Journal of Korean Institute of Intelligent Systems, Vol. 24, No. 4, August 2014, pp. 443-449
http://dx.doi.org/10.5391/JKIIS.2014.24.4.443

Infinite Relational Model 기반 Co-Clustering을 이용한 영화 추천

Movie Recommendation Using Co-Clustering by Infinite Relational Models

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

서울대학교 컴퓨터공학부
School of Computer Science & Engineering, Seoul National University

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 발견하였다. 공군집화 결과를 기초로 개인화 추천에서의 다양한 활용 방안을 논의한다.

키워드 : 추천 시스템, 군집 선호도, 연관 관계 분석, 영화 추천, Infinite Relational Model, 공군집화

Abstract

Preferences of users on movies are observables of various factors that are related with user attributes and movie features. For movie recommendation, analysis methods for relation among users, movies, and preference patterns are mandatory. As a relational analysis tool, we focus on the Infinite Relational Model (IRM) which was introduced as a tool for multiple concept search. We show that IRM-based co-clustering on preference patterns and movie descriptors can be used as the first tool for movie recommender methods, especially content-based filtering approaches. By introducing a set of well-defined tag sets for movies and doing three-way co-clustering on a movie-rating matrix and a movie-tag matrix, we discovered various explainable relations among users and movies. We suggest various usages of IRM-based co-clustering, espcially, for incremental and dynamic recommender systems.

Key Words : Recommender systems, Group preferences, Relational Analysis, Movie Recommendation, Infinite Relational Model, Co-Clustering

1. 서 론

추천 시스템의 응용 분야에서 영화 추천 시스템은 1990년대 초중반의 인터넷의 활성화 초기부터 매우 중요한 비중을 차지하였다. MovieLens, Netflix, IMDb 등의 대규모 영화 정보 및 별점 정보를 기반으로 하여 자동 추천 시스템은 연구개발과 응용 분야 모두에서 괄목한 만한 성장을 하였다. 그러나 개인의 영화에 대한 선호도를 정량화하고 이를 기반으로 개인화된 추천을 수행하는 문제는 여전히 난제로 남아 있다.

추천 시스템의 기본 요건은 사용자의 추천 대상 아이템에 대한 선호도 예측이며, 다수의 사용자와 아이템 간의 연관 관계에 대한 다양한 모델이 연구되었다[1]. 협업 필터링(collaborative filtering) 방식의 추천 시스템은 수집된 평점 데이터를 기반으로, 평점 분포가 유사한 사용자 또는 아이템 정보를 종합하여 추천을 하거나, 사용자와 아이템 간의 연관 관계에 대해 모델을 구축한 후 이를 기반으로 추천을 시도한다. 내용 기반 추천(content-based filtering) 방식에

접수일자: 2014년 3월 9일
심사(수정)일자: 2014년 4월 1일
게재확정일자 : 2014년 6월 2일
† Corresponding author
본 논문은 미래창조과학부의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734), 산업통상자원부의 재원으로 한국산업기술평가관리원의 지원(KEIT-10035348, KEIT-10044009)을 일부 받았음.

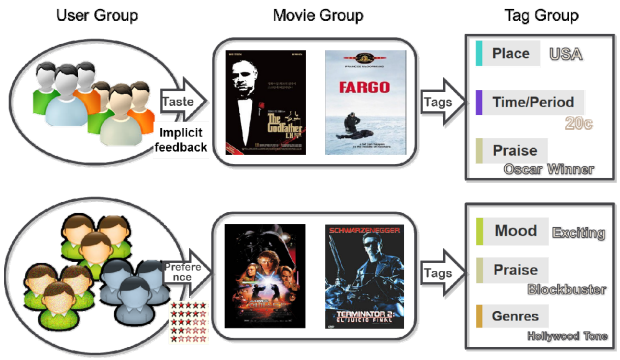


그림 1. 명시적 또는 비명시적으로 표현되는 영화 선호 데이터에 영화의 태그 정보를 추가하고 연관성 분석을 수행하여, 사용자/영화/태그 간의 개념적 연관 관계를 발견하고 이를 영화 추천에 활용할 수 있다.

Fig. 1. By relational analysis on preference data, either explicit or implicit, and additional movie tag data, we can discover various conceptual connections among user/movie/tag groups and utilize this information for movie recommendation.

서는 평점 데이터 외에 사용자의 인구 통계학적 정보, 아이템의 다양한 속성 등을 추가로 수집하여 사용자와 아이템 간의 연관 관계에 대해 보다 명시적인 모델을 구축한다. 지식 기반 추천(knowledge-based recommendation) 방식의 경우 추천 대상 아이템에 대한 내용 외에도 해당 분야의 전문적 지식을 가공하여 규칙과 제한조건 등을 생성하고 이를 기반으로 보다 안정적인 추천을 시도한다. 각 방식별로 다양한 장단점이 있으나 공통적으로 겪는 어려움은 평점 데이터로 대표되는 선호도 정보가 매우 희박하다는 점이다.

