메타데이터를 이용한 음악 추천 기법

이혜인 · 윤성대

부경대학교

Music Recommendation Technique Using Metadata

Hye-in Lee · Sung-dae Youn

Pukyong National University

E-mail : hilee.hyein@gmail.com

요 약

최근 디지털 음반시장의 성장으로, 들을 수 있는 음악의 양이 기하급수적으로 늘어나고 있다. 이로 인해 온라인 음원 서비스 이용자들은 마음에 드는 음악을 선택하는데 어려움을 겪고, 많은 시간을 낭비하게 되었다. 본 논문에서는 온라인 음원 서비스 이용자들이 겪는 선택의 어려움을 최소화하고, 낭비되는 시간을 줄이기 위한 추천 기법을 제안하고자 한다. 제안하는 기법은 개인정보의 이용 없이 아이템을 추천할 수 있는 아이템 기반 협업필터링 알고리즘을 사용한다. 더 정확한 추천을 위해 음원의 메타데이터를 이용하여 사용자의 선호도를 예측하고 선호도가 높은 Top-N개의 음악을 최종적으로 추천한다. 실험을 통해 제안하는 기법이 메타데이터를 이용하지 않을 때보다 추천 성능이 향상되는 것을 확인하였다.

ABSTRACT

Recently, the amount of music that can be heard is increasing exponentially due to the growth of the digital music market. Because of this, online music service users have had difficulty choosing their favorite music and have wasted a lot of time. In this paper, we propose a recommendation technique to minimize the difficulty of selection and to reduce wasted time. The proposed technique uses an item - based collaborative filtering algorithm that can recommend items without using personal information. For more accurate recommendation, the user's preference is predicted by using the metadata of the music source and the top-N music with high preference is finally recommended. Experimental results show that the proposed method improves the performance of the proposed method better than it does when the metadata is not used.

키워드

추천시스템, 협업필터링, 음악추천, 메타데이터

I. 서 론

우리는 정보의 홍수 시대에 살고 있다. 넘치는 정보들을 하나씩 확인해가며 필요한 정보인지 아닌지를 선택하기에는 시간이 오래 걸린다. 이런 시간을 단축하기 위해 추천시스템이 도입되면 많은 사람들이 편리함을 느끼고 이용할 것이다.

디지털 음반시장 또한 최근 엄청난 성장으로, 인터넷 음원서비스 사용자들은 선택의 어려움을 느끼고 있다. Lee와 Lee(2004)의 연구에서도 인터넷상에 제시되는 아이템이 많아질수록 사용자는 선택의 어려움을 느낀다고 보고된바 있다[1]. 따라서 이를 해결하기 위한 많은 연구들이 진행되고 있다.

본 논문에서는 개인정보 이용의 어려움을 고려하여, 사용자의 개인정보를 이용하지 않고 다양한 아이템을 추천할 수 있는 장점을 지닌 아이템 기반 협업필터링을 이용한 추천기법을 제안하고자 한다.

논문의 구성은 다음과 같다. 2장에서는 추천기법에 사용된 관련연구에 대해서 기술하고, 3장에서는 제안하는 추천기법을 설명하고 4장에서는 실험과 추천기법 성능을 비교한다. 마지막으로 5장에서는 결론과 향후 과제를 기술한다.

II. 관련 연구

2.1. 메타데이터

메타데이터는 "데이터에 관한 데이터"를 의미하며, 데이터의 하나 이상의 측면에 관한 정보를 제공하는 데이터로 정의된다. 메타데이터는 특정 데이터를 쉽게 추적하고 작업할 수 있는 데이터에 대한 기본정보를 요약하는데 사용된다.

