한국지능정보시스템학회논문지 제11권 제1호 2005년 6월(pp.107~121)

클러스터링 기반 사례기반추론을 이용한 웹 개인화 추천시스템

홍태호
부산대학교 상과대학 경영학부 조교수
(hongth@pusan.ac.kr)

이희정
부산대학교 상과대학 경영학부
(deweyes@pusan.ac.kr)

서보밀
숙명여자대학교 경상대학 경영학부 조교수
(bmsuh@sookmyung.ac.kr)

최근, 추천시스템과 협업 필터링에 대한 연구가 학계와 업계에서 활발하게 이루어지고 있다. 하지만, 제품 아이템들은 다중값 속성을 가질 수 있음에도 불구하고, 기존의 연구들은 이러한 다중값 속성을 반영하지 못하고 있다. 이러한 한계를 극복하기 위하여, 본 연구에서는 추천시스템을 위한 새로운 방법론을 제시하고자 한다. 제안된 방법론은 제품 아이템에 대한 클러스터링 기법에 기반하여 다중값 속성을 활용하며, 정확한 추천을 위하여 협업 필터링을 적용한다. 즉, 사용자 간의 상관관계만이 아니라 아이템 간의 상관관계를 고려하기 위하여, 사용자 클러스터링에 기반한 사례기반추론과 아이템 속성 클러스터링에 기반한 사례기반추론 모두가 협업 필터링에 적용되는 것이다. 다중값 속성에 기반하여 아이템을 클러스터링 함으로써, 아이템의 특징이 명확하게 식별될 수 있다. MovieLens 데이터를 이용하여 실험을 하였으며, 제안된 방법론이 기존 방법론의 성능을 능가한다는 결과를 얻을 수 있었다.

논문접수일 : 2005년 2월 게재확정일 : 2005년 6월 교신저자 : 서보밀

1. 서론

최근 웹의 급격한 발전으로 인하여 인터넷에서의 전자상거래가 더욱 활발하게 진행되고 있다. 인터넷을 통하여 쇼핑을 할 뿐만 아니라, 영화나 음악 서비스도 받을 수 있게 되었다. 이러한 전자상거래의 빠른 성장으로 인하여 기업과 고객 모두는 새로운 상황에 직면하게 되었다. 전자상거래 영역에서의 기업들은 더욱 치열해진 경쟁으로 인하여 생존에 어려움을 느끼고 있다. 반면, 고객들은 인터넷의 방대한 정보로 인하여 정보 홍수에 빠지게 되어 정보에 대하여 까다로운 요구를 하고 있다. 이를테면, 개인화 또는 맞춤화된 정보의 제공을 원하고 있는 것이다.

이러한 상황으로 인하여, 일대일 마케팅(one-to-one marketing), 웹 개인화(Web personalization), 고객관계관리(Customer Relationship Management: CRM)와 같은 새로운 마케팅 전략의 필요성이 실무에서뿐만 아니라 연구에서도 강조되고 있다(Berson et al. 2000; Changchien and Lu, 2001; Sarwar at al. 2000; Yuan and Chang, 2001). 맞춤화와 개인화 서비스는 인터넷 쇼핑몰이나 웹 서비스 제공자에게 있어 중요한 성공요인이기 때문에 대량 맞춤(mass customization)을 실현할 수

* 이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.

있는 웹 기반의 추천시스템(recommendation system)에 대한 관심이 더욱 높아지고 있는 것이다. 추천시스템이 널리 이용되는 이유는 크게 4가지로 정리할 수 있다. 첫째, 대량 맞춤을 통하여 일대일 마케팅이 가능하다. 둘째, 웹 개인화와 맞춤화 서비스가 가능하다. 셋째, 정보 과부하(information overlord) 문제를 해결할 수 있다. 넷째, 추천시스템은 판매자와 구매자 모두에게 부가가치를 제공한다.

이러한 장점들 때문에 GroupLens 연구소를 비롯하여 학계에서는 추천시스템의 알고리즘과 정보 필터링에 관한 다양한 연구를 진행하고 있으며, Amazon, CDnow, 삼성몰, 예스24와 같은 유수의 전자상거래 업체에서도 추천시스템은 기업의 경쟁도구로 이용되고 있다. 특히, 협업 필터링(Collaborative Filtering: CF) 기법을 이용한 추천시스템이 정보 과부하 문제의 해결에 강점을 가지고 있기 때문에 실무에서 많이 사용되고 있다. 따라서, 본 연구에서는 실무와 학계에서 활발히 진행되고 있는 추천시스템의 추천 능력을 개선할 수 있는 새로운 모델을 제안하고자 한다.

본 연구의 목적은 새로운 개발방법론을 통하여 대량 맞춤을 실현 가능하게 하는 개인화된 추천시스템을 구축하여 보다 정확하고 맞춤화된 정보를 고객에게 제공하는 것이다. 제안된 추천시스템 모델을 인터넷 비즈니스뿐만 아니라 고객관리를 중요시하는 모든 분야에 적용하여 향상된 고객관계관리 실현이 가능하도록 하고자 한다.

협업 필터링과 추천 알고리즘에 관한 많은 연구들이 국내외에서 진행되어오고 있으나, 다중값 속성(multi-valued attribute)을 지닌 아이템을 반영한 연구는 거의 없었다. 본 연구에서는 추천시스템의 핵심인 추천 능력을 개선하기 위하여 아이템의 다중값 속성을 이용하고자 한다. 데이터마이닝 기법 중의 하나인 클러스터링 기법과 사례기반추론(Case-Based Reasoning: CBR)을 통하여 아이템의 다중값 속성을 응용하는 새로운 방법론을 제시할 것이다. 즉, 아이템 속성 클러스터링에 기반한 사례기반추론을 통하여 사용자 간 연관성과 아이템 간 연관성을 함께 고려한 협업 필터링 기법을 제안하고자 한다. 본 연구에서는 GroupLens 연구소에서 제공하는 MovieLens 데이터를 이용하였다.