추천 시스템 구축에 필요한 선호도 데이터에 거의 예외 없이 희박성 문제가 수반되는 데에 대해 군집화(clustering) 기반의 모델링을 자연스럽게 적용해볼 수 있다. 희박성이 상대적으로 적은 사용자 또는 아이템 군집 단위에서 1차적인 분석을 수행한 후 개별 사용자/아이템 수준의 분석을 추가로 시도함으로써, 추천의 정밀도를 높이고 나아가 해당 아이템 시장에 대한 거시적인 패턴 파악도 기대할 수 있다[2]. 협업 필터링 분야에서는 사용자와 아이템 군집을 동시에 탐색하는 이중 군집화(bi-clustering) 또는 공군집화(co-clustering) 기반의 추천 기법도 다양하게 제시되고 있다[3][4]. 협업 필터링을 근간으로 내용 기반 추천 방식을 결합하는 하이브리드 기법에서도 군집화 기법이 다양한 방식으로 활용되고 있다[5][6]. 군집화 기반 방식에서의 가장 중요한 한계는 군집의 수를 실험자가 지정해야 한다는 점이다. 더불어, 단순한 군집화 기법만으로는 군집 수준에서의 연관성 분석이 힘든 단점이 있다.

본 논문에서는 군집화 기반의 연관 관계 추출 및 추천 기법으로서 Infinite Relational Model (IRM)[7]에 주목한다. IRM은 비모수적 베이지안 기법을 이용한 다중 개념 탐색 프레임웍으로서 소개된 알고리즘이다. 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, 일반화 및 확장된 공군집화(co-clustering) 모델로도 다룰 수 있는 장점이 있다.

영화 추천 측면에서 IRM은 다양한 장점이 있다. 우선 추천에 필요한 다양한 종류의 데이터를 동시에 다룰 수 있다는 점이다. 공통적으로 쓰이는 평점 데이터와 함께, 비명시적 선호도 데이터, 사용자 프로파일, 영화 프로파일 등의 다중 데이터에 대해 동시에 공군집화를 수행할 수 있다. 두 번째 장점으로 동종 및 이종의 군집 간 연간 관계를 자동으로 탐색할 수 있는 특성이 있다. [7]에서는 단일 행렬에 대한 단순 군집화뿐만 아니라, 질병 및 증상 관련 온톨로지(ontology) 학습, 국가 간의 다양한 정치적 관계 규명 등의 결과를 보인 바 있다. 영화 추천의 경우 영화 및 사용자를 표현하는 다중 정보에서 선호와 관련된 다양한 연관 관계를 추출할 수 있다.

본 논문에서는 그림 1에서와 같이, 영화에 대한 사용자들의 명시적 또는 비명시적 선호도 패턴과 영화에 대한 태그 기반의 메타 데이터에 IRM 기반의 공군집화를 적용하여, i) 연관된 이종 군집 정보를 기반으로 한 거시적인 추천 시스템 구축 및 ii) 영화의 선호도와 관련하여 데이터에 내재된 의미적 연관성 '발견'을 시도한다.

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.

이하 논문은 다음과 같은 순서로 구성되었다. 제2장에서는 IRM 기법에 대해 소개하며, 특히 공군집화 기법으로서 추천에 활용 가능한 특징에 초점을 둔다. 제3장에서는 군집화 기반의 추천시스템에 대해 정리하고, 영화의 메타 데이터를 도입하여 영화 선호 패턴에 연관된 구조 탐색 방법을 정리한다. 제4장에서 IRM 기반의 공군집화 기법을 실제 데이터에 적용하여 군집 단계에서 발견한 주요 연관 관계를 보고한 후, 제5장에서는 실제 추천에서의 다양한 활용 방안을 논의한다. 제6장에서 요약 및 결론을 도출한다.

2. IRM을 이용한 공군집화

Infinite Relational Model (IRM)[7]은 데이터에서 유의미한 개념(concept)을 발견하는 것을 목표로 하는 비모수적 베이지안(nonparametric Bayesian) 모델이다. 1차 술어(unary predicates)로 표현되는 개념 간 또는 요인(type) 간의 연관 관계를 발견할 수 있는 기법으로서, 모델에 설정한 모든 요인 및 연관 관계에 대한 군집화를 수행하며, 이는 기계학습 분야에서 공군집화(co-clustering)로 표현하는 기법과 일치하되, 2차 행렬 이상의 다양한 데이터 군에 대해서도 적용 가능한 보다 일반적인 모델이다.

IRM의 목적함수는 여러 요인 간의 관계 R에 대한 관측 데이터를 기초로 다음의 사후확률을 최대화하는 클러스터 조합 z를 찾는 것이다:

$$P(z|R) \propto P(R|z)P(z), \qquad (1)$$

개체 i에 대한 클러스터 할당 분포의 사전 확률(prior) $P(z)$는 식 (2)로 표현되는 중국식당 프로세스(Chinese restaurant process, CRP)[8]에 기반하여 갱신된다.