예를 들어, 디지털 이미지에는 그림의 크기, 색상 깊이, 이미지 해상도, 이미지 생성시의 셔터 속도 및 기타 데이터를 설명하는 메타 데이터가 포함될 수 있다. 텍스트 문서의 메타 데이터에는 문서의 길이, 작성자, 문서 작성 시기 및 문서 요약에 대한 정보가 포함될 수 있다. 웹 페이지 내의 메타 데이터는 페이지 콘텐츠에 대한 설명과 콘텐츠에 연결된 키워드를 포함할 수 있다[2].

2.2. 최소-최대 정규화

측정단위는 데이터 분석에 영향을 줄 수 있다. 측정단위에 종속된 문제점을 방지하기 위해서는 데이터를 정규화 또는 표준화 해야 한다. 이러한 과정은 해당 데이터가 -1~1 또는 0~1과 같은 작은 범위 내에 위치하도록 한다. 최소-최대 정규화는 원 데이터에 대해 선형 변환을 한다. 속성 X에 대한 최솟값과 최댓값을 $\min_X$, $\max_X$라하고 정규화 범위 최솟값과 최댓값을 $\min_{scale}$, $\max_{scale}$ 라 하면 속성 X에 대한 정규화는 식 (1)과 같다.

$$X'_i = \frac{X_i - \min_X}{\max_X - \min_X}(max_{scale} - min_{scale}) + min_{scale} \quad (1)$$

정규화 된 값 X'_i는 $\min_{scale}$, $\max_{scale}$ 범위에 속하게 된다[3].

2.3. 코사인 유사도

추천시스템에서 아이템 간 유사도를 측정할 때 코사인 유사도를 사용한다. 온라인 쇼핑몰 amazon.com 에서도 아이템 간 유사도를 구할 때 코사인 유사도를 사용한다[4].

먼저 각의 코사인을 계산하고, 그런 다음 각을 0~180도 범위로 환산하는 아크코사인 함수를 적용해 코사인 유사도를 측정할 수 있다. 두 벡터 x와 y가 주어졌을 때 이들이 만드는 각의 코사인은 내적 x, y를 x와 y의 L2 norm(원점으로부터의 유클리드 거리) 로 나눈 값이다. 코사인 유사도는 식 (2)로 구할 수 있다[5].

$$similarity = \cos\theta = \frac{X_i \cdot Y_i}{\|X_i\|\|Y_i\|} = \frac{\sum_{i=1}^{n}(X_i \times Y_i)}{\sqrt{\sum_{i=1}^{n} X_i^2} \times \sqrt{\sum_{i=1}^{n} Y_i^2}} \quad (2)$$

2.4. 아이템 기반 협업 필터링

협업 필터링 시스템(collaborative filtering system)은 사용자 혹은 아이템 간 유사도를 기반으로 아이템을 추천하는 추천시스템이다. 사용자간 유사도를 기반으로 추천하는 시스템을 사용자 기반 협업 필터링이라고 하고, 아이템 유사도를 기반으로 추천하는 시스템을 아이템 기반 협업 필터링이라고 한다[6]. 유사도 계산은 자카드 거리, 코사인 거리 등을 사용한다[4].

III. 제안하는 기법

음악 추천을 위한 추천기법은 다음 네 단계로 구성된다. 음악의 메타데이터 추출, 청취횟수 데이터 정규화, 메타데이터를 이용한 음악 간 유사도 측정, 선호도 예측 및 음악추천 순이다. 추천 알고리즘은 아이템 기반 협업필터링을 사용한다.

3.1. 음악 메타데이터 추출

음악 간 유사도 측정 시 메타데이터를 이용하기 위해서 음악의 메타데이터를 추출한다. 음악 식별자인 TrackID, SongID, 기본정보인 제목, 발행연도를 추출하고 오디오 분석정보인 Tempo, Loudness를 추출한다.

