Infinite Relational Model 기반 Co-Clustering을 이용한 영화 추천

김병희 · 장병탁
Byoung-Hee Kim and Byoung-Tak Zhang

서울대학교 컴퓨터공학부
E-mail: {bhkim, btzhang}@bi.snu.ac.kr

요 약

사람의 영화에 대한 선호도에는 개인의 특성과 영화의 속성을 기반으로 하는 다양한 요인이 연관되어 있다. 영화 추천을 위한 사용자-영화-선호도 연관 관계의 분석 기법으로서, 다중 개념 탐색 기법의 특성을 지닌 infinite relational model (IRM)의 활용 가능성을 확인하고, 이를 기초로 영화 선호 유형에 따른 사용자-영화 군집을 탐색한다. 별점으로 표현되는 명시적인 선호도 데이터에 영화 컨텐츠 관련 메타데이터를 추가하여 학습 데이터를 구성하고, 이에 IRM을 적용하여 공군집화(co-clustering)를 수행한 결과, 해석 가능한 다양한 명시적 연관 관계를 얻을 수 있었다. 공군집화 결과를 기초로 개인화 추천에서의 동적 모델 구축 방안을 논의한다.

키워드 : 추천 시스템, 군집 선호도, 연관 관계 분석, 영화 추천, Infinite Relational Model, Co-Clustering

1. 서 론

추천 시스템의 응용 분야에서 영화 추천은 인터넷의 활성화 초기부터 매우 중요한 비중을 차지하였다. MovieLens, Netflix, IMDb 등의 대규모 영화 정보 및 별점 정보를 기반으로 하여 자동 추천 시스템은 연구개발 분야와 응용 분야에서 괄목한 만한 성장을 하였다. 그러나, 개인의 영화에 대한 선호도를 정량화하고 개인화된 추천을 수행하는 문제는 여전히 난제로 남아 있다.

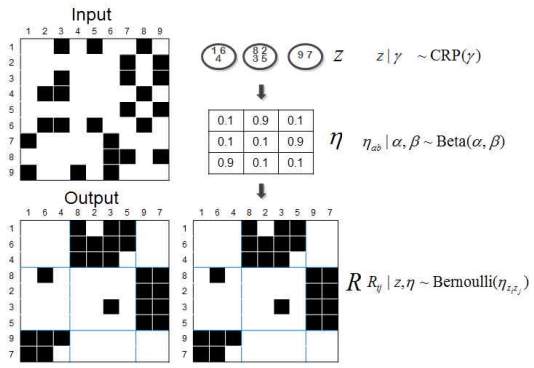
영화 추천의 기본 요건은 사용자-영화-선호도 연관 관계 파악에 있다. 본 논문에서는 Infinite Relational Model (IRM)[1]을 이용하여 이러한 연관 관계를 자동 추출한 결과를 정리한다. IRM은 비모수적 베이지안 기법을 이용한 다중 개념 탐색 프레임웍이다. 다양한 요인(type) 간의 연관 관계(relation)를 개념(concept) 수준에서 추출하는 것을 목표로 하며, 두 요소 간의 다중 연관 관계, 셋 이상 요소 간의 고차 연관 관계 분석 등의 활용이 가능하다. 또한, 기존의 공군집화(co-clustering)의 일반화된 모델로도 다룰 수 있는 장점이 있다.

영화 평점 데이터에서 공군집화를 기반으로 하는 다양한 추천 기법이 연구되었으나[2][3], IRM 기반 공군집화와는 달리 군집의 수를 미리 정해야만 하는 한계가 있다. 협업적 필터링에 다양한 사용자 및 아이템 요인을 추가하는 혼합 추천기법 또한 꾸준히 연구되고 있으나[4][5], 다중/고차 연관 관계의 탐색에는 한계가 있다.

영화의 메타데이터를 추가로 고려하여, 사용자-영화-선호도 간의 연관 관계를 보다 명시적으로 설명하고자, IRM을 이용한 공군집화를 수행하였다. 영화의 메타데이터를 고려한 공군집화 기법과 실험 결과를 보인다. 실험을 통해 파악한 다양한 연관 관계를 기반으로 영화 추천의 여러 난제에 대한 적용 방안을 소개한다.

