등[4]에 의해 아이템 사이의 유사도를 이용하여 사용자의 증가에 따른 계산 복잡도 문제를 해결하면서 추천의 질을 높일 수 있는 아이템 기반(item-based) 협동적 여과 알고리즘이 제안되었고, Condliff[3], Popescul[8], Hoffman[6] 등에 의해 확률 이론을 바탕으로 하는 다양한 모델 기반(model-based) 추천 방법에 대한 연구도 수행되었다.

본 연구에서 제안하는 그래프 기반 협동적 여과는 기본적으로는 아이템 기반(item-based) 협동적 여과 알고리즘에 가까우나, 선호도 정보를 누적하여 아이템들 사이의 연관관계를 그래프로 표현함으로써 오프라인에서 아이템들의 유사도 계산을 수행할 필요가 없으며, 대용량 데이터에서 실시간 추천이 가능한 장점이 있다.

3. 그래프 기반 추천 알고리즘

3.1 기본 개념

일반적인 음악 서비스 사이트에서 모든 사용자로부터 풍부한 명시적 선호도(평가) 정보를 기대하기는 어렵다. 따라서 원활한 음악 추천 서비스를 구현하기 위해서는 음원의 구매나 실행 횟수 등과 같은 암시적인(implicit) 정보만을 이용하여 실시간으로 음악을 추천할 수 있는 기법이 필요하다. 또한 수백만의 사용자와 아이템을 다루게 되므로 일반적인 사용자 유사도 기반, 혹은 아이템 유사도 기반의 협동적 여과 기법을 적용하기 어렵다.

본 논문에서 제안하는 그래프 기반 추천 기법은 다수 사용자들의 음악 구매 기록(download), 음악 실행 횟수(play count), 개인 음악 목록(play list) 등 암시적 선호 정보들로부터 음악 사이의 연관 관계를 가중치 그래프로 표현하여 저장하고, 이러한 그래프를 이용하여 추천 대상 사용자에 대한 음악의 추천을 빠르게 수행하는 방법이다.

그래프 기반 추천 기법은 사용자의 암시적 선호 정보가 그래프 구조에 누적됨으로써 추천 실행 시간에 유사도 계산 등의 과정을 수행할 필요가 없어 추천 계산 시간이 사용자의 수나 아이템의 수에 비례하여 늘어나지 않는다. 또한 암시적인 선호 정보만으로 계산되는 아이템 사이의 연관 관계를 이용하므로 특정 사용자의 선호 정보가 희소하더라도 추천이 가능하게 되어, 종래의 협동적 여과 방식의 단점인 확장성 문제와 희소성 문제를 해결할 수 있다.

3.2 추천 알고리즘

그래프 모델을 사용하여 음악을 추천하는 과정은 크게 네 단계로 나눌 수 있다. 첫째, 각 음악 사이의 연관 관계를 그래프 모델로 표현하고 저장한다. 둘째, 추천 대상 사용자의 취향에 맞을 가능성이 있는 음악들을 저장된 그래프를 이용하여 찾아낸다. 셋째, 찾아진 음악들에 대하여 해당 사용자의 선호도 예측 값을 그래프 데이터를 이용하여 계산한다. 넷째, 산출된 예측 값을 이용하여 추천 대상 음악에 순위를 정하여 사용자에 적합한 음악을 추천한다.

1) 그래프 표현 및 선호 정보 저장

그래프에서 각 정점 M_i 는 각 음악을 나타내고, 정점사이의 간선 (M_i, M_j)은 각 음악 사이의 연관성을 나타낸다. 각 정점에는 각 음악에 대한 사용자들의 선호도 총합 $C(M_i)$가 저장된다. $C(M_i)$는 M_i 에 대한 총 구매 수, 다운로드 수, 실행 회수 또는 이들의 조합 등이 될 수 있다. 각 간선에는 각 간선이 연결하는 두 정점에 해당하는 음악의 동일 사용자에서의 동시 출현 빈도 총합 $C(M_i, M_j)$가 저장된다. 이때 동시 출현 빈도는 동일 사용자에서 발생한 음악들의 쌍을 나열하여 이 쌍들의 수로서 계산한다.

