ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

잠재적 속성 선호도를 이용한 협업 필터링의 데이터 희소성 문제 개선 방법

Method to Improve Data Sparsity Problem of Collaborative Filtering Using Latent Attribute Preference

권 형 준[1*] 홍 광 석[1]
Hyeong-Joon Kwon Kwang-Seok Hong

요 약

본 논문에서는 협업 필터링의 선호도 예측 정확성의 저하를 초래하는 전통적 문제점 중 하나인 데이터 희소성 문제에 강인한 잠재적 속성 선호도 기반 협업 필터링 방법(Latent Attribute Rating-based Collaborative Filtering, LAR_CF)을 제안한다. 기존의 협업 필터링은 객체의 유사성을 판단하기 위한 특징벡터로써 사용자가 명시적으로 평가한 선호도만을 이용하며, 해당 문제 개선을 위해 속성을 사용하는 연구들은 범용적으로 사용하기 어려웠다. 이웃 기반 필터링에 근본을 두는 LAR_CF는 기존의 명시적 선호도와 함께 유사도 평가의 대상이 되는 두 객체의 고유한 속성을 특징벡터로 삼기 때문에 명시적 선호도의 수가 적어서 발생하는 데이터 희소성 문제를 개선하여 선호도 예측 정확도를 향상시키며, 속성의 종류에 구애받지 않고 손쉽게 적용할 수 있는 장점을 가진다. LAR_CF의 유효성 평가를 위해서 MovieLens 100k 데이터세트 및 해당 데이터세트에 사용된 속성정보를 활용하여 일반적 성능 실험과 인공적 데이터 희소성 실험에서 선호도 예측 정확도를 평가한 결과, 제안하는 방법이 데이터 희소 조건에서 선호도 예측 정확도를 향상시킬 수 있음을 확인하였다.

주제어 : 협업 필터링, 속성 선호도, 추천 시스템

ABSTRACT

In this paper, we propose the LAR_CF, latent attribute rating-based collaborative filtering, that is robust to data sparsity problem which is one of traditional problems caused of decreasing rating prediction accuracy. As compared with that existing collaborative filtering method uses a preference rating rated by users as feature vector to calculate similarity between objects, the proposed method improves data sparsity problem using unique attributes of two target objects with existing explicit preference. We consider MovieLens 100k dataset and its item attributes to evaluate the LAR_CF. As a result of artificial data sparsity and full-rating experiments, we confirmed that rating prediction accuracy can be improved rating prediction accuracy in data sparsity condition by the LAR_CF.

☞ keyword : Collaborative Filtering, Attribute Preference, Recommender System

1. 서 론

정보통신 기술의 발전은 정보 제공자와 소비자의 경계를 완화하여 온라인상에서 공유되는 정보의 양을 급속히 증가시켰다. 이러한 변화는 사용자가 선택할 수 있는 기회를 높이는 장점을 제공하고 있지만, 사용자 개인이 선호하는 정보를 찾기 위해 필요한 시간과 비용을 증가시키고 있으며 이에 따라 사용자 개인의 성향에 맞는 정보를 자동적으로 선별하여 제공하는 추천 시스템(Recommender Systems, RS)에 관한 연구가 활발하다[1]. 특히, 다수의 지식정보를 이용한 정보 제공의 개인화 개념 중 하나인 집단지성(Collective Intelligence)의 이론을 가장 잘 반영한 것으로 알려진 협업 필터링(Collaborative Filtering, CF) 기술은 오늘날 RS를 구축하기 위한 최적의 방법으로 채택되고 있다[2].

사용자 및 사용자가 선호도를 평가하기 위한 대상을 통칭하는 아이템 중 하나를 객체로 삼아 객체들 사이의 유사성을 이용하는 이웃 기반 협업 필터링(Neighbor-based Collaborative Filtering, NBCF)은 사용자 자신이 과거에 경험했던 아이템에 대하여 주관적인 관점에서 명시적으로 평가한 선호도 점수를 이용하여 사용자가 경험하

1 School of Information and Communication Engineering, Sungkyunkwan University, Suwon, 440-746, South Korea
* Corresponding Author (katsyuki@skku.edu)
[Received 25 January 2013, Reviewed 8 March 2013(R2 5 June 2013), Accepted 19 July 2013]

지 못한 아이템에 대한 선호도를 예측하는 기술이다[3]. 추천 대상 객체와 유사한 성질을 가진 객체의 집단을 발견하기 위해서 추천 대상 객체와 유사성 판단 대상 객체 사이의 특징벡터를 이용하여 유사성을 판별하는데, 두 객체 사이에 공통적으로 존재하는 명시적 선호 점수의 개수가 충분하지 못하면 객체들 사이의 유사성 도출이 불가능해지며, 도출되더라도 그 유사성을 신뢰하기 어려워진다. 즉, 특징벡터의 차수가 적을수록 객체 사이의 유사성을 발견하기 어렵게 되기 때문에 선호도 예측 정확도가 하락하는 현상을 보인다. 이것을 CF의 선호도 예측 정확도를 저해하는 주된 문제 중 하나인 데이터 희소성(Data Sparsity) 문제라 한다[3, 4]. 기존의 연구들은 본 문제를 NBCF로 극복하기 어려운 문제로 규정짓고 이를 해결하기 위해서 인공지능, 패턴인식, 통계적 기계 학습 기법을 사용하는 모델 기반 협업 필터링(Model-based Collaborative Filtering, MBCF) 방법에 관해 연구하고 있지만 NBCF가 가진 실시간 예측 시의 장점을 포기해야 하기 때문에 NBCF의 장점을 유지하면서도 데이터 희소성 문제에도 강인한 선호도 예측 방법에 관한 연구가 요구되었다[5].