2. 관련 연구

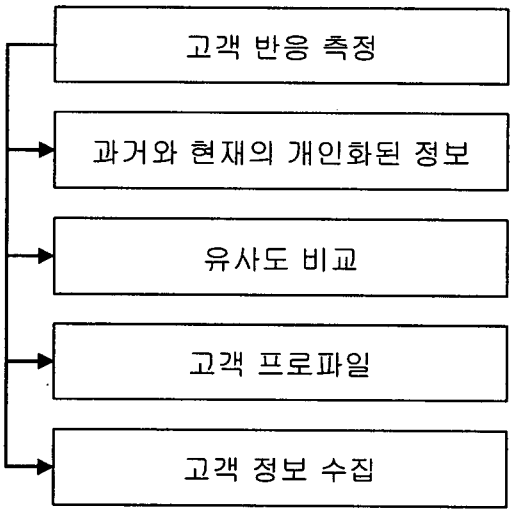
2.1 추천시스템

추천시스템은 고객이 관심을 가지는 상품에 관한 정보나 인구통계학적 정보, 과거 구매행동 분석을 토대로 고객의 요구에 맞는 상품을 추천해주는 시스템이다(Sarwar et al. 2001). 또한, 고객들이 구매하고자 하는 상품을 쉽게 찾을 수 있도록 도와주는 정보 필터링 기술이기도 하다(Schafer et al. 1999). 추천시스템은 인터넷을 기반으로 고객 개개인에 대하여 일대일 마케팅을 가능하게 하는 e-CRM의 한 분야로서, Amazon, CDnow 등 해외의 유수 전자상거래 사이트에 적용되고 있으며 Ringo 음악 추천이나 Bellcore 비디오 추천에도 이용되고 있다. Amazon이나 CDnow, eBay와 같은 전자상거래 기업들은 그들이 현재 보유한 고객을 유지하기 위하여 보다 개인화된 서비스를 제공하려고 끊임없이 노력하고 있는 것이다.

전자상거래에서 고객 개개인에게 맞춤화된 개인화 서비스가 강조되고 있다(Schafer et al. 2001). 개인화(personalization)는 웹 사이트 상에서 어떤 특정 고객만의 고유하고도 특정한 요구에 대

하여 민첩하게 대응할 수 있는 것을 의미한다. Mobasher et al.(2002)은 웹 개인화를 인터넷 사용에서 개개인의 고객 관심이나 취미에 따라 반응하는 활동으로 정의하고 있다. 개인화 서비스가 중요한 이유는, 고객의 입장에서는 상품 검색의 노력을 줄일 수 있고 기업 입장에서는 적합한 상품 추천으로 인하여 전자상거래 사이트에 대한 고객 충성도를 높여 줄뿐만 아니라 고객과의 유대감을 형성시킬 수 있기 때문이다. 이에 따라, 개인화 서비스는 고객관계관리 측면에서도 중요하게 인식되고 있다(Kim et al. 2002; Mild and Natter, 2002).

Adomavicius and Tuzhilin(2001)은 [그림 1]과 같이 개인화의 과정을 5단계로 제안하고 있다. 먼저 이후 단계를 모니터 하기 위하여 고객들의 반응을 측정하고, 수집된 고객 정보를 통하여 다시 고객 반응을 알 수 있다.



[그림 1] 개인화 과정

2.2 개인화 기법

추천시스템을 위한 개인화 기법으로는 다음과 같은 3가지가 있다(Kuo and Chen, 2001):

(1) 규칙 기반 필터링(rule-based filtering)
인구통계학적 정보나 개인신상 정보를 사용자에게 질문하여 그 응답에 맞는 규칙을 기반으로 하여 추천하는 개인화 기법이다.

(2) 학습 에이전트(learning agent)
웹 사이트 방문기록 및 횟수, 접속장소, 시간 등 일종의 로그 파일 분석을 통하여 사용자의 속성, 습관, 개인의 선호를 추적하는 학습 에이전트를 이용하는 개인화 기법이다.

(3) 협업 필터링(collaborative filtering)
고객이 선호하는 패턴과 유사한 다른 고객들의 선호도를 이용하여 고객에게 관련된 서비스를 추천하는 개인화 기법이다.

이 중 협업 필터링은 추천시스템 분야에서 가장 성공적인 추천기법으로 전자상거래 기업에 가장 널리 이용되고 있다(Konstan et al. 1997). 협업 필터링의 핵심은 추천대상 고객과 유사한 고객들을 찾고, 선호도가 유사한 고객군에서 높이 평가한 아이템을 추천하는 것이다.

추천시스템에 이용되는 협업 필터링에는 크게 두 가지 접근방법으로 나눌 수 있다. 사용자 기반 협업 필터링과 아이템 기반 협업 필터링이 그것이다(Herlocker et al. 1999). 사용자 기반 협업 필터링은 사용자 간의 유사성을 측정하여 선호도가 비슷한 다른 고객들이 평가한 상품을 기반으로 어떤 특정 고객이 선호할만한 상품을 추천하는 방식이다. 즉, 사용자 기반 협업 필터링 방식은 사용자와 사용자 간의 연관관계를 파악하는 것이 핵심이며, 이처럼 다른 사용자 집단의 패턴을 기반으로 추천하는 알고리즘이라는 의미에서 '사회적 필터링'이라고 부르기도 한다. 사용자 기반 협업 필터링에서 사용자 간의 연관성을 토대로 어떤 특정 고객과

선호도가 비슷한 이웃들을 선정하는 기법으로는 클러스터링, 최근접 이웃 추출법, 베이지안 네트워크 등이 있다.