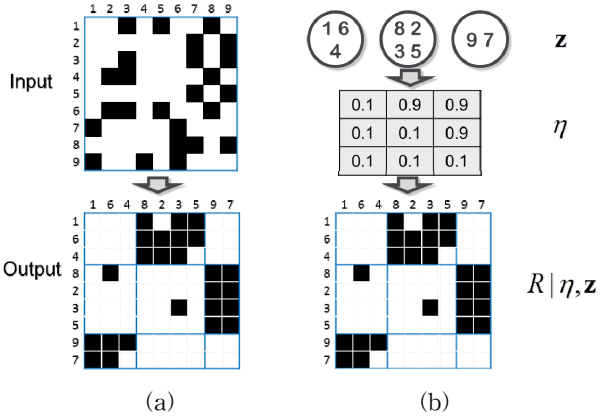


그림 2. IRM을 이용한 연관 관계 모델 구성. (a) 입력 및 출력 행렬 (b) 생성 모델로서의 IRM의 구조

Fig. 2. Constitution of relational model based on IRM. (a) input and output matrix (b) generative process of an IRM model

$$P(z_i = a|z_1,...,z_{i-1}) = \begin{cases} \frac{n_a}{i-1+\gamma} & n_a > 0 \\ \frac{\gamma}{i-1+\gamma} & a\text{는 신규 클러스터} \end{cases}, \quad (2)$$

CRP는 다음과 같이 클러스터링을 할 대상(object)이 순서대로 주어질 때 클러스터의 생성 과정을 표현하는 확률과정(stochastic process)이다. 최초 개체가 하나의 클러스터를 형성하는 과정에서 새로운 개체 i가 도입될 때마다, 기존의 클러스터의 크기(포함된 개체의 수, n_a)에 비례하여 i번째 개체가 기존 클러스터에 포함되거나, 새로운 클러스터를 형성하게 된다. 새로운 클러스터가 형성될 확률은 매개 변수 γ로 조정한다. 이론적으로 셀 수 있는 무한개의 클러스터를 생성할 수 있다.

CRP 사전 확률을 이용하여 가능한 모든 클러스터 조합에 대해 앙의 확률을 부여할 수 있으며, 적절한 우도 및 사후 확률에 대한 모델 설정을 통해, 데이터에 적합한 클러스터 수가 자동으로 결정되는데 기여한다.

기본적인 적용 사례로서 사람들(T) 간의 선호(R) 여부를 $R: T\times T\rightarrow\{0,1\}$로 표현했을 때 그림 2와 같이 군집 및 군집 간 선호 관계를 발견할 수 있다. 이 사례에서와 같이 0 또는 1의 이진 데이터를 대상으로 한 IRM 모델은 식 (3)과 같은 생성 모델로 표현된다:

$$\begin{aligned} z|\gamma &\sim \mathrm{CRP}(\gamma) \\ \eta(a,b)|\alpha,\beta &\sim \mathrm{Beta}(\alpha,\beta) \\ R(i,j)|z,\eta &\sim \mathrm{Bernoulli}(\eta(z_i,z_j)), \end{aligned} \quad (3)$$

a,b는 클러스터의 인덱스이며, 사람들 간에 선호 관계가 발생하는지 여부는 전적으로 해당 인물에 대한 클러스터 할당 내역에 의해서만 결정된다는 가정이 수반된다(세 번째 수식). $R(i,j)$는 인물 i가 인물 j를 선호함을 의미한다고 해석할 수 있다. η의 각 항에 대해 켤레 사전 확률을 적용하기 때문에 식 (4)와 같이 우도(likelihood)를 계산할 수 있다:

$$P(R|z) = \prod_{a,b} \frac{\mathrm{Beta}(m_{ab}+\alpha, \overline{m}_{ab}+\beta)}{\mathrm{Beta}(\alpha,\beta)}. \quad (4)$$

m_{ab} 및 $\overline{m}_{ab}$는 각각 클래스 a와 b 사이의 연결선에 부여된 가중치가 1, 0인 관측 횟수이다. 보통 두 요인 사이의 연관관계에 대한 대칭성 가정 하에 Beta 함수의 두 파라미터 값을 동일하게 설정하며($\alpha=\beta$), 학습 과정에서 β의 척도를 탐색하는 과정이 병행된다. 관측값이 빈도로 표현되는 경우에는 Dirichlet-multinomial 모델을, 연속값의 경우에는 Normal- Inverse-Gamma 분포를 적용한다.

개체별 클러스터 부여 변수 z가 주어진 경우 $\eta(a,b)$의 최대 사후확률(maximum *a posteriori*)은 식 (5)와 같이 계산된다:

$$\frac{m_{ab}+\beta}{\overline{m}_{ab}+m_{ab}+2\beta}. \quad (5)$$

IRM 기반의 추론은 식 (4)에 의해 계산되는 현 단계의 사후 확률에서 MCMC (Markov chain Monte Carlo)를 이용한 샘플링을 이용하거나, 사후확률의 최빈값을 탐색하여 진행할 수 있다. 최적의 클러스터 설정을 찾기 위해 한 개체의 클러스터 할당을 변경하거나, 클러스터를 나누거나 합치는 과정을 거치게 되며, 이는 그림 2(a)와 같이 입력 행렬의 행과 열을 섞어 최적의 블록을 찾는 과정으로 표현된다.