이를 위한 방법으로 사용자 또는 아이템의 속성을 이용하는 다양한 접근법이 연구되고 있다[6-12]. 속성의 정보는 사용자가 직접 평가하는 선호도와는 다르게 비정량 및 비정형 데이터이기 때문에 이를 처리하기 위해 정형화시키는 방법들이 다양하게 제시되었다. 그러나 협업 필터링의 적용 분야에 따라 사용자 또는 아이템의 속성이 다양해지는 만큼 선호도 예측에 반영하기 위한 방법도 속성에 따라 특화되어 다르게 연구되어 왔기 때문에 속성의 종류를 막론하고 범용적으로 사용할 수 있는 방법이 요구된다.

이에 본 논문에서는 NBCF의 데이터 희소성 문제 개선을 위해 객체의 속성을 이용하되 속성의 종류에 구애받지 않고 어떠한 속성에도 적용할 수 있는 일반화된 방법으로 잠재적 속성 기반 협업 필터링 방법(Latent Attribute-based Collaborative Filtering, LAR_CF)을 제안한다. 각 사용자 또는 아이템의 속성을 바탕으로 아이템 속성에 대한 사용자의 선호도 또는 사용자 속성에 대한 아이템의 선호도를 추출하고 특징벡터로 삼는다. 기존의 선호도 특징벡터와의 선형적 결합을 통해 특징벡터의 차수를 증가시켜서 이를 사용자 사이의 유사성 평가에 사용하여 특징벡터의 부족에 기인하는 데이터 희소성 문제를 개선함과 동시에, 속성의 특성에 상관없이 다양한 분야에서 활용되는 협업 필터링에 범용적으로 사용할 수 있도록 고안되었다.

논문의 나머지 구성은 다음과 같다. 2장에서 CF에 대한 개념과 기존 연구를 살펴보고, 3장에서 잠재적 속성 기반 협업 필터링 방법에 관해 설명한다. 4장에서는 제안 방법의 성능 측정 실험 및 결과를 소개하고 5장에서 결론을 맺는다.

2. 관련 연구

방법론적으로 크게 NBCF 및 MBCF 방법으로 구분되는 협업 필터링은 특정 사용자가 아직 경험하지 않은 아이템에 대한 선호도를 다른 사용자의 선호도를 참조하여 예측하는 것으로, 오늘날 RS의 구축에 최우선적으로 고려되는 기술임을 이미 설명하였다. 잘 알려진 소프트컴퓨팅 알고리즘을 CF에 적용하는 MBCF는 초기에 나이브 베이즈 분류(Naive Bayes Classification)와 같은 다소 기초적인 확률 모델이 적용되었으나, 최근에는 지지 벡터 머신(Support Vector Machine, SVM), 주성분 분석(Principal Component Analysis, PCA), 인공 신경망(Artificial Neural Network, ANN), K-평균 클러스터링(K-means Clustering) 등 전통적인 인공지능, 패턴인식 및 확률적 알고리즘들이 적용되어 왔다[2].

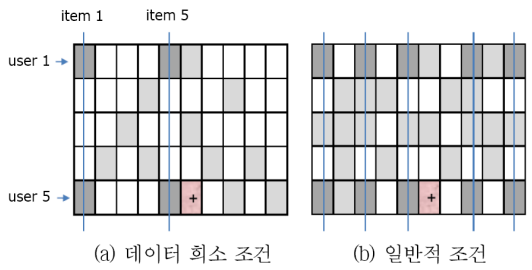
본 논문에서 제안하는 LAR_CF의 근간인 NBCF는 세부적으로 사용자 또는 아이템 중 어느 객체를 기준으로 선호도 예측을 수행할 것인지에 따라서 사용자 기반 및 아이템 기반 방법으로 나뉜다. NBCF가 사용자의 선호도를 예측하는 절차를 단계적으로 구분하면 세 단계로 구분할 수 있는데, 사용자 기반 방법과 아이템 기반 방법의 전반적인 프로세스는 동일하며 단지 유사도 측정 대상이 사용자인지 또는 아이템인지에 관한 것만 차이가 있다. 본 논문의 전체적인 부분에서 사용자 및 아이템 모두를 지칭할 때는 객체로 표현하고자 하며, 아래의 절차는 사용자 기반 방법을 기준으로 설명하였다.

첫 번째 절차는, 예측 대상 사용자가 아직 선호도 평가를 수행하지 않은 모든 아이템들을 예측 대상 아이템으로 선정하고, 예측 대상 사용자를 제외한 다른 모든 사용자 중에서 예측 대상 아이템에 선호도 평가를 수행한 사용자들의 목록을 만든다. 두 번째 절차는, 만들어진 목록에 포함된 모든 이웃과 예측 대상 사용자가 공통적으로 평가한 아이템 목록을 만들고, 해당 아이템 목록에 대응되는 선호도 목록을 특징벡터로 활용하여 예측 대상 사용자와 다른 사용자 사이의 유사도를 산출한다. 이 때 고려될 수 있는 유사성 척도는 공통 아이템의 선호도 목록이 선형적 관계에 있으므로 피어슨 상관계수(Pearson

Correlation Coefficient), 스피어만(Spearman) 상관계수 및 코사인 계수(Cosine Coefficient) 등을 사용할 수 있다[2, 5]. 피어슨 상관계수와 코사인 계수는 특징이 정규분포를 따른다고 가정할 수 있는 모수적 특성을 따를 때, 스피어만 상관계수는 비모수적일 때 고려되는 유사성 척도이다. 세 번째 절차는, 예측 대상 사용자와 유사도가 높은 상위 *k*개의 객체인 최근접 이웃과 해당 유사도 및 그들이 예측 대상 아이템에 평가한 선호도를 이용하여 가중 평균 또는 편차의 가중 평균으로 예측 대상 아이템에 대한 예측 대상 사용자의 선호도를 예측한다.