사용자 기반 협업 필터링은 선호도가 비슷한 유사 고객들이 동일하게 평가한 상품에 대하여 상대적으로 높은 예측력을 보이고 있고(Konstan et al. 1997), 새로운 상품에 대한 예측도 가능하며, 데이터가 많은 경우에는 다른 기법에 비해 상대적으로 정확한 예측을 한다는 장점을 가지고 있다. 그러나, 사용자 기반 협업 필터링의 한계점은, 둘 또는 그 이상의 고객이 모두 평가를 내린 동일 상품이 있어야 하는데, 고객들이 서로 이질적인 평가를 한 상품에 대해서는 단지 두 고객 사이에서만 상관관계를 구하므로 예측의 정확성이 떨어질 가능성이 있다는 것이다. 또한, 고객의 집단이 커지면 규모의 문제(scalability problem)가 발생하게 되어 연산처리를 많이 해야 한다는 단점도 가지고 있다.

협업 필터링의 또 하나의 방법으로는 아이템 기반 협업 필터링이 있다. 아이템 기반 협업 필터링은 대부분의 사람들이 과거에 자신이 좋아했던 상품과 유사한 상품을 선호하는 경향이 있고, 반대로 싫어하거나 선호하지 않았던 상품과 유사한 상품은 선호하지 않는 경향이 있다는 점을 근간으로 하고 있다. 아이템 기반 협업 필터링은 아이템 간의 유사성, 즉 고객이 선호도 등급을 입력한 기존 상품들과 추천하고자 하는 상품들 간의 유사성을 측정하여 어떤 특정 고객이 어떤 상품을 선호하는지 예측하여 추천하는 방식이다. 즉, 예측하고자 하는 상품과 유사한 상품들에 대하여 고객이 높은 점수로 평가하였다면 그 상품도 높게 평가할 것이며, 반면에 낮은 점수로 평가하였다면 그 상품도 역시 낮은 점수로 평가할 것이라고 예측하는 것이다. 이 방법은 상품들 간의 유사도를 계산하기 위하여 두 상품 모두에 선호도를 입력한 고객들의 선호도를 사용한다. 그러나, 고객들 간의 유사도가 전혀 고려되지 않기 때문에, 어떤 특정 고객과 전혀 선호도가 다른 사용자들의 평가를 기반으로 한다면 상품들 간의 상관관계 정확도가 떨어지게 되어 추천시스템의 예측력과 추천능력이 떨어지게 될 수 있다.

지금까지 사용자 기반 협업 필터링과 아이템 기반 협업 필터링의 단점을 보완하고 추천시스템의 핵심인 예측력과 추천능력을 향상시킬 수 있는 협업 필터링과 추천 알고리즘에 관한 많은 연구들이 진행되어 왔다. Sarwar et al.(2001)은 데이터의 희박성(sparsity)과 규모의 문제를 지닌 사용자 기반 협업 필터링을 보완하기 위해 아이템에 기반한 협업 필터링 추천 알고리즘을 제시하였다. 박지선 등(2002)은 아이템 기반 협업 필터링의 문제점을 보완하고자, 유사 고객군을 찾은 다음 그들이 평가한 상품들 간의 유사도를 구해 추천하는 방식인 2-way 협업 필터링을 이용한 연구를 제시하였다. 또한, Kim et al.(2002)은 K-means 클러스터링 기법을 추천 알고리즘에 적용하였고, Li and Kim (2003)은 클러스터링 기법을 아이템 기반 협업 필터링에 응용하였다. 이 외에도, Roh et al.(2003)의 연구에서는 클러스터링을 위하여 SOM을 사용하고 유사도를 찾기 위하여 최근접 이웃법을 이용하여 기존의 협업 필터링 방법과 비교 분석하여 예측 성능의 우수함을 설명하였다. 또한, 김재경 등(2003)은 전자상거래 쇼핑몰에 웹 데이터마이닝과 클러스터링에 기반한 협업 필터링 추천에 관한 방안을 제시하였다. Weng and Liu(2004)는 고객이 새로운 제품을 아직 구매하지 않았기 때문에 기존의 장바구니 분석과 협업 필터링 분석을 통해서는 새로운 제품을 추천할 수 없다고 보고 제품의 특징에 기반한 추천을 제안하였다.

3. 연구모형

지금까지 협업 필터링을 이용한 추천시스템에 관한 많은 연구들이 진행되고 있으며, 특히 최근접 이웃법(nearest neighbor method)이나 K-means 클러스터링, SOM 클러스터링 기법들이 적용되어 왔다. 하지만, 아이템의 다중값 속성을 활용하여 아이템을 클러스터링 하고, 사용자별 클러스터(cluster)를 통하여 추천하는 방법에 관한 연구는 거의 없었다. 기존의 연구에서는, 어떤 아이템이 다중값 속성을 가지고 있음에도 불구하고, 오직 단일값 속성(single-valued attribute)으로만 고려하여 아이템을 분류하였다. 이렇게 특정 값을 가지는 속성인 것으로 분류를 하게 되면, 다중값 속성의 성격이 희석되어 추천 알고리즘이 아무리 우수하다고 할지라도 결국에는 추천 예측력이 떨어지게 된다.