3. 공군집화 기반의 연관 관계 개념망 구성

3.1 영화 추천 기법

영화의 선호도는 개인의 속성과 영화의 속성을 기반으로 하는 다양한 요인의 연관 관계에 관련되어 있다. 개인의 속성으로는 성별, 나이, 수입과 같은 인구 통계학적 정보와 선호 키워드, 감정 상태와 같은 개성적인 특성 등이 있다. 영화의 속성으로는 감독, 주연배우, 개봉 첫 주 관객 수 등과 같은 영화 컨텐츠 외적인 메타 데이터와 영화의 장르, 시대적 배경, 배경 음악 등과 같은 컨텐츠 내적 요소 등이 있다. 이러한 요인에 의해 표현되는 선호도 정보는 일반적으로 명시적인 5점 별점 데이터로 수집되는 경우와 평점을 부여했는지, 관람을 했는지 등과 같이 비명시적인 데이터로 획득하는 경우로 크게 구분할 수 있다.

영화 추천을 위해서는 특정 개인의 취향을 파악하여 영화별 선호도를 예측하기 위한 모델이 필요하다. 데이터 기반의 자동화된 추천 시스템의 대표적 기법인 협업 필터링 방식의 경우, 부분적으로 관측된 평점 행렬 $R\in\mathbb{R}^{N_u\times N_v}$을 입력받아 관측되지 않은 (사용자 u, 영화 v) 쌍의 평점 r_{uv}을 예측하는 것을 목표로 한다.

협업 필터링 방식은 크게 메모리 기반 기법과 모델 기반 기법의 두 가지 세부 방식으로 구분된다[9]. 메모리 기반의 기법의 경우, 모든 평점 행렬이 메모리에 저장되며 영화 간 또는 사용자 간의 유사도 정보를 핵심으로 추천을 한다. 모델 기반의 협업 필터링의 경우, 평점 행렬을 입력으로 모델을 생성하고, 이 모델을 이용하여 관측되지 않은 평점을 예측한다. 클러스터링 모델[3], 베이지안 네트워크 모델[10], 강화학습 모델[11] 등의 다양한 기계학습 기반 모델이 적용된 바 있으나, 성능과 응용성 측면에서 가장 성공한 방식은

은닉요인(latent factor) 모델 기반의 접근법이다.

은닉요인 모델은 사용자와 영화를 공통의 K차원 은닉요인 공간으로 사상하여 두 요인 간의 관계를 표현한다. 은닉요인 공간에서 영화 v는 $q_v \in R^K$, 사용자 u는 $p_u \in R^K$로 표현되며, 사용자의 평점은 $\hat{r}_{uv} = q_v^T p_u$로 예측한다. 다양한 은닉요인 모델 중에서 지난 십여 년간 기법의 유연성 및 우수한 성능을 특징으로 하는 다양한 행렬 분해(matrix factorization)기법이 특히 중점적으로 개발되었다[12][13][14].

내용 기반 추천 시스템에서는 평점 데이터 외에도 사용자와 아이템 각각에 대해 추가의 프로파일을 구축하여 추천에 활용한다[15]. 아이템 프로파일은 도메인에 특화된 정보를 기반으로 아이템의 특성을 표현하며, 사용자 프로파일은 사용자의 기존의 추천 관련 행동과 인구통계학적 정보를 기초로 사용자의 선호도를 표현하는 모델이다. 사용자가 기존에 선호한 아이템과 유사한 아이템을 추천하며 특히 평점 데이터가 없는 경우에도 아이템 프로파일을 기반으로 한 추천이 가능하다.

본 논문에서는 평점 행렬 또는 간접적 선호도 정보 데이터와 함께 영화의 태그 기반 프로파일을 도입하여 공통의 사용자-영화 간 선호관계 연관성을 모델링하는 혼합 추천 기법을 적용한다. 평점 행렬에 내재된 저차원의 은닉 요인을 공군집(co-cluster)으로서 탐색하되, 해당 요인에 연관된 영화의 추가 요인의 구성 및 연관성을 함께 탐색하여 영화 선호에 대한 개념망을 구성하는 것을 목표로 한다. 이러한 목적에 IRM은 매우 적합한 특성을 가진 도구이다.

3.2 메타데이터 기반 공군집화 및 연관 관계 탐색

3.1절에서 언급하였듯이 영화 선호에 대한 개념망 구성을 목표로, 영화의 메타데이터 기반 프로파일을 추가로 도입하여 공군집화를 수행하고, 선호도 관련 연관 관계를 종합적으로 탐색한다. 사용자, 영화 및 영화의 메타데이터 세 가지 요인을 기반으로 다음과 같이 연관 관계 모델을 설정하였다. R_1은 사용자의 영화에 대한 선호 관계를 이진 계수로 표현한 것이며, R_2는 태그 기반의 메타 데이터가 지정 영화에 포함되는지 여부를 표현한다.

$$R_1 : T_1 \times T_2 \rightarrow \{0,1\} \text{ (선호 관계)},$$
$$R_2 : T_2 \times T_3 \rightarrow \{0,1\} \text{ (포함 관계)}$$
(T_1: 사용자, T_2: 영화, T_3: 영화의 메타데이터)

영화의 메타데이터로는 영화의 다양한 속성에 대한 체계적 태그 기반 시스템으로서 지니닷컴의 Entertainment Genome (EG)[16]을 사용한다. EG는 영화의 속성을 12가지 genome (태그의 유형에 해당), 이천여 개의 gene (태그에 해당)으로 표현하며, 지니닷컴에서 제공하는 비디오 컨텐츠 추천 서비스의 기반이 되는 핵심 분류체계(taxonomy)이다.