어떤 사용자 u 가 새로운 음악 M_i 를 다운로드하거나, 선곡 리스트에 넣거나, 실행하게 되면 사용자 u 의 변경된 음악 선호도 값 $\Delta C_u(M_i)$을 결정한 다음 그래프 데이터의 해당 음악에 대한 선호도 총합 $C(M_i)$를 $C(M_i)+\Delta C_u(M_i)$로 변경한다. 다음에 해당 사용자의 선호 음악 리스트에 있는 다른 모든 음악들 M_j 에 대하여 만일 $C_u(M_j) > C_u(M_i)$ 이면 그래프 데이터의 $C(M_i, M_j)$ 와 $C(M_j, M_i)$를 $C(M_i, M_j)+\Delta C_u(M_i)$와 $C(M_j, M_i)+\Delta C_u(M_i)$로 각각 변경한다. 이러한 작업이 사용자가 선곡 리스트를 만들거나, 음악을 다운로드하거나 실행할 때마다 수행되어

음악들 사이의 연관관계를 표현하는 그래프가 만들어지고 지속적으로 갱신된다. 그래프에서 $C(M_i)$ 값이 클수록 많은 사용자로부터 선호되어지는 인기있는 음악이며, $C(M_i, M_j)$ 값이 클수록 두 음악은 동시에 선호될 가능성이 높은 음악이 된다. 이러한 그래프는 무방향 가중치 그래프(undirected weighted graph)로서 총 음악 수에 해당하는 개수의 리스트로 저장될 수 있다.

2) 그래프 정보를 이용한 선호도 예측 및 추천

추천 대상 사용자 u 가 선정되면 해당 사용자의 선호 음악 리스트를 읽어 온다. 추천 대상 사용자가 $M_1..M_n$ 의 n 개 음악을 선호하는 것으로 알려졌으며, 각 음악에 대한 선호도가 $C_u(M_1)..C_u(M_n)$ 이라고 하자. 추천 대상 음악은 이 사용자가 이미 선호하는 것으로 알려진 음악들과 높은 연관성을 가지는 것으로서 이들은 위에서 정의한 그래프로부터 얻을 수 있다. 즉, 추천 대상 음악은 $M_1..M_n$ 과 간선으로 연결된 인접 정점에 해당하는 음악들로서, 정점 $M_1..M_n$ 각각으로부터 인접 정점 리스트를 읽으면 얻을 수 있다. 이들을 $A_1..A_m$ 이라 하자.

앞에서 검색한 각 음악 A_j 에 대한 추천 대상 사용자의 선호도 예측 값 $P_u(A_j)$를 해당 사용자가 이미 선호하는 것으로 알려진 음악 M_i 에 대한 선호도와 M_i 와 A_j 의 연관도로부터 식 2와 같이 계산한다. 이러한 계산은 검색된 인접 음악의 수 m 만큼 수행된다.

$$P_u(A_j) = \sum_{i=1}^{n} C_u(M_i)\, W(M_i, A_j) \quad (2)$$

이때 $W(M_i, A_j)$는 M_i 와 A_j 의 연관도로서 $MI(M_i, A_j)$ (Mutual Information)을 사용하면 그래프의 정점 및 간선 테이터에 저장된 값들로부터 바로 계산할 수 있다.

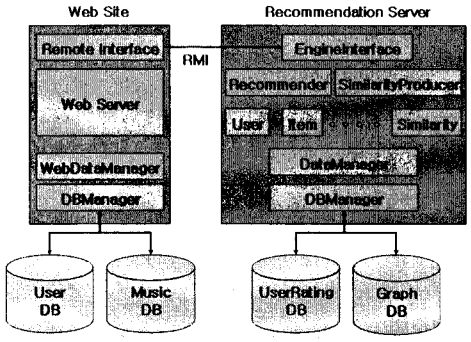
위의 식에 따라 $P_u(A_1)..P_u(A_m)$ 이 얻어지면, 이 값들을 내림차순으로 정렬하고, 추천하고자 하는 상위 K 개의 음악을 선택함으로써 최종적인 추천 음악 리스트가 완성된다. 이러한 추천의 계산 시간은 각 사용자의 평균 선호 아이템 수와 그래프의 정점당 평균 간선 수에 의해 결정되며, 총 사용자 수나 총 아이템 수에는 영향 받지 않는다.

3.3 시스템 구성

본 논문에서 구현한 음악 추천 시스템은 Windows server 상에서 Java로 구현되었으며, MySQL 데이터베이스를 사용한다. 추천 시스템이 사용하는 데이터는 사용자 선호도 데이터, 곡 정보 데이터, 선호도 데이터로부터 생성되는 음악 간의 연관관계 그래프 데이터 등이다. 그림 1은 그래프 기반 음악 추천 시스템의 구성도이다.