따라서, 본 연구에서는 사용자 기반 협업 필터링과 아이템 기반 협업 필터링의 한계를 보완하고자, 다중값 속성을 활용하여 아이템을 클러스터링 하고, 사용자별로 클러스터링 된 소속 그룹에 따라 사례기반추론을 이용하여 아이템을 추천하는 방법론을 제안하고자 한다.

본 연구에서는 3가지의 방법론을 비교, 분석할 것이다. 먼저, 가장 일반적인 협업 필터링 방법인 전체 사용자를 기반으로 하는 사례기반추론 협업 필터링 방법(CBR_CF)을 적용하고, 다음으로는 사용자의 선호도별로 클러스터링 된 소속 군집 내에서 사례기반추론을 수행하는 협업 필터링 방법(UC_CBR_CF)을 적용한다. 이 2가지 방법론은 본 연구에서 제안하고자 하는 방법론에 대한 벤치마크로서 역할을 할 것이다. 마지막으로, 본 연구에서 제안하는 새로운 추천시스템 개발 방법론으로 사용자별 클러스터링 사례기반추론과 아이템 속성별 클러스터링에 기반한 사례기반추론을 결합한 하이브리드 추천시스템(UC_IC_CBR_CF)을 적용한다.

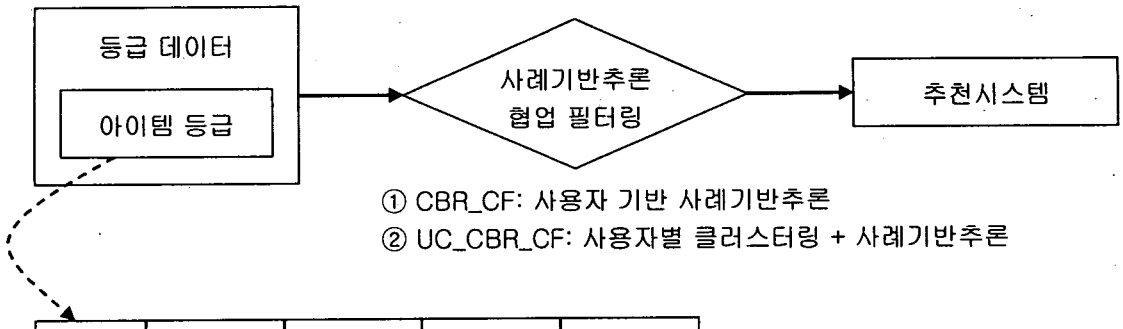
첫 번째, CBR_CF 모델은, 사용자들을 클러스터링 하지 않고, 각 아이템에 등급을 매긴 여러 사용자들을 기반으로 사례기반추론을 이용하여 최근접 이웃을 추출하는 협업 필터링 추천방법이다.

두 번째, UC_CBR_CF 모델은 사용자별로 클러스터링을 하여 사용자의 소속 그룹을 결정하고, 그에 따라 소속 그룹의 다른 사용자들이 선호하는 아이템을 사례기반추론을 통하여 추천하는 협업 필터링을 말한다. UC_CBR_CF의 과정은 다음과 같다:

(1) 사용자별로 클러스터링 한다.
(2) 사용자의 소속 그룹을 정한다.
(3) 소속 그룹 내에서 사례기반추론을 통하여 유사한 사용자들을 추출한다.
(4) 유사한 사용자들이 선호하는 아이템을 추천한다.

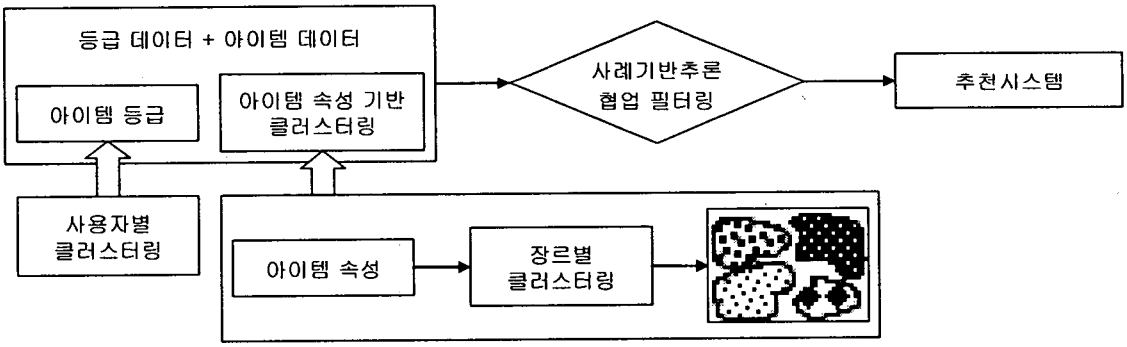
[그림 2]는 사용자가 아이템에 대하여 평가값을 매긴 등급 데이터를 가지고 사례기반추론 협업 필터링을 이용하여 추천하는 과정을 나타내고 있다. CBR_CF 모델은 사용자별 클러스터링이 없는 전체 사용자에 대한 사례기반추론 협업 필터링 과정을 말하는 것이고, UC_CBR_CF 모델은 사용자별로 클러스터링 한 다음 사례기반추론 협업 필터링을 수행하는 것이다.

세 번째는, 하이브리드 모델인 사용자별 클러스터링과 아이템별 클러스터링에 기반한 사례기반추론 협업 필터링을 UC_IC_CBR_CF 모델로 제시한다. 제안하고자 하는 UC_IC_CBR_CF 모델의 추천과정은 다음과 같은 단계를 거치며, [그림 3]과 같이 표현된다.