선호도 데이터와 EG 정보를 근간으로 R_1, R_2에 대한 IRM 기반 모델링 결과의 사례를 그림 3에서 살펴볼 수 있다. 공군집화를 통해 발견한 사용자, 영화 및 EG의 군집을 추가로 기재하였다. 예시한 학습 결과에서 사용자 군집 u3는 영화 그룹 m2를 선호하며, m2에는 g6 그룹의 gene이 집중적으로 포함되어 있는 것을 파악할 수 있다. 이와 같이 군집 단위에서 사용자, 영화, 메타데이터 간에 군집 단계의 보다 개념적인 연관 관계를 발견할 수 있다.

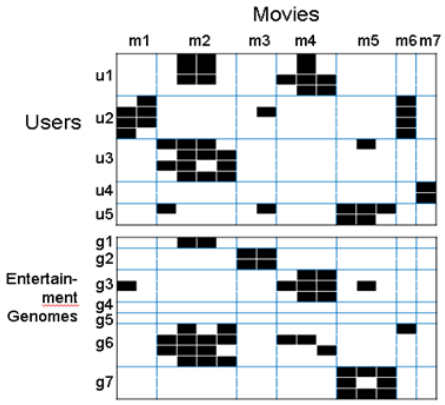


그림 3. 사용자-영화 평점 행렬 및 영화-EG기반 메타데이터에 IRM을 적용하여 얻을 수 있는 공군집화 결과 예시.

Fig. 3. An example result of co-clustering by IRM on user-movie rating matrix and Entertainment-genome-based metadata for movies.

4. 공군집화 실험 및 연관 관계 탐색

4.1 실험 구성

IRM을 이용한 공군집화를 통해 영화 선호에 대한 연관 관계를 발견하기 위하여 다음과 같이 실험을 구성하였다.

영화의 평점 정보로는 대표적인 영화 추천 평가용 데이터 중 하나인 MovieLens 100k 데이터셋을 사용하였다. MovieLens 100k 데이터셋은 943명 사용자의 1682개 영화에 대한 10만 개의 1~5점 별점 사례로 구성되어 있다. 1682개의 영화 중 메타데이터로서 EG를 추출할 수 있는 1601개의 영화를 선별하였다. 평점 정보는 명시적 정보와 비명시적 정보의 두 가지 형태로 구성하였다. 명시적 정보로는 MovieLens 데이터의 평점 정보를 그대로 적용하며, 비명시적 정보로는 영화에 평점을 부여하는 행위 자체가 선호도의 반영이라는 가정 하에, 평점이 부여된 모든 관측값을 1로 설정한 행렬을 구성하였다. 영화별 EG 데이터는 지니닷컴에서 온라인 검색을 수행하고 영화별 gene을 수집하여 구성하였다. 수집한 전체 EG의 수는 874개이다. EG의 각 gene에 부여된 상위 분류체계로서의 genome 정보는 이 실험에서는 직접 활용하지 않았다. 영화별 EG 데이터는 모두 52,726개의 영화별 태그로 구성되었다.

IRM 수행 코드는 [7]의 저자가 공개한 C언어 코드를 활용하였다[17]. 추론 알고리즘은 경사면 등반(hill-climbing)을 적용하고, 하이퍼파라미터가 학습 과정 중 갱신되도록 설정하였다. 전체 모델의 평가 점수는 각 요소별 현재의 클러스터 구성에 대한 로그 확률값의 총합으로 표현하며, 부분 극점 도달시 랜덤 초기화를 통해 탐색을 반복하고, 여러 극점 중 점수가 최대인 모델을 선택한다. 이 모델을 분석하여 영화 추천 데이터의 이면에 담긴 고차 연관 관계를 탐색한다.

4.2 실험 결과 및 분석

사용자, 영화 및 영화 태그로서의 EG, 세 가지 요인의 두 가지 연관 관계 R_1, R_2에 대한 IRM의 수행 결과 최적의 모델에서 발견한 군집의 수는 표 1과 같다. 탐색 공간의 복잡

표 1. MovieLens 100k 및 영화 메타데이터로서 지니닷컴의 EG에 IRM 기반 공군집화 적용 결과 (score=-206457.754)
Table 1. IRM-based co-clustering results on MovieLens 100k set and movie tag set based on Entertainment Genome by jinni.com (score=-206457.754)

Factors (types)	Size (# inst.)	Number of Clusters	Cluster Size (min/max/avr)
User (T_1)	943	24	1/286/39.3
Movie (T_2)	1601	27	9/158/59.3
EG (T_3)	874	110	1/212/7.9

성으로 인해 반복학습 회수를 늘릴 경우 보다 좋은 해를 발견할 가능성이 있지만, 연관 관계의 '발견' 측면에서는 충분히 활용 가능한 결과이다. 사용자 군집과 EG 군집 중에는 단일 원소 클러스터도 포함되어 있으며, 세 요인 공히 매우 거대한 규모의 클러스터도 발견되었다. 요인별 클러스터 크기의 분포를 그림 4에 정리하였다.