상술된 NBCF의 성능을 저해하는 주된 문제점은 연산의 확장성(Scalability), 데이터의 희소성(Sparsity) 및 콜드-스타트(Cold-start) 문제에 기인한다[6]. 확장성 문제는 NBCF에서 사용자의 수나 아이템의 수가 증가할수록 예측에 필요한 계산량이 지수적으로 증가하여 예측 시간이 오래 걸리는 문제이다. 이는 NBCF의 태생적인 문제이므로 개선하기가 매우 어렵기 때문에 MBCF가 대안으로써 제시된 것이다. 그러나 MBCF는 실시간으로 축적되는 사용자의 선호도를 즉각적으로 활용하기 어렵고, 기존에 잘 알려진 알고리즘을 사용하기 때문에 부가적인 정보를 적용하는 것 외에는 특별한 성능 개선 방법이 제시되지 못했다. 때문에 실세계 적용 관점에서는 NBCF가 더 정확한 예측 정확도를 보이게 된다[7]. 즉, 모델 생성 시간을 제외한 예측 속도 측면에서는 MBCF가 우위에 있으며, 예측 정확도 및 실세계 적용 측면에서는 NBCF가 우위에 있다.

제안방법인 LAR_CF가 개선하려는 데이터 희소성 문제는 사용자 사이의 유사도 계산 시에 공통 선호도의 개수가 적을 때 유사도 측정이 정확히 되지 않아서 선호도 예측 성능이 저하되는 현상을 지칭하며, 콜드-스타트 문제는 사용자 사이의 유사도 계산 시에 공통 선호도의 개수가 적을 때 유사도 측정이 정확히 되지 않는 문제와 함께 데이터베이스에 신규 진입한 사용자 또는 아이템에 대해 추천 프로세스가 동작할 수 없는 현상을 말한다. 콜드-스타트 문제가 선호도 개수가 부족하여 생기는 문제를 포함한다는 점이 데이터 희소성 문제와 동일한 맥락에 있다. 더욱 원론적인 면에서 두 문제를 구분하면 데이터 희소성 문제는 데이터베이스에 포함된 선호도 데이터 양의 관점에서 보는 것이고 콜드-스타트 문제는 신규 사용자 또는 아이템의 관점에서 본 것이다. 다수의 사용자 중 하나의 사용자만 선호도 데이터가 매우 적은 경우라면 데이터베이스의 희소성 문제는 발생하지 않을지라도 하나의 사용자에 대해서는 콜드-스타트 문제가 발생하게 되는 것이다.



(그림 1) 데이터 희소성 문제의 시각화
(Figure 1) Visualization of Data Sparsity Problem

사용자가 아이템에 선호도 평가를 수행하지 않아서 생기는 빈 공간을 임의의 수로 채울 수는 없기 때문에 적은 수의 특징벡터로도 안정된 예측 정확도를 보일 수 있는 방법이 연구되고 있다[8-14]. (그림 1)에 공통 선호도의 의미와 데이터 희소성 문제가 발생하는 경우를 단편적으로 보였다. 가로축은 아이템, 세로축은 사용자를 의미하며, 음영으로 표기된 부분은 선호도가 존재하는 것이며, 진한 음영은 공통 선호도에 해당되는 선호도이다. 음영으로 표기되지 않은 부분은 선호도가 존재하지 않음을 나타낸다. “+” 로 표기된 위치의 선호도를 예측하기 위해서 사용자 5와 사용자 1의 유사도를 측정할 때 두 사용자 사이의 공통 선호도를 이용하는데 (그림 1) (a)의 경우는 공통 선호도의 개수가 2개가 된다. 공통 선호도가 5개인 (그림 1) (b)의 조건과 비교하면 상대적으로 유사성이 잘 드러나지 않게 되어 데이터 희소성 문제에 의한 예측 정확성의 하락 현상이 발생한다.

NBCF의 데이터 희소성 문제를 해결하기 위한 주된 기술적 접근 방법은 유사성 척도에 관한 연구 및 사용자와 아이템의 유사도를 모두 이용하는 하이브리드 형태의 연구로써, 최근의 연구 중 하나는 PIP(Proximity - Impact - Popularity) 유사성 척도이다[8]. 기존의 유사성 척도가 유사도 측정 대상 객체들 사이의 특징벡터 차수가 적을 때 유발시키는 여러 가지 산술적 문제점을 해결하여 선호도 예측 정확도를 향상시키기 위한 목적으로 제안되었다. 이외에도 선형 회귀 기법을 응용한 Slope One 알고리즘[9], 아이템 기반 방법과 사용자 기반 방법을 결합한 하이브리드 접근법에 관한 연구[10], 반복적 유사도 갱신을 통한 예측 결과의 조정[11], 다양한 유사성 척도의 융합을 통한 NBCF에 관한 연구[12] 및 행렬 인수분해에 기반하는 양방향 유사도[13] 등이 제안되었다.