사용자	아이템1	아이템2	...	아이템n
34	5	4	...	2
67	4	3	...	4
98	3	4	...	1
125	3	4	...	2

[그림 2] 사례기반추론 협업 필터링 과정



[그림 3] UC_IC_CBR_CF 모델의 추천과정

(1) 아이템 속성별로 클러스터링 한다.
(2) 사용자별로 클러스터링한다.
(3) 사례기반추론을 통하여 유사한 이웃을 추출한다.
(4) 유사한 고객이 선호한 상품을 추천한다.

CBR_CF 모델과 UC_CBR_CF 모델이 사용자 간의 유사성만을 고려하는 반면에, 제안 모델인 UC_IC_CBR_CF 모델에서는 아이템 간의 유사성도 같이 고려한다. 즉, 유사한 아이템끼리 그룹화하기 위하여 아이템의 속성 기반으로 클러스터링하는 것이다. 클러스터링을 한 다음, 어떤 사용자가 어떤 종류의 아이템을 더 선호하는지 그 소속 그룹을 정하고, 정해진 그룹 내에서 사례기반추론을 통하여 유사한 사용자들이 선호하는 아이템을

추천해주는 추천시스템 개발방법론이다.

본 연구에서 하이브리드 모델을 이용하려는 이유는 사례기반추론을 통하여 상품을 추천하기 위해서는 탐색공간(search space)이 내부적으로 동질적일수록 유리하기 때문이다. 본 연구모형은, 사용자별 아이템별 클러스터링을 통하여 동질적인 탐색공간들을 우선 정의하고 각 탐색공간에서 유사한 이웃을 찾음으로써, 사례기반추론의 성과를 높일 수 있을 것으로 예상된다.

4. 실험 및 결과

4.1 데이터

본 연구에서는 협업 필터링을 이용한 추천시스템에 관하여 많은 연구를 수행하고 있는 GroupLens 연구소에서 공개한 MovieLens의 웹 기반 추천시스템의 영화 데이터 셋을 이용하여 실험을 수행하였다. 이 데이터 셋은 총 943명의 사용자가 1,628편의 영화에 대하여 1~5점 척도 점수로 부여한 총 100,000개의 선호도 점수로 구성되어 있다. 각 사용자들은 적어도 20편 이상의 영화에 대하여 등급 점수를 산정하였고, 영화의 장르는 다큐멘터리, 드라마, 판타지, 공포, 뮤지컬, 필름 느와르, 미스터리, 로맨스, 공상과학, 스릴러, 전쟁, 웨스턴 등의 총 18가지로 구분되어 있다.

4.2 실험과정 및 결과

4.2.1 클러스터링

앞의 연구모형에서 설명한 바와 같이, 본 연구에서는 아이템의 다중값 속성을 활용한 추천시스템 개발방법론에 초점을 맞추고 있다. 이를 위하여 본 실험에서는 영화 데이터(아이템 데이터)를 이용하여 장르 속성(아이템 속성)을 기반으로 K-means 클러스터링을 수행하였다. 장르 속성별로 클러스터링을 수행하는 이유는 영화가 본래 하나의 장르 특성만을 가진 것이 아니라, 여러 장르의 특성을 가지고 있기 때문이다. 예를 들어, '토이스토리'는 어린이, 코미디, 애니메이션의 3가지 장르 특성을 가지고 있는 것이다. 이와 같이 각 영화는 여러 장르의 특성을 지니고 있기 때문에, 하나의 장르로 단정하기에는 애매모호한 점이 많다. 하지만, 장르 속성별로 클러스터링을 수행하면 보다 포괄적인 장르(예: 액션이면서 SF적인 전쟁 영화)로 재정의되어 어떤 종류의 영화인지 알 수 있는 것이다. 추천시스템의 핵심은 사용자가 선호하는 영화를 파악하고 이와 유사한 패턴을 가진 다른 고객이 선호하는 영화를 추천하는 것이다. 장르 속성별로 클러스터링을 수행하면 유사한 영화를 보다 근접하게 찾아 추천할 수 있게 되어 추천 예측력을 높일 수 있다.

총 1,628편의 영화를 18개 장르로 구분한 원본 데이터를 가지고 사전 군집 수를 정하여 클러스터링 하게 된다. 본 연구에서는 K-means 클러스터링의 보다 정확한 수행을 위하여, 군집의 수를 각각 4, 5, 6개로 변화시키며 실험을 수행하였으며 또한 계층적 클러스터링 방법을 통한 실험도 수행하였다. 군집의 수를 변화시킨 비계층적 방법과 계층적 방법을 통한 실험에서 얻어진 결과에 따라 군집의 수를 4개로 지정하여 K-means 클러스터링을 수행하였다.

데이터의 전처리(preprocessing)을 위하여 원본 데이터의 18개 장르 중에서 11개의 장르만을 실험 데이터로 사용하였다. 18개의 장르 중 특히 '다큐멘터리'나 '필름 느와르' 같은 장르들은 전체 영화 1,628편 중에서 1~2편 정도 밖에 없었기 때문에,

전체 영화들의 장르 속성을 대표할 수 없는 예외적 사례(outlier)들이다. 이러한 예외적 사례들은 모집단을 대표하지 못하고 있으며 실험을 왜곡시킬 수 있기 때문에(여운승, 2000; Hair et al. 1998), 전처리 과정에서 제거하여 11개의 장르만으로 실험을 수행한 것이다.