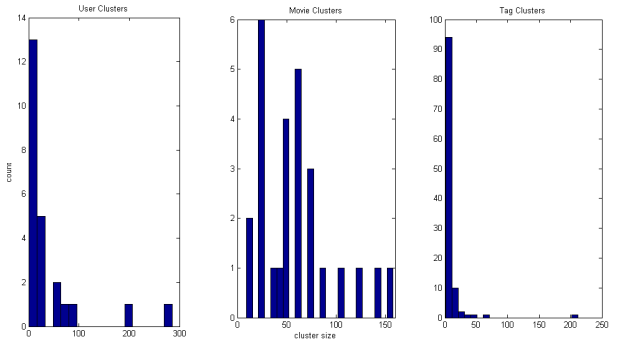


그림 4. 표 1에 정리한 결과 모델에서 사용자, 영화, 영화 태그별 군집의 크기 분포
Fig. 4. Distributions of the cluster sizes on users, movies, and movie tags, extracted from the learned model in table 1

학습 결과 생성된 공군집 중에서 고밀도의 블록($\eta_{\text{rating}} > 0.6$, $\eta_{\text{movie-EG}} > 0.9$)을 기반으로 발견한 선호 영화에 관련된 연관 관계 사례를 그림 5에 정리하였다. 303번 사용자는 헐리우드풍의 블록버스터 영화를 선호하는 유사한 사용자들과 11번 사용자 그룹을 형성하는 것을 확인할 수 있다. 비슷한 맥락에서, 사용자의 비선호 패턴에 대한 EG 관점에서의 설명이 가능하다.

한편, 비명시적 정보로 해석한 평점 정보를 적용하여 실험을 수행하고 그 중 점수가 가장 높은 모델을 분석한 결과 사용자 및 영화 클러스터의 수는 각각 84, 51개로 명시적 평점을 적용한 경우보다 많고 균일하였으며, 반면 태그 클러스터의 수는 53개로서 명시적 경우보다 대체로 큰 단위의 클러스터가 형성되었다. 영화 선호도를 태그 중심으로 설명하려 할 때, 이진화된 선호 여부에 대해 보다 균일한 개념 집합을 통해 설명할 수 있을 것이라고 해석할 수 있다.

표 2에 명시적 평점 및 비명시적, 이진화된 선호 정보에서 추출한 선호 및 비선호 관련 연관 관계의 예를 정리하였다. 영화 클러스터를 중심으로 한 연관 관계에 초점을 두었으며, EG 클러스터에 포함된 태그를 기초로 연관 관계에 대해 분석한 결과를 추가로 기재하였다.

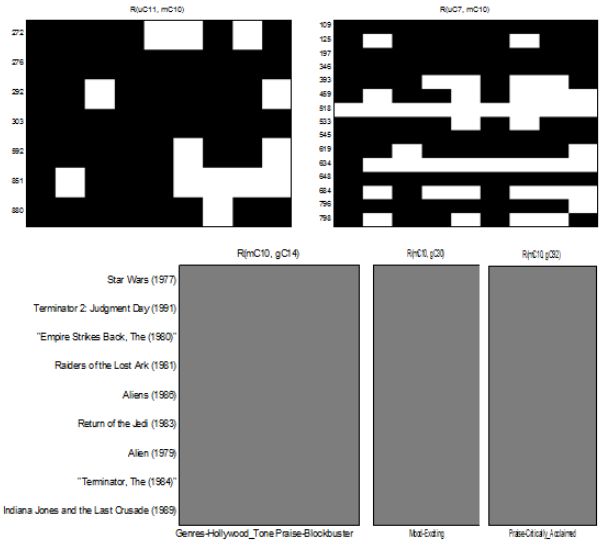


그림 5. 사용자그룹 7, 11은 영화그룹 10을 선호하며, 이 영화그룹은 EG의 14, 30, 92번 그룹의 특성(헐리우드풍, 흥분 무드, 주요 영화제 수상후보)을 강하게 가지고 있다
Fig. 5. User groups 7 and 11 show preferences on movie group 10, which has features of EG (tag) group 14, 30, and 92 (Hollywood tone, blockbuster, exciting mood, critically acclaimed)

표 2. MovieLens 100k set 및 Entertainment Genome 정보를 입력으로 IRM을 이용한 공군집화 수행 결과 발견한 요인의 클러스터 간 연관 관계 예시. 평점 정보를 명시적 정보로 그대로 활용한 경우와 비명시적 정보로 해석한 경우의 두 사례로 구분된다.
Table 2. Samples of discovered relations among clusters for users, movies, and movie tags (EG) after co-clustering by IRM with MovieLens 100k and Entertainment Genome tag data as input

Pattern Type	User Cluster ID	Movie Cluster ID	EG Cluster ID	Description of Co-clusters	Movie Titles
Like	7, 11	10	14, 30, 92	A group of users who like popular Hollywood SF from 70s~90s	Star Wars, Terminator
Dislike	3, 16	5	88	A group of users who do not like horror movies	Nadia, I Know What You Did Last Summer
Dislike	3, 4	7, 14	3	A group of users who do not like humorous or comic movies	Billy Madison, Nutty Prof.
Dislike	3, 15	20	55, 110	Users who dislike realistic or semi-serious movies	Wedding Bell Blues, French Twist
Implicit-Rating	21	27	33	A group who show interest on movies on friendship	Grand Day Out, Close Shave
Implicit-Rating	5,14	21	15, 26, 46	These people are interested in Oscar-awarded movies with 20th America as their setting	Godfather 1, FARGO

5. 연관 관계 개념망 기반 추천 방안

4절에서 IRM을 이용한 공군집화를 통해 영화의 선호도에 내재된 개념 수준의 다양한 연관 관계를 발견할 수 있음을 확인하였다. 이 절에서는 이러한 결과를 활용한 추천 방안에 대해 논의한다.