유사성 척도의 개발 이외의 다른 기술적 접근 방법 중 하나는 예측 알고리즘의 개발이다. 그래프 이론을 적용한

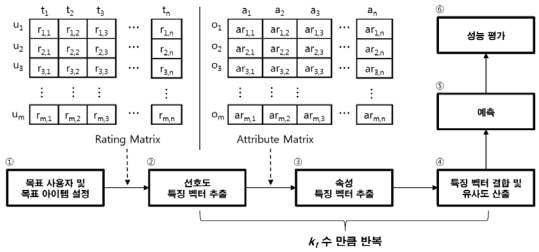
예측 알고리즘인 [14]는 예측 대상 사용자와 다른 모든 사용자 사이의 유사성 측정을 통해 선정된 상위 k개에 해당하는 이웃들을 깊이 0으로 삼고, 이웃의 이웃을 깊이 1, 다음 차수의 이웃을 깊이 2로 삼는 방법을 적용하여 예측 대상 사용자와 직접적인 이웃은 아니더라도 각 최근접 이웃들과 유사한 깊이를 파라미터로 삼는 간접적 이웃 선정 방법을 제시하였다.

기술적 문제 이외의 접근 방법으로는 본 논문에서 다루는 것과 같이 사용자 또는 아이템의 속성을 이용하는 방법이 있다[15, 16]. 성별, 나이, 직업 등의 사용자 속성과 장르, 아이템에 대한 경험자 수 및 직업 등의 아이템 속성은 사용자의 주관적 선호도 평가 점수가 정량적 데이터인 것과 달리 비정형 데이터에 속하기 때문에 속성 정보를 이용하여 성능을 개선하려는 최종 목표 아래에서 어떠한 속성정보를 어떻게 정형화 시키는지에 관한 내용을 다루었다[17, 18]. 그러나 기존의 연구들은 실험용 데이터세트에 적합하게끔 일부의 특정 속성들만을 정형화하여 활용하기 위한 방법에 관해 다루고 있어서 손쉽게 적용하기 어렵거나 널리 사용되기 힘든 문제가 존재한다. 즉, 특정 데이터세트에 최적화되어 개발된 방법이어서 다양한 데이터세트에 적용되기 어렵다. 본 논문에서 제안하는 LAR_CF는 속성을 이용하여 데이터 희소성 문제를 해결함과 동시에, 데이터세트가 포함하는 속성의 종류에 상관없이 범용적으로 활용될 수 있는 방법이어서 기존의 연구보다 확장 및 활용이 용이하다.

3. 잠재적 속성 기반 협업 필터링 방법

LAR_CF는 사용자 또는 아이템의 속성을 이용하려는 것으로 기존의 연구에서도 객체의 속성을 이용하는 방법을 시도한 바 있음을 설명하였다[15-19]. 그러나 기존의 연구들은 사용자의 속성 및 아이템의 속성이 될 수 있는 다양한 종류의 속성들 중 일부에 특화되어서 제한적으로 유효한 방법이었다. LAR_CF는 속성의 종류를 막론하고 아이템이 가지는 속성에 대한 사용자의 선호도를 이끌어내어 활용성 측면의 개선을 위해서 일반화 방법을 제시함과 동시에 데이터 희소성 문제를 개선하려는 것으로, 그 방법을 (그림 2)에 나타내었다.

제안하는 방법은 총 6단계로 구성되며 2단계부터 4단계까지는 최근접 이웃 수를 의미하는 k 만큼 이웃을 변경하며 반복 수행된다. LAR_CF는 기존의 NBCF 프로세스에 객체의 속성 정보를 이용하는 절차를 더하여 객체 사이의 유사도를 보다 정확하게 산출하려는 것을 목적으로



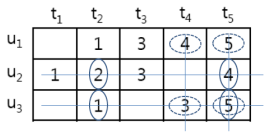
(그림 2) LAR_CF의 처리 절차
(Figure 2) Process of LAR_CF

하는 방법이다. LAR_CF를 설명하기 위해 수반되는 기호 표기법을 (표 1)에 정리하여 나타내었다. 기본적으로 사용자가 아이템에 명시적으로 평가한 선호도 점수를 특징벡터로 삼는 기존의 프로세스에 객체의 속성에 대한 특징벡터를 새로이 도출하여 두 가지 특징벡터를 결합함으로서 더욱 정확한 유사도 산출을 꾀한다.

(표 1) Notation
(Table 1) Notation

기호	의미
$t,\ t_i$	임의의 아이템 및 아이템 i
$u,\ u_i$	임의의 사용자 및 사용자 i
$o,\ o_i$	임의의 객체 및 객체 i
$r_{i,j}$	사용자 i가 아이템 j에 평가한 선호도
$ar_{i,j}$	아이템 i의 속성 j 값
k	최근접 이웃 수
k_m	최대 최근접 이웃 수
$sim(i,j)$	객체 i 및 객체 j 의 유사도
$P(i,j)$	사용자 i의 아이템 j에 대한 예측 선호도

① 목표 사용자 및 목표 아이템 설정 단계에서는 예측 대상이 되는 사용자 및 아이템을 설정한다. 협업 필터링에서의 선호도 예측의 목적은 사용자가 경험하지 않은 아이템에 대한 선호도를 예측하고 높게 예측된 아이템을 사용자에게 추천하려는 것이다. 예를 들어, 최대 3명의 사용자가 최대 5개의 아이템에 대해 평가한 선호도 매트릭스의 예를 나타낸 (그림 3)에서 비어 있는 공간의 선택이 목표 사용자 및 목표 아이템을 결정한다. u_2 가 t_4 에 평가하기 위한 공간이 비어 있으므로 이를 예측한다고 가정하면, 목표 사용자는 2가 되며 목표 아이템은 4가 되므로, 예측값인 $P(2,4)$ 를 구하는 것이 최종적인 목표가 될 것이다.