장르 클러스터링 수행 결과, 총 4개의 군집으로 분류되었다. 이를 정리하면 <표 1>과 같다:

다음으로는 이 영화들에 대하여 선호도가 비슷한 사용자별로 분류하기 위한 사용자별 클러스터링 분석을 실시하였다. 이를 위하여 각 군집마다 무작위로 5편씩 총 20편의 영화를 선정하고, 이 영화들에 대하여 등급을 부여한 131명의 사용자들을 클러스터링 하였다. 특정 장르의 영화만 집중 선호하는 사람이 아닌 모든 장르의 영화를 다양하게 본 사람을 선별하기 위하여, 각 영화마다 평가를 골고루 한 사용자를 선별하였다. 결측치(missing value)를 줄이기 위해서는, 20편의 영화 중 최소 15편 이상을 평가한 사람을 대상으로 정하였다. 이렇게 선정된 총 131명의 사용자에 대하여 사용자별 클러스터링 분석을 수행한 결과, 4개의 사용자 군집으로 분류되었다. 선호도의 평균값을 통하여 <표 2>와 같이 각 사용자 군집의 특징을 살펴볼 수 있다.

사용자 군집 1은 범죄적 요소가 가미된 무서운 영화(Th)와 로맨틱하고 서정적인 영화(Ro)를 선호하는 군집이고, 사용자 군집 2는 액션 성격의 스케일이 큰 영화(Ac)와 스릴이 있는 영화(Th)를 좋아하는 군집이다. 사용자 군집 3은 액션 성격의 영화(Ac)와 로맨틱한 서정적 영화(Ro)에는 비교적 낮은 평가를 보이는 군집이다. 사용자 군집 4는 다른 사용자 군집보다 모든 장르의 영화에 대하여 전체적으로 높은 평가를 보이고 있음을 알 수 있다.

한편, 사용자별 클러스터링 분석을 통하여 도출된 각 군집들이 서로 이질적인 집단인지 살펴보기 위하여 분산분석(ANOVA)을 수행하였다. 이에 대한 결과는 <표 3>에서 보는 바와 같이, 대부분의 영화에 대하여 군집 간에 유의한 차이가 있음을 알 수 있다.

<표 1> 장르 속성별 클러스터링 수행 결과

군집 1	액션 성격의 스케일이 큰 영화(Ac)
군집 2	범죄적인 요소가 있는 오싹하고 무서운 영화(Th)
군집 3	로맨틱한 서정적 드라마 영화(Ro)
군집 4	신나게 웃을 수 있는 코믹한 영화(Ch)

<표 2> 사용자 군집별 선호도 평균값

영화 군집	사용자 군집 1	사용자 군집 2	사용자 군집 3	사용자 군집 4
Ac	3.295	3.838	2.980	4.247
Th	3.951	3.952	3.861	4.265
Ro	3.820	3.478	3.097	3.967
Ch	3.048	3.700	3.355	4.088

<표 3> 분산분석 결과

변수 명		제곱합	자유도	제곱평균	F 통계량
Ac1	집단 간	39.781	3	11.100	37.811***
	집단 내	44.539	127	0.351	
	합계	84.320	130		
Ac2	집단 간	37.433	3	12.478	24.242***
	집단 내	65.370	127	0.515	
	합계	102.803	130		
Ac3	집단 간	34.116	3	11.372	15.190***
	집단 내	95.080	127	0.749	
	합계	129.196	130		
Ac4	집단 간	34.344	3	11.448	19.967***
	집단 내	72.814	127	0.573	
	합계	107.158	130		
Ac5	집단 간	26.731	3	8.910	22.662***
	집단 내	49.936	127	0.393	
	합계	76.667	130		
Th1	집단 간	2.399	3	0.800	0.858
	집단 내	118.401	127	0.932	
	합계	120.800	130		
Th2	집단 간	9.829	3	3.276	5.283***
	집단 내	78.760	127	0.620	
	합계	88.589	130		
Th3	집단 간	11.516	3	3.839	8.285***
	집단 내	58.839	127	0.463	
	합계	70.355	130		
Th4	집단 간	1.865	3	0.622	2.389*
	집단 내	33.048	127	0.260	
	합계	34.913	130		
Th5	집단 간	22.000	3	7.333	9.918***
	집단 내	93.900	127	0.739	
	합계	115.900	130		
Ro1	집단 간	8.797	3	2.932	3.298**
	집단 내	112.924	127	0.889	
	합계	121.721	130		
Ro2	집단 간	19.562	3	6.521	11.273***
	집단 내	73.464	127	0.578	
	합계	93.026	130		
Ro3	집단 간	50.715	3	16.905	23.077***
	집단 내	93.033	127	0.733	
	합계	143.748	130		

변수 명		제곱합	자유도	제곱평균	F 통계량
Ro4	집단 간	8.416	3	2.805	4.953***
	집단 내	71.939	127	0.566	
	합계	80.355	130		
Ro5	집단 간	3.548	3	1.183	1.703***
	집단 내	88.179	127	0.694	
	합계	91.727	130		
Ch1	집단 간	26.955	3	8.985	13.752***
	집단 내	82.977	127	0.653	
	합계	109.933	130		
Ch2	집단 간	12.163	3	4.054	5.487***
	집단 내	93.837	127	0.739	
	합계	106.000	130		
Ch3	집단 간	22.894	3	7.631	15.856***
	집단 내	61.122	127	0.481	
	합계	84.016	130		
Ch4	집단 간	17.186	3	5.729	8.037***
	집단 내	90.524	127	0.713	
	합계	107.710	130		
Ch5	집단 간	32.823	3	10.941	19.364***
	집단 내	71.758	127	0.565	
	합계	104.581	130		