3.2. 청취횟수 데이터 정규화

청취자마다 선호하는 곡의 청취 횟수가 다르기 때문에 청취 횟수를 선호도 수치로 변환하기 위한 기준이 필요하다. 따라서 청취횟수를 최소-최대 정규화하여 1~10까지의 선호도로 구분한다. 음악 i에 대한 청취자 u의 정규화된 선호도 $P_{u,i}$는 식 (1)에 대입하여 식 (3)로 계산할 수 있다.

$$P_{u,i} = \frac{L_{u,i} - \min(L_u)}{\max(L_u) - \min(L_u)}(S_{\max} - S_{\min}) + S_{\min} \quad (3)$$

$L_{u,i}$는 음악 i에 대한 청취자 u의 청취 횟수이다. $\min(L_u)$는 청취자 u가 가장 적게 청취한 횟수이다. $\max(L_u)$는 청취자 u가 가장 많이 청취한 횟수이다. $S_{\max}$는 맵핑하고자 하는 스케일의 최댓값 10이다. $S_{\min}$는 맵핑하고자 하는 스케일의 최솟값 1이다. $\max(L_u)$와 $\min(L_u)$가 같은 청취자의 데이터는 제외한다. 왜냐하면 분모가 0이 되어 계산이 불가능해지기 때문이다.

3.3. 메타데이터를 이용한 음악 간 유사도 측정

청취자가 많이 들은 음악 순으로 M곡을 선정하고, 25곡 이상의 음악을 청취한 청취자 N명의 데이터를 사용한다.

음악 M곡과 청취자 N명에 대한 음악X청취자 매트릭스를 만들고 정규화된 선호도 데이터를 매트릭스에 채워 넣는다. 표 1은 음악X청취자 매트릭스의 예시이다.

표 1. 음악(i_x)X청취자(u_y) 매트릭스.

	i_1	i_2	i_3	i_4	i_5	…	i_{99}	i_{100}
u_1	3					…		1
u_2	3	3		7		…	2	
u_3			5			…		
⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮
u_{499}		10				…		
u_{500}				3	2	…		

먼저 목표 청취자를 정하고, 해당 청취자가 청취한 음악과 청취하지 않은 음악을 구분한다. 그리고 식 (2)을 이용하여 코사인 유사도를 계산한다. 유사도는 하나의 음악 당 N명의 청취자 선호도 수치를 벡터화한 후 계산한다. 그리고 음악의 tempo와 loudness 수치도 벡터화한 후 코사인 유사도를 구한다. 각 유사도 값을 1:1 비중으로 하여 가산한 값을 최종 유사도 값으로 저장한다.

3.4. 선호도 예측 및 음악 추천

예측한 선호도 값을 이용하면 최종적으로 생성할 추천목록의 음악 중 Top-N개의 음악을 선택할 수 있다. 식 (4)는 예측 선호도 값을 구하는 수식이다[7].

$$P_{u,i} = \frac{\sum_{all\ similar\ items,N}(s_{i,N} * R_{u,N})}{\sum_{all\ similar\ items,N}(|s_{i,N}|)} \quad (4)$$

$P_{u,i}$는 목표 청취자 u가 청취하지 않은 음악 i에 대한 예측 선호도이다. $s_{i,N}$ 은 청취자 u가 청취한 음악 중 음악 i와 가장 유사하다고 판단되는 상위 n개의 음악 N과의 유사도이다. $R_{u,N}$은 청취자 u가 청취한 음악 N의 선호도이다.

예측 선호도 $P_{u,i}$을 구하기 위해 유사도가 높은 상위 n개의 음악을 3~10으로 변경하면서 계산한다. 예측 선호도가 높은 순으로, 최종 추천하는 음악의 개수를 5~30까지 5단위로 변경하면서 추천목록을 생성한다.

IV. 실험 및 결과

4.1. 실험 데이터

실험에 사용된 메타데이터는 Million Song Dataset의 summary 데이터와 장르 속성데이터를 이용하여 추출했다. summary 데이터는 음악의 제목, 식별자, 발행연도 등과 같은 기본정보와 오디오분석 정보로 구성되어있다. 장르 속성데이터는 음악의 식별자와 장르로 구성되어있다. 본 실험에서는 추출한 데이터와 장르 속성데이터가 모두 존재하는 음악에 대한 데이터를 사용했다.