2. IRM을 이용한 공군집화

Infinite Relational Model (IRM)은 데이터에서 유의미한 개념(concept)을 발견하는 것을 목표로 하는 비모수적 베이지안 모델이다. 1차 술어(unary predicates)로 표현되는 개념 간 또는 요인(type) 간의 연관 관계 R을 발견할 수 있는 기법으로서, 모델에 설정한 모든 요인 및 연관 관계에 대한 군집화를 수행하며, 이는 기계학습 분야에서 공군집화(co-clustering)로 표현하는 기법과 일치하



(a) (b) (c)

그림 1. IRM을 이용한 연관 관계 발견. (a) 입력 및 출력 행렬 (b-c) 생성 모델로서의 IRM 개념 및 식

되, 2차 행렬 이상의 다양한 데이터군에 대해서도 적용 가능한 보다 일반적인 모델이다.

기본적인 사례로서 사람들(T) 간의 선호(R) 여부를 $R: T \times T \rightarrow \{0,1\}$로 표현했을 때 그림 1과 같이 군집 및 군집 간 선호 관계를 발견할 수 있다. IRM의 목적함수는 다음의 사후확률을 최대화하는 z를 찾는 것이다:

$$P(z|R) \propto P(R|z)P(z),$$

$$P(R|z) = \prod_{a,b} \frac{B(m_{ab} + \alpha \overline{m}_{ab} + \beta}{B(\alpha, \beta)},$$

m_{ab} 및 $\overline{m}_{ab}$는 각각 클래스 a와 b 사이의 연결선에 부여된 가중치가 1, 0인 관측 횟수이고, $B(\bullet, \bullet)$는 Beta 함수이다. 개체 i에 대한 클러스터 할당은 다음의 식으로 표현되는 중국식당 프로세스(chinese restaurant process, CRP)로 정해진다:

$$P(z_i = a | z_1, ..., z_{i-1}) = \begin{cases} \frac{n_a}{i-1+\gamma} & n_{a>0} \\ \frac{\gamma}{i-1+\gamma} & a \text{ is a new class} \end{cases}$$

3. 메타데이터 기반 사용자-영화-선호도 연관 관계 탐색

3절에서는 영화 추천에 필수적인 사용자 및 영화의 프로파일을 IRM 기반 공군집화를 이용하여 생성하는 방법에 대해 소개한다. 사용자와 영화의 선호도 관계를 명시적으로 부연 설명하기 위하여, 영화의 메타데이터를 도입한다. 즉 다음의 모델을 IRM에 적용한다:,

$$R_1 : T_1 \times T_2 \rightarrow \{0,1\} \text{ (prefers)},$$

$$R_2 : T_2 \times T_3 \rightarrow \{0,1\} \text{ (contains)}$$

(T_1: 사용자, T_2: 영화, T_3: 영화의 메타데이터)

영화의 메타데이터로 영화의 다양한 속성에 대한 체계적 시스템으로서 jinni.com의 Entertainment Genome (EG) [6]을 사용한다. EG는 영화의 속성을 12가지 genome, 이천여 개의 gene으로 표현하며, 지니닷컴의 비디오 컨텐츠 추천의 기반이 되는 핵심 taxonomy이다.

IRM을 통해 파악할 수 있는 사용자-영화-선호도 연관 관계의 예를 그림 2에서 살펴볼 수 있다. 예를 들면 사용자 군집 u3는 영화 그룹 m2를 선호하며, m2에는 g6 그룹의 gene이 집중적으로 포함되어 있다.

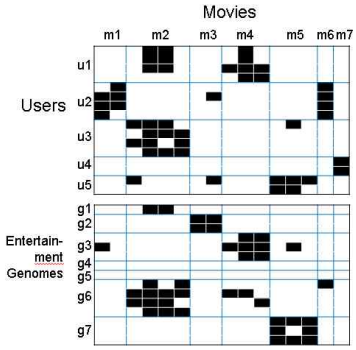


그림 2. IRM 기반 공군집화를 적용하여 발견한 사용자-영화-선호도 연관 관계. 블록별로 검게 채워진 영역이 클수록 블록에 연관된 서로 다른 요소의 그룹 간에 강한 연관 관계가 있다고 판단할 수 있다.

4. 실험 및 분석

대표적인 영화 추천 데이터로서 MovieLens 100k 데이터셋을 활용하여 IRM 기반 공군집화를 적용한 결과를 분석한다. MovieLens 100k 데이터셋은 943 사용자의 1682개 영화에 대한 1~5점 별점 10만 개로 구성되어 있다. 1682개의 영화 중 EG 추출 가능한 1601개의 영화를 사용한다. EG는 지니닷컴에서 영화별 gene을 수집하여 데이터를 구성하였다.