Web server는 웹에서의 UI를 제공하며, 음악 검색 및 추천, 플레이/다운로드에 대한 응답을 담당한다. Remote Interface는 음악 추천, 플레이, 다운로드 및 평가 등, Recommendation Server에서 처리하는 기능들에 대한 원격 메소드 호출을 담당한다.

Engine Interface는 Web server 측에서 호출된 메소드에 응답하는 모듈로서, 웹에서 호출되는 메소드를 크게 두 가지로 구분하여, 선호도 정보의 갱신에 관련된 메소드는 Similarity Producer로, 음악 추천에 관련된 메소드는 Recommender로 처리를 위임한다. Similarity Producer는 가중치 갱신에 관련된 메소드를 실질적으로 처리해주는 모듈로서, 플레이, 다운로드와 같은 암시적인 행위에 대해서 그래프 데이터의 정점 및 간선 값을 갱신한다. Recommender는 추천 계산을 수행하는 모듈로서, 그래프 데이터를 바탕으로 특정 사용자의 특정 음악에 대한 선호도 예측 값을 계산하고 사용자에 대한 추천 음악 리스트를 생성한다. Data Manager는 Recommendation Server에서 사용되는 각종 객체(User, Item, Rating, Similarity등)를 생성할 때 데이터베이스에서 적절한 테이블을 읽어서 객체에 값을 설정해주는 기능과, 객체의 현재 상태를 데이터베이스 테이블에 기록하는 역할을 담당하며, DB Manager는 데이터베이스에 대한 접속 및 질의를 담당한다.



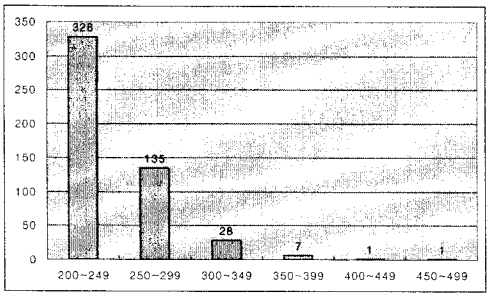
(그림 1) 그래프 기반 음악 추천 시스템 구성도

4. 실험

4.1 실험 데이터

본 논문에서 제안한 그래프 기반 추천 방식을 실험하기 위해 FunCake 사이트의 실제 data를 이용하였다. FunCake은 온라인 음악 서비스 사이트로서 스트리밍 서비스와 다운로드 서비스를 동시에 제공한다. FunCake 사이트에서 실험에 사용할 수 있도록 제공받은 실험용 데이터 중 본 논문의 실험에서 사용한 데이터는 2005년 7월부터 2006년 3월까지 FunCake 사이트에서 판매된 곡들에 대한 데이터로, 사용자의 개인정보를 제외한 사용자 ID, 곡 ID, 구매일이 나타나 있다.

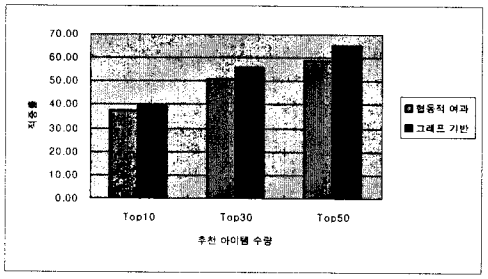
전체 데이터 집합에 있는 총 사용자 수는 821,562명이고, 총 곡의 수는 553,118개이며, 총 판매 수는 약 51만 건이다. 본 논문에서는 추천 정확도 실험을 위해 이 중 판매량 기준 상위 500개의 곡과 구매량 기준 상위 500명의 사용자를 추출하여 500x500 사이즈의 소규모의 실험용 데이터를 만들어 사용하였다. 즉, 선호도 데이터는 암시적인 선호도(구매)만 있는 500x500 불리언 행렬이 된다. 소규모 실험용 데이터 집합의 사용자 평균 연령은 26.5이며, 사용자 1인당 평균 구매 회수는 243.7, 표준편차는 36.8이다. 구매 회수에 따른 사용자 분포는 그림 2와 같다.