IRM 활용시 비모수적 베이지안 기법의 공통적 특성인 '긴 학습 및 추론 시간'을 고려해야만 한다. 학습 데이터의

크기가 대략 100k(평점)+50k(태그)인 4절 실험의 경우, 리눅스-AMD Opteron 2.6GHz(8코어)-램 256G 서버에서 24~27시간 정도 소요가 되었다. 추천 시스템의 여러 단계에서 IRM 및 관련 기법을 온라인 단계 작업의 요소로 포함시키는 것은 현실적이지 않으며, 오프라인 단계에서 추천에 필요한 핵심 정보의 '발견'을 위한 도구로서의 활용이 의미가 있을 것이다. 여러 요인의 연관 관계 개념망이 추천 시스템의 패러다임 별로 어떻게 활용 가능한지 정리해보자.

가장 직접적인 활용 방법은 공군집화 기반의 협업 필터링 방식 추천 기법[3][4][18]과 동일한 전략을 적용하는 것이다. 즉, IRM 을 공군집화 탐색의 도구로 적용하고, 그 결과 발견한 사용자 및 아이템 군집 별로 별도의 모델을 구성한 후 추천시 모델을 통합하는 방식의 활용이 가능하다.

내용 기반 추천 방식 측면에서는 아이템 군집별 부분공간 표현(subspace representation) 및 사용자 군집 단위의 프로파일링 구축에 적용할 수 있다. 온라인 상에서는 각 추천 대상 아이템 또는 사용자에 대해 가장 유사한 군집을 탐색하거나, 각 군집에 대한 소속 가능 확률을 계산한 후, 해당 군집별 선호 패턴 모델을 취합하는 추천을 수행한다. 이와 관련하여 협업-필터링과 내용 기반 방식의 혼합 추천 기법 성과를 참고할 수 있다[5][6][19].

지식 기반(knowledge-based) 추천 시스템의 경우, 활용 가능한 다양한 요인에 대한 데이터를 수집하고 IRM을 적용하여, 사용자-아이템 간 연간 관계에 대한 지식베이스를 구축하거나 제한조건을 설정하는 데에 적용 가능하다.

비명시적 피드백 기반 추천의 경우 IRM을 통해 범용(generic) 사용자 모델을 구현할 수 있다. 공군집을 전체 사용자 공간을 포괄하는 유형 구분 단위(subtyping set)로서 설정하고, 비명시적 피드백 패턴에서 각 공군집으로의 사상함수(mapping function)를 학습하는 것이다. 이는 간접적 피드백 정보에서 선호도를 추론하기 위한 기저 함수 추출의 관점으로도 볼 수 있다.

한편, 보다 고도화된 연관 관계 개념망 학습 기법으로서 IRM 기법을 제안한 저자의 후속 연구에 주목할 만하다. [20]에서는 다중 피쳐 요인 간의 군집 내 분화 및 군집 구조 단위의 온톨로지 구축 관점에서 IRM의 확장된 모델이 소개되었다. [21]에서는 IRM에서와 같이 개체, 특성값, 연관 관계를 기초 재료로 개념 단계에 대한 모델을 구축하되 술어논리(predicate logic)을 통해 다양한 도메인에서 공통적으로 적용 가능한 모델을 구축할 수 있다는 제안을 한다. [22]에서는 기 학습한 모델을 기초로 새로운 분류(군집)과 개체를 생성하는 기법을 인지심리학 관점에서 정리하였다. 이러한 접근법은 추천에서의 콜드스타트(cold-start) 문제 해법으로서 활용 가능성이 크다.

6. 결론

본 논문에서는 비모수적 베이지안 기반의 Infinite Relational Model (IRM)을 이용한 공군집화를 영화 추천 분야에 적용하여, 사용자의 영화에 대한 선호 관계를 파악하고, 이를 추천에 활용하는 방안을 다루었다. 구체적인 예시로서 영화에 대해 명시적으로 선호도를 표현한 평점 데이터 또는 비명시적 선호도 데이터를 근간으로 하고, 영화의 태그 형태의 메타데이터로서 Entertainment Genome (EG)을 도입하여, 영화 선호 요인에 대한 설명 및 사용자 그룹의 특성을 발견한 결과를 정리하였다.