청취 횟수 데이터로는 Million Song Dataset의 공식 사용자 데이터 집합인 The Echo Nest의 음악ID-사용자ID-청취횟수 데이터를 사용했다. 추천을 위해 추출한 메타데이터에 존재하는 음악을 청취한 데이터만 사용했으며, 추출한 청취 데이터를 최소-최대 정규화하여 1~10범위의 선호도 값으로 변환했다.

청취자가 많이 들은 음악 순으로 100곡을 선정하고, 25곡 이상의 음악을 청취한 청취자500명의 데이터 10,272건의 선호도 데이터를 사용했다. 실험을 위해 train dataset을 70%, test dataset을 30%로 각각 7190건과 3082건으로 나누었다.

4.2. 추천 기법 성능비교

train dataset을 사용하여 음악 간 유사도와 예상 선호도를 계산하고, 목표 청취자에게 추천할 음악 목록을 생성했다. 생성된 예상 선호도와 추천 음악 목록을 test dataset와 비교하여 추천기법의 성능을 확인했다. 음악 선호도 예측 시 유사한 음악 개수를 10개로 지정했을 때 추천시스템 성능이 가장 우수했다. 유사음악 개수를 10으로 고정하고, 예상 선호도가 가장 높은 음악 Top-N을 5에서 30까지 5단위로 변경하면서 전통적인 아이템기반 협업필터링 기법으로 추천했을 때와 제안하는 추천기법으로 추천했을 때의 성능을 비교했다.

성능의 척도로는 정밀도, 재현율, F1척도를 이용했다. 정밀도는 추천시스템에서 추천된 아이템들 중 실제로 사용자가 선호한 아이템의 비율이고, 재현율은 사용자가 선호한 아이템들 중 실제로 추천된 아이템의 비율을 나타낸다. 정밀도와 재현율은 식 (5),(6)로 구할 수 있다[8].

$$\text{정밀도} = \frac{TP}{TP+FP} \quad (5) \qquad \text{재현율} = \frac{TP}{TP+FN} \quad (5)$$

여기서 True positive(TP)는 추천한 아이템 들 중 실제로 선호한 아이템 개수, False positive(FP) 는 추천한 아이템 중 선호하지 않은 아이템 개수이다. 따라서 $TP+FP$ 는 총 추천한 음악의 개수가 된다. 그리고 False negative(FN)는 추천하지 않았지만 선호한 아이템 개수이다. 따라서 $TP+FP$는 청취자가 청취한 음악의 개수가 된다. 식 (7)은 F1척도를 나타낸다[9].

$$F1 = \frac{2 \times \text{정밀도} \times \text{재현율}}{\text{정밀도} + \text{재현율}} \quad (6)$$

표 2는 추천하는 음악개수 Top-N의 변화에 따른 정밀도, 재현율, F1척도 값을 나타낸다. 그림 2, 3, 4는 표 2를 각각 그래프로 나타낸 것이다.

Top-N이 30에 가까워지면 전통적인 추천기법과 제안하는 추천기법의 성능차이가 줄어들지만, 청취자가 쉽게 선택할 수 있는 범위인, Top-N이 5~15인 범위에서는 제안하는 기법의 F1척도가 각각 0.032, 0.021, 0.0142 만큼 우수한 성능을 보였다.

표 2. 추천 음악개수(Top-N)에 따른 전통적인 협업필터링 기법과 제안하는 기법의 성능비교.