***: $p \leq 0.01$, **: $p \leq 0.05$, *: $p \leq 0.10$

4.2.2 사례기반추론

사례기반추론을 위하여 앞에서 설명한 총 131명의 사용자를 실험 데이터로 사용하였다. 몇몇 사용자들은 1편에서 5편 정도의 영화에 등급을 부여하지 않은 경우도 있었으며, 이런 경우에는 다른 사용자들이 부여한 영화 등급의 평균값으로 결측치를 대체하였다. 사용자 131명의 데이터 중 80%는 모형의 구축을 위하여 학습용 데이터(training data)로 사용하고, 20%는 모형 검증을 위한 검증용 데이터(testing data)로 사용하였다. 유사도를 결정하는 방법으로는 최근접 이웃 추출법을 사용하였고, 한 번의 추론에 사용되는 이웃의 수는 5개로 결정하였다. 유사한 사례들이 예측한 값을 가중평균하여 추천하고자 하는 영화의 실제값과 비교하였다.

추천시스템의 협업 필터링에서 예측 성과를 측정하는 지표로서 MAE(Mean Absolute Error)나 MAPE(Mean Absolute Percent Error), RMSE (Root Mean Squared Error)가 가장 보편적으로 이용된다(Herlocker et al. 1999; Sarwar et al. 2001). 이 3가지 지표를 통하여 각 모델의 성과를 살펴보면 <표 4>와 같다. CBR_CF는 RMSE가 0.683, MAE가 0.493, MAPE가 12.90%이었다. UC_CBR_CF는 RMSE가 0.561, MAE가 0.429, MAPE가 10.78%로 전체적으로 CBR_CF보다 실제값과 예측값 간의 차이가 적음을 보이고 있다.

<표 4> 모델별 성과 비교

모델	RMSE	MAE	MAPE (%)
CBR_CF	0.683	0.493	12.90
UC_CBR_CF	0.561	0.429	10.78
UC_IC_CBR_CF	0.500	0.350	9.16

특히, 본 연구에서 제안한 모델인 UC_IC_CBR_CF는 RMSE, MAE, MAPE가 각각 0.500, 0.350, 9.16%로 CBR_CF와 UC_CBR_CF보다 모두 낮음을 나타내고 있다. 따라서, 제안된 모델(UC_IC_CBR_CF)이 기존의 두 모델(CBR_CF, UC_CBR_CF)보다 더 정확한 예측력을 지니고 있음을 보여준다.

각 모델 간 오차의 차이를 살펴보고 통계적 유의성을 검증하기 위하여 대응표본 t-검정(paired t-test)를 수행하였다. 오차는 다음과 같은 식에 의해 계산된다:

$$E_i = |R_i - P_i|$$

여기서, E는 각 모델의 오차를 나타내고 R은 실제의 선호도 값, P는 예측값이 된다. 대응표본 t-검정의 결과는 <표 5>와 같다. <표 5>에서 양의 t 값은 행측 모델의 오차가 열측 모델의 오차보다 작음을 의미한다. 즉, 행측 모델이 열측 모델보다 예측력이 우수하다는 것을 의미한다. 제안 모델인 UC_IC_CBR_CF와 CBR_CF 간의 t 값을 보면 5% 수준에서 유의하게 나타났고, UC_IC_CBR_CF와 UC_CBR_CF 간의 t 값도 5% 수준에서 유의하게 나타났다. 따라서, 제안된 모델이 기존의 다른 모델보다 통계적으로 우수함을 알 수 있다.

5. 결론

추천시스템과 협업 필터링에 관한 선행연구의 대부분은 높은 예측 성능을 지닌 예측 알고리즘에 초점을 두었다. 본 연구에서는 아이템이 단일값 속성이 아닌 다중값 속성을 지닌 점에 초점을 맞추어, 아이템의 속성을 활용한 협업 필터링 기법을 제안하였다. 또한, 아이템 속성별 클러스터링과 사용자별 클러스터링에 동시에 기반한 사례기반추론 방법을 제시하였다. 여러 장르의 특징을 지니고 있는 영화들을 K-means 클러스터링 기법으로 클러스터링 한 결과, 4개의 군집으로 분류되었다. 다음으로 사용자를 클러스터링 한 후 사용자의 소속 그룹 내에서 사례기반추론을 통하여 유사 선호도를 가진 이웃들의 평가를 토대로 영화를 추천하였

<표 5> 모델 간 대응표본 t-검정

	CBR_CF	UC_CBR_CF
UC_CBR_CF	t = 0.806	
UC_IC_CBR_CF	t = 1.875**	t = 1.784**

**: $p \leq 0.05$

다. 기존의 연구결과를 보면, 하이브리드 모델을 사용할 때 각 방법론들의 장점을 이용한 상승효과(synergy effect)를 통하여 모델의 설명력 제고(신택수, 홍태호, 2004), 데이터의 노이즈 제거(Shin and Han, 2001), 학습 표본의 선정(Shimodaira, 1996) 등의 효과가 나타나게 된다. 하지만, 이러한 하이브리드 모델을 개발할 때에는, 각 방법론의 장점들을 활용하여 새로운 효과가 기대되지 않을 때에는 방법론의 복잡성, 오류의 증폭 등의 어려움이 발생할 여지가 매우 높다. 본 연구에서는 사용자와 아이템이라는 2차원 상의 데이터를 동시에 클러스터링 함으로써 보다 정확한 클러스터링 방법을 제안하고자 하였다. 실증분석 결과, 제안된 방법이 기존의 클러스터링 없이 전체 사용자 기반의 사례기반추론 협업 필터링과 사용자 클러스터링 기반의 사례기반추론 협업 필터링 방법보다 MAE, MAPE, RMSE에서 더 우수한 성과를 나타내어 예측력이 더 뛰어남을 보였으며, 모델 간 대응표본 t-검정을 통해 성과 차이가 통계적으로 유의함을 검증하였다.