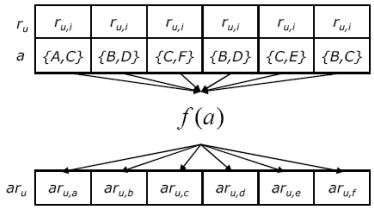


(그림 3) 선호도 매트릭스의 예
(Figure 3) An Example of Preference Matrix

② 선호도 특징벡터 추출 단계에서는 어떤 객체를 중심으로 예측을 수행할 것인지에 따라서 목표 객체와 하나의 이웃 객체 사이의 특징벡터를 추출한다. 사용자 기반 방법을 채택한다면 단계 ① 에서 설정한 목표 사용자 및 목표 아이템에 선호도 평가 기록이 있는 다른 하나의 이웃 사용자 사이의 선호도를 추출하며, 아이템 기반 방법을 채택한다면 단계 ① 에서 설정한 목표 아이템 및 동일한 사용자에 의해 선호도 평가된 다른 하나의 이웃 아이템 사이의 선호도를 추출한다. (그림 3)의 $r_{2,4}$를 예측하기 위해서 $P(2,4)$ 를 구하기 위한 방법을 예로 들어 설명하면, 사용자 기반 방법의 경우에는 목표 아이템 t_4에 평가한 u_1 및 u_3이 최근접 이웃에 포함될 수 있으므로 최대 최근접 이웃 수 $k_m=2$ 가 된다. 특징벡터는 목표 사용자와 이웃 사이의 공통 선호도를 추출하는데 아이템 기반 방법을 고려하는 경우에는 목표 사용자 u_2 가 평가한 목표 아이템 t_4 를 제외한 다른 아이템인 t_1, t_2, t_3, t_5 가 최근접 이웃에 포함될 수 있으므로 최대 최근접 수 $k_m=4$ 가 되며, 사용자 기반 방법과 마찬가지로 목표 아이템과 이웃 아이템 사이의 공통 선호도를 취하여 특징벡터로 삼는다. 사용자 기반 및 아이템 기반 방법은 각각의 최대 최근접 수인 k_m 만큼 이웃을 바꾸어가며 반복 수행된다. 따라서, (그림 3)의 데이터세트 조건에서 u_2 와 u_3 의 유사도 산출을 위해 본 단계에서 추출된 특징벡터는 u_2={2, 4} 및 u_3={1, 5} 가 된다. 또한, t_4 와 t_5 의 유사도 산출을 고려한다면 본 단계에서 추출된 특징벡터는 t_4={4, 3} 및 t_5={5, 5} 가 될 것이다. 이를 (그림 3)에 표기하였다.

③ 속성 특징벡터 추출 단계에서는 각 객체가 가진 속성에 대한 특징을 유사도에 반영하기 위한 벡터를 추출하기 위한 것이다. 이를 위해서는 속성 매트릭스를 필요로 하며, 이는 각 객체가 가지고 있는 속성에 대한 정보를 담고 있다. 사용자 기반 방법을 고려하는 경우에 사용자가 가질 수 있는 속성은 매우 다양하며 특징벡터로 삼아 유사성 도출에 적용할 수 있다.

예를 들어, {나이, 성별, 거주지, 직업} 등을 특징벡터로 고려할 수 있다. 아이템이 가질 수 있는 속성 또한 아이템의 종류에 따라 장르, 주연배우 목록, 제작년도 등 많은 속성들을 이용할 수 있음은 물론이다. 단, 속성 특징벡터는 수치로 정량화되어야 하며 선호도 특징벡터와 동일한 구간으로 매핑 되도록 설계되어야 한다.



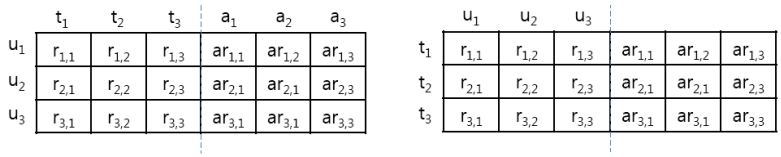
(그림 4) 속성 특징벡터 생성의 예
(Figure 4) An Example of Generation of Attribute Feature Vector

본 논문에서 제안하는 LAR_CF에 의해 개발된 속성 특징벡터 생성의 예를 (그림 4)에 나타냈다. 공통 선호도 6개에 대해서 각 공통 선호도 아이템이 가질 수 있는 속성의 목록이 {A, B, C, D, E, F} 중에서 복수의 속성을 가질 수 있는 상황을 가정하였다. $f(a)$는 사용자가 아이템에 대해 부여한 선호도를 입력받아 각 아이템이 가진 속성에 대한 선호도를 출력하는 함수이며, 최종적으로 {A, B, C, D, E, F} 속성에 대한 사용자의 속성 특징벡터 ar_u 를 출력하는 것이다. 즉, $f(a)$에 따라서 속성 선호도의 생성 매커니즘이 변경될 수 있다. 본 연구에서는 (그림 4)와 같은 속성 특징벡터 도출 방법에 기반하여 $f(a)$를 식 1과 같이 설계하였다. a_j는 임의의 아이템이 가진 속성들을 나타낸다. 각각의 속성 j 에 대한 특징벡터 a_j 에 대하여 속성 j 를 가진 모든 공통 평가 아이템 I 의 n 개 선호도 평균인 E 를 취한다.

$$f(a) = E[a_j] = \frac{1}{n}\sum_{i=1}^{n} r_{u,i} \quad when \quad i \ni a_j \qquad (1)$$

속성 특징벡터를 사용자 기반 방법에 적용하게 되면 사용자의 선호도 평가와 무관하게 기본적으로 데이터베이스에 특징벡터가 존재하는 것이므로, 본 단계에 의해서 두 사용자 사이에 공통 선호도가 존재하지 않더라도 유사도 도출이 가능해지기 때문에 협업 필터링의 데이터 희소성 문제를 개선할 수 있는 장점을 가진다. 또한, 아이템 기반 방법에서는 아이템이 가질 수 있는 속성에 의해 아이템의 속성에 대한 사용자의 잠재적 선호도를 특징벡터로 추출할 수 있는 장점을 가진다.