(그림 2) 구매 회수에 대한 사용자 분포

4.2 실험 결과

본 실험에서는 벡터 유사도를 이용한 사용자 기반 협동적 여과 방식과 본 논문에서 제안한 그래프 기반 협동적 여과 방식의 추천 정확도를 비교하였다. 실험에서는 우선 각 사용자의 구매 기록에서 데이터를 하나씩 차례로 제외하면서 나머지 데이터를 기반으로 추천을 수행하였다. 추천된 top-N 아이템 리스트에 제외된(선호하는 것으로 알고 있는) 아이템이 포함되면 적중(hit)이라고 하고, 각 사용자의 적중 비율을 총 평균한 적중률(hit ratio)로 성능을 비교하였다. 실험 결과는 그림 3과 같다.



(그림 3) 추천 정확도 비교

실험 결과 상위 10개의 아이템을 추천한 경우에 사용자 기반 협동적 여과 방법은 37.2%의 적중률을 나타내었고, 그래프 기반 협동적 여과 방법은 39.6%의 적중률을 나타내어 2.4%의 적중률 향상을 나타내었다. 추천 아이템의 수량을 증가시키게 되면 협동적 여과 방식과 그래프 기반 방식의 성능 격차는 더욱 크게 나타났다.

이러한 실험 결과는 본 논문에서 제안하는 그래프 기반 협동적 여과에 의한 추천 방법이 희소성 문제를 해결하고 빠른 추천 계산이 가능하다는 장점이 있을 뿐 아니라 실제 추천 정확도에 있어서도 더 나은 성능을 보일 수 있음을 시사한다. 앞으로 다양한 희소성을 갖는 데이터 집합을 이용한 성능 비교 실험과, 80만 사용자, 50만 아이템의 전체 데이터를 이용한 추천 정확도 및 추천 시간 실험을 통하여 이러한 장점들을 명확하게 확인할 수 있을 것이다.

5. 결론

본 논문에서는 협동적 여과 방식의 단점인 확장성 문제와 희소성 문제를 해결할 수 있는 그래프 기반 협동적 여과 기법을 사용하는 음악 추천 시스템을 소개하였다. 그래프 기반 추천 방식은 사용자들의 암시적인 선호 정보로부터 아이템들 사이의 연관관계를 그래프로 표현하여 관리함으로서 특정 사용자의 평가 정보가 희소하더라도 추천이 가능하고, 추천 계산 시간을 크게 단축하여 대용량 데이터에서 실시간 추천이 가능하다.

FunCake 사이트로부터 제공받은 데이터의 일부로 실험을 수행한 결과 본 논문에서 소개한 음악 추천 시스템은 추천 정확도에 있어서도 더 나은 성능을 보이는 것을 확인하였다. 향후 데이터 희소성에 따른 성능 변화에 대한 분석과 대규모 데이터를 이용한 추천 성능 실험을 수행할 예정이며, 암시적인 선호도 정보에 각 곡의 장르, 아티스트 등 추가적인 내용 정보를 함께 이용하여 추천의 정확도를 높이는 방안을 연구할 필요가 있다.

6. 참고문헌

[1] D. Billsus and M. J. Pazzani, "Learning Collaborative Information Filters," *Proceedings of the 15th International Conference on Machine Learning*, 1998.

[2] J. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, 1998.

[3] M. Condliff, D. Lewis, D. Madigan, and C. Posse, "Bayesian Mixed Effects Models for Recommender Systems," *Proceedings of ACM SIGIR '99 Workshop on Recommender Systems*, 1999.

[4] M. Deshpande and G. Karypis, "Item-Based Top-N Recommendation Algorithms," *ACM Transaction on Information Systems*, Vol.22, No.1, 2004.

[5] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering," *Proceedings of ACM SIGIR '99*, 1999.

[6] T. Hoffman, "Latent Semantic Models for Collaborative Filtering," *ACM Transaction on Information Systems*, Vol. 22, No. 1, 2004

[7] J. Konstan, B. Millr, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News," *Communications of the ACM*, Vol.40, No.3, 1997.

[8] A. Popescul, L. Ungar, D. Pennock, and S. Lawrence, "Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments," *Proceedings of 17th Conference on Uncertainty in Artificial Intelligence*, 2001.

[9] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of Dimensionality Reduction in Recommender System--A Case Study," *Proceedings of the ACM WebKDD Workshop*, 2000.

[10] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," *Proceedings of the 10th International WWW conference*, 2001.