IRM의 입력 데이터 및 수행 결과 얻은 군집의 수는 표 1과 같다. 학습 결과 생성된 공군집 중에서 고밀도 ($\eta_{rating} > 0.6$, $\eta_{mg} > 0.9$)의 블록을 기반으로 발견한 선호 영화에 관련된 연관 관계를 그림 3에 정리하였다. 비슷한 맥락에서, 사용자의 비선호 패턴에 대한 EG 관점에서의 설명이 가능하다.

5. 논의 및 결론

본 논문에서는 IRM을 이용한 공군집화를 적용하여, 사용자의 영화에 대한 선호 관계를 파악하고, 영화의 메타데이터로서 Entertainment Genome을 도입하여, 선호

표 1. MovieLens 100k 및 지니닷컴의 EG를 활용한 공군집화 결과 (score=-206457.754)

요인	크기	군집수	군집 크기 (최소/최대/평균)
사용자 (T_1)	943	24	1/286/39.3
영화 (T_2)	1601	27	9/158/59.3
EG (T_3)	874	110	1/212/7.9

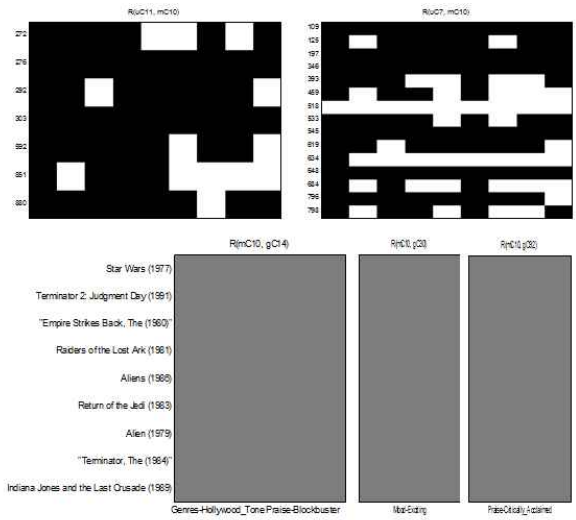


그림 3. 사용자그룹 7, 11은 영화그룹 10을 선호하며, 이 영화그룹은 EG의 14, 30, 92번 그룹의 특성(헐리우드풍, 흥분모드, 주요 영화제 수상후보)을 강하게 가지고 있다

에 대한 설명 및 사용자 그룹의 특성을 명시적으로 파악한 결과를 정리하였다.

실제 추천 환경에서는, 별점과 같은 사용자의 명시적인 선호도 지표를 확보하는 것은 고비용 작업이며 사용자의 수고를 필요로 한다는 점에서 역효과를 낼 수 있다. 이러한 이유로, 사용자의 선호도를 추정할 수 있는 다양한 간접적 피드백 정보를 확보하고 활용하는 것이 필요하다. 그러나, 간접적 피드백 정보에서 사용자의 선호도를 예측하는 일은 매우 도전적인 과제이다[7].

IRM 공군집화 기반의 연관관계 분석 결과는 다음과 같이 간접적 피드백 정보 기반의 선호도 예측에 활용될 수 있다. 범용(generic) 사용자 모델로서, 사용자 군집 및 군집과 EG 간의 연관 관계를 활용한다. 특히, 개인의 영화에 대한 간접적 피드백 정보에서 선호도를 추론하기 위한 basis로서 공군집화 결과를 활용할 수 있다.

감사의 글

이 논문은 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734, Videome), 정부(산업통상자원부)의 재원으로 한국산업기술평가관리원의 지원(KEIT-10035348 (mLife), KEIT-10044009)을 일부 받았음.

참고문헌

[1] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, “Learning Systems of Concepts with an Infinite Relational Model,” AAAI 2006, pp. 381 - 388.
[2] B. Xu, J. Bu, C. Chen, and D. Cai, “An Exploration of Improving Collaborative Recommender Systems via User-item Subgroups,” WWW 2012, pp. 21 - 30.
[3] N. Mirbakhsh and C. X. Ling, “Clustering-based factorized collaborative filtering,” RecSys 2013, pp. 315 - 318.
[4] C. Wang and D. M. Blei, “Collaborative topic modeling for recommending scientific articles,” KDD 2011, pp. 448 - 456.
[5] D. Agarwal and B.-C. Chen, “fLDA: matrix factorization through latent dirichlet allocation,” WSDM 2010, pp. 91 - 100.
[6] Entertainment Genome: http://www.jinni.com/info/entertainment-genome.html
[7] F. Ricci, L. Rokach, and B. Shapira, “Introduction to Recommender Systems Handbook.,” in Recommender Systems Handbook, 2011, pp. 1 - 35.