IRM 기반 공군집화는 특히 최신 추천 시스템 연구의 핵심 도전 과제인 '간접적 피드백 정보에서 사용자의 선호도를 예측' 문제에서, 연관된 다양한 보조 정보를 활용하여 접근할 수 있는 기법으로서 주목할만한다. 후속 연구에서는 영화의 태그 정보 이외에도 사용자의 문맥과 연관된 데이터를 추가로 활용하고, 연관 관계에 대한 보다 심도 있는 탐색과 분석에 주안점을 두고자 한다.

References

[1] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: An Introduction*, Cambridge University Press, New York, NY, USA, 2010.

[2] L. H. Ungar, D. P. Foster, E. Andre, S. Wars, F. S. Wars, D. S. Wars, and J. H. Whispers, "Clustering Methods for Collaborative Filtering," in *AAAI Workshop on Recommendation Systems*, 1998, pp. 114 - 129.

[3] B. Xu, J. Bu, C. Chen, and D. Cai, "An Exploration of Improving Collaborative Recommender Systems via User-item Subgroups," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 21 - 30.

[4] N. Mirbakhsh and C. X. Ling, "Clustering-based factorized collaborative filtering," in *Proceedings of the 7th ACM conference on Recommender systems - RecSys '13*, 2013, pp. 315 - 318.

[5] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, 2011, pp. 448 - 456.

[6] D. Agarwal and B.-C. Chen, "fLDA: matrix factorization through latent Dirichlet allocation," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 2010, pp. 91 - 100.

[7] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning Systems of Concepts with an Infinite Relational Model," in *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006, pp. 381 - 388.

[8] J. Pitman, *Combinatorial Stochastic Processes,* Springer-Verlag, Berlin, 2006.

[9] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 1998, pp. 43 - 52.

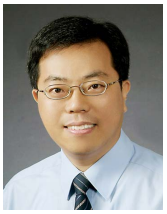
[10] D. Heckerman, D. Chickering, C. Meek, R. Rounthwaite, and C. Kadie, "Dependency networks for inference, collaborative filtering, and data visualization," *Journal of Machine Learning Research*,

vol. 1, pp. 49 - 75, 2001.
[11] B.-T. Zhang and Y.-W. Seo, "Personalized web-document filtering using reinforcement learning," *Applied Artificial Intelligence*, vol. 15, pp. 665 - 685, 2001.
[12] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of dimensionality reduction in recommender system-a case study," *ACM WebKDD 2000 Web Min. ECommerce Work.*, vol. 1625, pp. 264 - 8, 2000.
[13] R. Salakhutdinov and A. Mnih, "Probabilistic Matrix Factorization," in *Advances in Neural Information Processing Systems 20*, 2008, pp. 1257 - 1264.
[14] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30 - 37, 2009.
[15] P. Lops, M. De Gemmis, and G. Semeraro, Content-based Recommender Systems: State of the Art and Trends, In Recommender Systems Handbook (pp. 73 - 105), 2011.
[16] Entertainment Genome: http://www.jinni.com/info/entertainment-genome.html
[17] IRM Code: http://www.psy.cmu.edu/~ckemp/code/irm.html
[18] J. Lee, S. Kim, G. Lebanon, and Y. Singer, "Local Low-Rank Matrix Approximation," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, vol. 82 - 90, pp. 82 - 90.
[19] A. Popescul, L. H. Ungar, D. M. Pennock, and S. Lawrence, "Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 2001, pp. 437 - 444.
[20] C. Kemp, A. Perfors, and J. B. Tenenbaum, "Learning overhypotheses with hierarchical Bayesian models," *Developmental science*, vol. 10, no. 3, pp. 307 - 21, May 2007.
[21] C. Kemp, "Exploring the conceptual universe," *Psychological Review*, vol. 119, no. 4, pp. 685 - 722, Oct. 2012.
[22] A. Jern and C. Kemp, "A probabilistic account of exemplar and category generation," *Cognitive. Psychology*, vol. 66, no. 1, pp. 85 - 125, Feb. 2013.

저 자 소 개

김병희(Byoung-Hee Kim)
2003년 : 서울대학교 컴퓨터공학부 공학사
2006년 : 서울대학교 컴퓨터공학부 박사과정 수료
2006년: 독일 베를린공대 방문연구원
2006년～현재 : 서울대학교 컴퓨터공학부 연구원

관심분야 : Machine Learning, Artificial Intelligence, Probabilistic Graphical Models
Phone : +82-2-880-1847
E-mail : bhkim@bi.snu.ac.kr



장병탁(Byoung-Tak Zhang)
1986년 : 서울대학교 컴퓨터공학과 공학사
1988년: 서울대학교 컴퓨터공학과 공학석사
1992년: 독일 Bonn 대학교 컴퓨터과학 박사
1992년~1995년: 독일국립정보기술연구소(GMD, 현 Fraunhofer Institutes) 연구원
1997년~현재 : 서울대학교 컴퓨터공학부 교수 및 인지과학, 뇌과학, 생물정보학 협동과정 겸임교수
2003년~2004년: MIT 인공지능연구소(CSAIL) 및 뇌인지과학과(BCS) 객원교수
2012년~현재: 서울대학교 인지과학연구소 소장

관심분야 : Biointelligence, Cognitive Machine Learning, Molecular Evolutionary Computation-based Neurocognitive Information Modeling
Phone : +82-2-880-1833
E-mail : btzhang@bi.snu.ac.kr