	전통적 추천기법			제안하는 추천기법		
Top-N	정밀도	재현율	f1척도	정밀도	재현율	f1척도
5	0.0850	0.0596	0.0700	0.1163	0.0910	0.1021
10	0.0905	0.1344	0.1082	0.1087	0.1587	0.1290
15	0.0923	0.2026	0.1268	0.1030	0.2233	0.1410
20	0.0913	0.2696	0.1364	0.0974	0.2815	0.1447
25	0.0913	0.3295	0.1430	0.0942	0.3434	0.1479
30	0.0915	0.3988	0.1489	0.0934	0.4049	0.1518

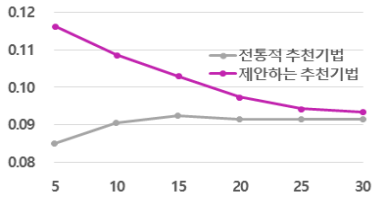

그림 2. 추천 음악개수에 따른 정밀도.

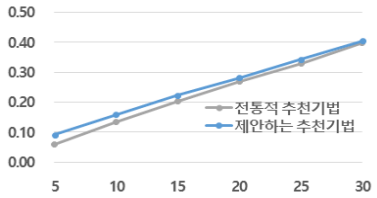

그림 3. 추천 음악개수에 따른 재현율.

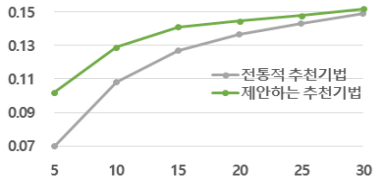

그림 4. 추천 음악개수에 따른 F1척도.

V. 결 론

본 논문에서는 청취자가 마음에 드는 음악을 선택하는 어려움을 줄이기 위해, 청취 횟수 데이터와 음원의 메타데이터를 사용하여 음악을 추천하는 기법을 제안했다. 데이터는 Million Song Dataset에서 추출한 메타데이터와 The Echo Nest의 음악ID-사용자ID-청취횟수 데이터를 사용하였으며 제안한 기법은 전통적인 아이템기반 협업필터링보다 더 나은 추천 성능을 보였다.

실제 음악추천에 적용하기 위해서는 음악 간 유사도를 빠르게 계산하고 추천목록을 생성할 수 있어야 한다. 빠른 계산과 추천을 위해 데이터를 적절하게 샘플링 하면 실제 음악추천의 성능 향상에 기여할 수 있을 것으로 기대된다. 또한 메타데이터의 장르 속성도 이용하여 속성별 최적의 가중치를 찾는다면 더더욱 좋은 결과가 있을 것이라고 예상된다.

참고문헌

[1] B. K. Lee, W. N. Lee, “The Effect of Information Overload on Consumer Choice Quality in an On-line Environment,” Psychology and Marketing, vol. 21, no. 3, pp. 159–183, Mar. 2004.

[2] P. Caplan, “Metadata fundamentals for all librarians,”American Library Association, pp. 1-11, 2003.

[3] D. Pyle, “Data preparation for data mining,” morgan kaufmann, pp. 224-256, 1999.

[4] G. Linden, B. Smith, and J. York, “Amazon. com recommendations: item-to-item collaborative filtering,” IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, Jan. 2003.

[5] M. S. Charikar, “Similarity estimation techniques from rounding algorithms,” Proceeding of the thiry-fourth annual ACM symposium on Theory of computing, pp.380-388, May. 2002.

[6] G. Adomavicius, A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” IEEE transactions on knowledge and data engineering, vol. 17, no. 6, pp.734-749, Jun. 2005.

[7] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-based collaborative filtering recommendation algorithms,” Proceeding of the 10th international conference on World Wide Web, pp.285-295, Apr. 2001.

[8] J. Lee, D. Lee, Y.-C. Lee, W.-S. Hwang, S.-W. Kim,“Improving the accuracy of top-N recommendation using a preference model,” Information Sciences, vol. 348, no. 20, pp. 290-304, Jun. 2016.

[9] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, “Analysis of Recommendation Algorithms for E-Commerce,” Processing of the 2nd ACM Conference on Electronic Commerce, pp.158-167, Oct. 2000.