④ 특징벡터 결합 및 유사도 측정 단계에서는 단계 ②



(a) 사용자 기반 방법 (b) 아이템 기반 방법

(a) User-based method (b) Item-based method

(그림 5) 특징벡터의 선형 결합

(Figure 5) Linear Combination of Feature Vectors

에서 추출된 선호도 특징벡터와 단계 ③에서 추출된 속성 특징벡터를 선형으로 결합한다. 사용자 기반 방법에서는 사용자 사이의 유사도를 측정하므로 사용자 속성이 사용되며 아이템 기반 방법에서는 아이템 사이의 유사도를 측정하므로 아이템 속성이 사용된다. (그림 5)는 선호도 특징벡터와 속성 특징벡터의 선형 결합을 보인 것이다.

(그림 5)의 (a)는 사용자 기반 방법일 때 속성 특징벡터가 결합된 매트릭스 형태를 나타내며, (그림 5)의 (b)는 아이템 기반 방법일 때 속성 특징벡터가 결합된 매트릭스 형태를 나타낸다. 특징벡터가 결합된 후에 유사도를 산출하는데, 선형적 관계에 있는 두 변수 사이의 유사도를 정량적으로 측정할 수 있는 모든 척도가 고려될 수 있음을 관련 연구에서 설명하였다.

보편적으로 사용되는 유사성 척도 중 하나는 피어슨 상관계수를 들 수 있다. 선형적 관계를 갖는 임의의 변수 X 및 Y 사이의 피어슨 상관계수로 유사도를 산출하는 과정에서 모집단은 불확실하므로 식 (2)와 같은 표본 상관계수를 이용한다. 동일한 조건에서 코사인 계수를 산출하는 방법은 식 (3)과 같다.

$$PCC(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{Y})^2}} \quad (2)$$

$$\cos(\vec{A},\vec{B})=\frac{\vec{A}\bullet\vec{B}}{\|\vec{A}\|*\|\vec{B}\|} \quad (3)$$

⑤ 예측 단계에서는 단계 ②부터 단계 ④까지 각 이웃을 대상으로 반복 연산하여 산출한 유사도를 바탕으로 목표 사용자의 목표 아이템에 대한 선호도를 예측한다. 사용자 기반 방법에서 일반적으로 사용되는 방법은 식 (4)에 나타낸 Bias-From-Mean Average 계산법이며, 아이템 기반 방법에서는 이를 식 (5)와 같이 변형하여 적용한다.

$$p(u,t)=\bar{r_u}+\frac{\sum_{i=1}^{n}sim(u,i)(r_{i,t}-\bar{r_i})}{\sum_{i=1}^{n}sim(u,i)} \quad (4)$$

$$p(u,t)=\bar{r_t}+\frac{\sum_{i=1}^{n}sim(t,i)(r_{u,t}-\bar{r_u})}{\sum_{i=1}^{n}sim(t,i)} \quad (5)$$

⑥ 성능 평가 단계에서는 예측된 선호도와 실제 사용자가 평가한 선호도의 오차에 대한 절대치를 반복적으로 구하여 그 평균을 취하는 절대 평균 오차(Mean Absolute Error, MAE) 또는 평균 제곱근의 오차를 취하는 평균 제곱근 오차(Root Mean Squared Error, RMSE)를 사용한다. 각 방법은 식 (6) 및 식 (7)과 같다. n 은 총 예측 횟수를 나타내며 p_i 는 I 번째 예측에서 예측된 선호도, q_i 는 i 번째 실제 선호도를 나타낸다.

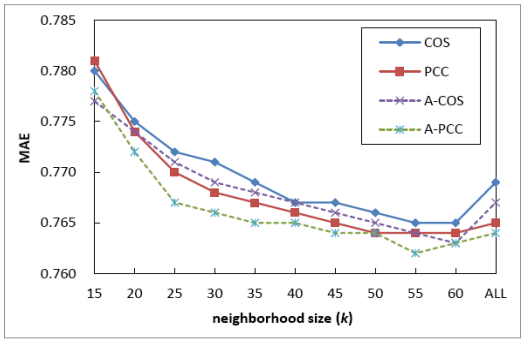
$$MAE=\frac{\sum_{i=1}^{n}|p_i-q_i|}{n} \quad (6)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i-q_i)^2} \quad (7)$$

4. 실험 및 결과

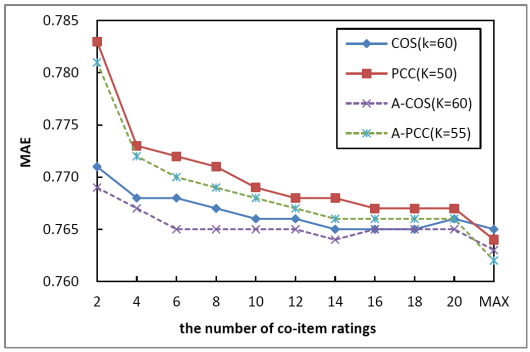
LAR_CF의 성능 평가를 위해 C 컴파일러를 이용하여 아이템 기반 방법이 적용된 실험용 NBCF 시뮬레이터를 구현하였다. 구현 조건은 3장에서 설명한 전반적인 내용 및 단계 ③의 식 (1)을 적용하였고, 실험을 위한 유사성 척도로써 2장 및 3장에서 소개한 피어슨 상관계수(PCC) 및 코사인 계수(COS) 유사성 척도를 사용하였다. LAR_CF를 적용한 것과 적용하지 않은 것의 선호도 예측 성능 비교를 통해서 LAR_CF의 유효성을 확인하고자 하였다.

실험용 데이터는 기존의 연구들에 의해 널리 사용된 바 있는 GroupLens의 MovieLens 100k 데이터세트 중, 데이터세트 제공자가 실험용 및 학습용 데이터로 분리 제공하는 U1을 사용하였다[19, 20]. 본 데이터세트는 1,682개의 영화 아이템에 943명의 사용자가 부여한 1부터 5까지의 정수형 선호도 데이터 100,000개로 구성되어 있으며, 각 사용자는 최소 20개의 영화에 대해 선호 점수를 부여한 데이터로 구성되어 있다. 영화는 16개 중 다수의 장르를 속성으로 가지고 있으므로, LAR_CF에 의해 최대 16개의 잠재적 속성 선호도가 특징벡터로 추가될 수 있다. 전체 데이터세트에서 임의로 20%의 데이터를 추출하여 실험용 데이터로 사용하였으며 나머지 80%의 데이터를 훈련용 데이터로 사용하여 2장 및 3장에서 설명한 바 있는 MAE를 성능 비교의 척도로 삼았다. 유사성 측정을 위한 최소 공통 선호도 개수인 2부터 개수를 점차 증가시키며 예측 정확도를 관찰하는 인공적 데이터 희소 조건 실험을 수행하면 LAR_CF가 데이터 희소성 문제에 강인한지 확인할 수 있다. 이를 위해서 데이터 희소성 실험 이전에 Full-rating 실험을 우선적으로 실시하여 각 유사성 척도 별 최적의 최근접 이웃 수인 k 값을 탐색하였다. 그 결과는 (그림 6)과 같다.



(그림 6) Full-rating 실험 결과
(Figure 6) Experimental Result Using Full-rating

(그림 6)의 A-COS 및 A-PCC는 코사인 계수와 피어슨 상관계수에 LAR_CF를 적용한 것이다. 실험 결과를 살펴보면, COS 및 A-COS는 k=60, PCC는 k=50, A-PCC는 k=55에서 최적의 성능을 보이는 가운데 최고 성능은 A-PCC의 0.762였다. 이를 COS의 0.765 및 PCC의 0.764와 비교하면 LAR_CF가 기존의 방법보다 더 나은 성능을 보임이 확인된 것이며 Full-rating 조건에서 피어슨 상관계수가 LAR_CF가 가장 효과적으로 적용되었음을 의미하는 것이다. 모든 사용자를 이웃으로 사용하는 ALL은 k의 수만큼 목표 사용자와 유사한 사용자를 예측에 활용하는 TOP-k 구조에서 모든 사용자를 이웃으로 사용하는 것에 해당한다. 그러므로, 유사하지 않은 사용자의 선호도까지 목표 사용자에 대한 목표 아이템의 선호도 예측에 활용하게 되기 때문에 TOP-k 방법이 적용되지 않는 것과 동일하므로 예측 정확도가 나빠지게 된다. Full-rating 실험을 통해서 각 유사성 척도 별로 최적의 k값을 찾는 이유이다.



(그림 7) 인공적 데이터 희소 조건 실험 결과
(Figure 7) Experimental Result of Artificial Data Sparsity Condition

(그림 7)은 Full-rating 실험 결과로부터 도출된 최적의 k값을 이용하여 선호도 특징벡터의 개수를 인공적으로 제한함으로서 데이터 희소 조건을 인위적으로 만들어 선호도 예측 성능 실험을 수행한 결과이다. 그러므로 (그림 7)의 MAX는 인공적 데이터 희소 조건이 적용되지 않은 것이고 (그림 6)의 유사성 척도 별 최고 성능과 동일하다. (그림 7)에서 관찰되는 바와 같이 LAR_CF에 의해서 기존보다 데이터 희소 조건에 강인해짐을 확인할 수 있다. 특히 코사인 계수를 이용하는 경우에 공통 선호도 수와 상관없이 전반적으로 가장 좋은 성능을 보임을 쉽게 확인할 수 있다. Full-rating 실험에서 좋은 성능을 보인 유사성 척도인 PCC가 데이터 희소 조건에서 나쁜 성능을 보이는 이유는 특징벡터의 수가 약 5개 미만인 경우에 잘못된 상관계수를 보이는 단점과 특징벡터의 수가 2개 미만인 경우에 유사도 측정이 불가능한 점 때문인데, 이를 뒷받침할 수 있는 근거는 기존의 연구들에서 발견된다[7, 8]. 즉, 피어슨 상관계수는 데이터 희소 문제에 매우 취약한 단점이 있으며, LAR_CF에 의한 특징벡터의 차수 증가에 의해 이를 개선하였음을 확인할 수 있다.

5. 결 론

본 논문에서는 NBCF에서 두 사용자의 유사성을 산출하고자 할 때, 객체 사이의 특징벡터 차수가 적을 때 유사도가 부정확하게 계산되거나 계산할 수 없는 상황에 봉착하는 데이터 희소성 문제를 개선하기 위해 객체의 속성을 이용하는 일반화된 방법인 LAR_CF를 제안하였다. 사용자의 주관적인 선호도 평가 결과로부터 알아낼 수 있는 공통 아이템의 속성 및 사용자 고유의 특성을 이용하면 LAR_CF를 효과적으로 적용할 수 있다. LAR_CF의 효과를 살펴보기 위해 영화 아이템의 장르 속성을 이용하여 Full-rating 및 인공적 데이터 희소 조건에서 실험한 결과, LAR_CF가 데이터 희소 조건에서 유효하였고 이러한 내용이 전반적 성능의 개선으로 이어짐을 확인하였다. 제안하는 방법은 일반화 관점에서 개발되었으므로 실제 적용 시에는 아이템의 속성과 사용자의 속성을 잘 결정하여 더 나은 유사도를 도출할 수 있도록 하는 것이 바람직할 것이다.

감 사 의 글

본 연구는 2010년도 정부(교육과학기술부) 재원 한국연구재단의 기초연구사업(2010-0021411)과 2013년도 정부(교육부)의 재원으로 한국연구재단의 중점연구소지원사업으로 수행된 연구임(NRF-2010-0020210).

참 고 문 헌(Reference)

[1] Joseph A. Konstan and Jhon Riedl, “Deconstructing Recommender Systems”, IEEE Spectrum, October 2012.

[2] Adomavicius G. and Tuzhilin, A., “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions”, IEEE Trans. Know. and Data Eng., Vol. 17 No. 6, pp. 734-749, 2005.

[3] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers and Jhon Riedl, “An Algorithmic Framework for Performing Collaborative Filtering” ACM SIGIR 22nd Int. Conf. Research and Development in Information Retrieval, pp. 230-237, 1999.

[4] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, “Using Collaborative Filtering to Weave an Information Tapestry”, Communications of the ACM, Vol. 35, No. 12, pp. 61-70 1992.

[5] John S. Breese, David Heckerman and Carl Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering”, Proc. the 14th Conf. Uncertainty in Artificial Intelligence, pp. 43-52, 1998.

[6] K. Goldberg, T. Roeder, D. Gupta and C. Perkins, “Eigentaste: A Constant Time Collaborative Filtering Algorithm”, Information Retrieval, Vol. 4, No. 2, pp. 133-151, 2001.

[7] H. J. Kwon and K. S. Hong, “Personalized Smart TV Program Recommender Based on Collaborative Filtering and a Novel Similarity Method”, IEEE Trans. Consum. Electron., Vol. 57, No. 3, pp. 1416-1423, 2011.

[8] H. J. Ahn, ″A New Similarity Measure for Collaborative Filtering to Alleviate the New User Cold-starting Problem″, Information Sciences, Vol. 178, No. 1, pp. 37–51, 2008.

[9] D. Lemire and A. Maclachlan, “Slope One Predictors for Online Rating-Based Collaborative Filtering”, Proc. 5th SIAM Int. Conf. Data Mining, pp. 471-475, 2005.

[10] Yu Li, Liu Lu and Li Xuefeng, “A Hybrid Collaborative Filtering Method for Multiple-interests and Multiple-content Recommendation in E-Commerce”, Expert Systems with Applications, Vol. 28, No. 1, pp. 67-77, 2005.

[11] Buhwan Jeong, Jaewook Lee and Hyunbo Cho, “Improving Memory-based Collaborative Filtering via Similarity Updating and Prediction Modulation”, Information Sciences, Vol. 18, No. 5, pp. 602-612, 2010.

[12] Jun Wang, Arjen P. de Vries and Marcel J. T. Reinders, “Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion”, Proc. 29th ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 501-508, 2006.

[13] Bin Cho, Jian-Tao Sun, Jianmin Wu, Qiang Yang and Zheng Chen, “Learning Bidirectional Similarity

for Collaborative Filtering", LNCS 5211, pp. 178-194, 2008.

[14] T. H. Kim and S. B. Yang, "An Improved Neighbor Selection Algorithm in Collaborative Filtering", IEICE Trans. Inform. and Syst., Vol. E88-D, No. 5, pp. 1072-1076, 2005.

[15] Souvik Debnath, Niloy Ganguly and Pabitra Mitra, "Feature Weighting in Content based Recommendation System Using Social Network Analysis", Proc. of the 17th Int. Conf. on World Wide Web, pp. 1041-1042, 2008.

[16] Karen H. L. Tso-Sutter, Leonardo Balby Marinho and Lars Schmidt-Thieme, "Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms", Proc. of the 2008 ACM Symposium on Applied computing, pp. 1995-1999, 2008.

[17] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar and David M. Pennock, "Methods and Metrics for Cold-start Recommendations", Proc. of the 25th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 253-260, 2002.

[18] Huang, Z., Chen, H. and Zeng, D. "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering", ACM Transactions on Information Systems, Vol. 22, No. 1, pp. 116-142, 2004.

[19] B. N. Miller, I. Albert, S.K. Lam, J.A. Konstan, J. T. Riedl, "MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System on Four Mobile Devices", Proc. of the 2003 Int. Conf. Intelligent User Interfaces, pp. 263 - 266, 2003.

[20] http://www.grouplens.org/node/73, MovieLens Data Sets, GroupLens Research in Department of Computer Science and Engineering at the University of Minnesota.

◐ 저 자 소 개 ◑

권 형 준

2006년 서울보건대학 컴퓨터정보과(학사)
2008년 성균관대학교대학원 전자전기컴퓨터공학과(석사)
2013년 성균관대학교대학원 전자전기컴퓨터공학과(박사)
2010년~현재 성균관대학교 IT융합연구원 정보통신기술연구소 연구원
관심분야 : 데이터 분석 및 마이닝, 패턴 예측과 인식, HCI
E-mail : katsyuki@skku.edu

홍 광 석

1985년 성균관대학교 전자공학과(학사)
1988년 성균관대학교대학원 전자공학과(석사)
1992년 성균관대학교대학원 전자공학과(박사)
1993년 서울보건대학 전산정보처리과 전임강사
1995년 제주대학교 정보공학과 전임강사
1997년~현재 성균관대학교 정보통신대학 교수
관심분야 : 오감인식, 융합 및 재현, HCI
E-mail : kshong@skku.ac.kr