본 연구에서는, 속성별 클러스터링을 통하여 영화를 장르별로 분류하였다. 데이터를 가지고 영화 장르를 구분할 때부터 통합된 장르를 바탕으로 유사한 영화를 찾아 추천하는 알고리즘이 보다 효과적일 것으로 생각한 것이다. 추천시스템을 구축할 때에는 알고리즘적인 것, 즉 논리적인 계산으로 유사도를 찾아 고객에게 추천하는 것이 핵심일 것이다. 하지만, 표면적으로 표현된 장르뿐만 아니라, 실제로 영화의 내용을 알고 그 내용을 기반으로 하여 고객의 선호 패턴을 예측하여 추천을 한다면 고객에게 더욱 가깝게 다가갈 수 있을 것이다.

본 연구의 결과에 따르면, 영화나 상품을 추천할 때 사용자 간의 유사성뿐만 아니라 아이템 간의 유사성, 즉 아이템의 속성을 함께 고려함으로써 보다 정확하게 예측하여 추천할 수 있다. 본 연구에서는 아이템의 속성 중 장르 속성만으로 클러스터링 하였으나, 향후 연구에서는 보다 다양한 속성을 활용하여 클러스터링 기반의 사례기반추론을 이용한 추천시스템 연구가 필요하겠다. 또한, 제안된 모델을 MovieLens의 영화 데이터가 아닌 실제의 전자상거래 쇼핑몰 데이터에 적용할 필요성이 있다. 쇼핑몰의 상품 또한 영화의 장르처럼 다중값 속성을 지니고 있기 때문이다. 향후에는 웹 마이닝과 아이템 속성 기반의 클러스터링 협업 필터링 기법을 활용한 추천시스템에 관한 연구가 필요하겠다.

추천시스템을 개발할 때는 효과적인 알고리즘 이용과 동시에 특정 영화를 좋아할만한 고객 요구의 분석을 통하여 고객의 니즈를 충족시킬 수 있는 마케팅 전략이 무엇보다 필요하겠다. 고객들에게 상품이나 서비스를 추천함에 있어, 데이터마이닝 기법에 의존한 개인화된 상품, 서비스 추천보다는 고객이 원하는 정보에 맞춘 고객 맞춤화 상품이나 서비스를 추천하여야 하겠다(Nunes and Kambil, 2001). 고객들은 인공지능에 의하여 자동적으로 제공된 정보보다는 고객 선호도 등의 직접적인 정보가 반영된 맞춤 정보를 원하고 있다. 따라서, 고객에게 상품이나 서비스를 추천하고 고객관계를 형성함에 있어, 데이터마이닝 기법과 추천알고리즘에만 의존하기보다는 고객 개개인의 성향과 특성을 파악하는 진정한 고객관계관리, 일대일 마케팅 실현이 더욱 시급하다고 할 수 있다.

참고문헌

[1] 김재경, 안도현, 조윤호, "인터넷 쇼핑몰을 위한 데이터마이닝 기반 개인별 상품추천방법

론의 개발", *한국지능정보시스템학회논문지*, 9권 3호(2003), 177-191.

[2] 박지선, 김택헌, 류영석, 양성봉, "추천시스템을 위한 2-way 협동적 필터링 방법을 이용한 예측알고리즘", *정보과학회지*, 29권 9, 10호(2002), 669-675.

[3] 신택수, 홍태호, "인공신경망과 로짓모형을 통합한 부실확률맵기반 신용등급화에 관한 연구", *회계저널*, 13권 3호(2004), 7-26.

[4] 여운승, *다변량 행동조사*, 민영사, 2000.

[5] Adomavicius, G., and A. Tuzhilin, "Using Data Mining Methods to Build Customer Profiles," *IEEE Computer*, Vol. 34, No. 2(2001), 74-82.

[6] Berson, A., K. Smith, and K. Thearing, *Building Data Mining Applications for CRM*, McGraw-Hill, New York, 2000.

[7] Changchien, S.W., and T.-C. Lu, "Mining Association Rules Procedure to Support On-Line Recommendation by Customers and Products Fragmentation," *Expert Systems with Applications*, Vol. 20, No. 4(2001), 325-335.

[8] Hair, J.F., R.E. Anderson, R.L. Tatham, and W.C. Black, *Multivariate Data Analysis*, 5th Edition, Prentice-Hall, Inc., New Jersey, 1998.

[9] Herlocker, J., J.A. Konstan, A. Borchers, and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering," *Proceedings of the 1999 Conference on Research and Development in Information Retrieval*(1999).

[10] Kim, J.K., Y.H. Cho, W.J. Kim, J.R. Kim, and J.H. Suh, "A Personalized Recommendation Procedure for Internet Shopping Support," *Electronic Commerce Research and Applications*, Vol. 1, No. 3/4(2002), 301-313.

[11] Kim, T.-H., Y.-S. Ryu, S.-I. Park, and S.-B. Yang, "An Improved Recommendation Algorithm in Collaborative Filtering," *Lecture Notes in Computer Science*, No. 2455(2002), 254-261.

[12] Konstan, J.A., B. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, and J. Riedl, "GroupLes: Applying Collaborative Filtering to Usenet News," *Communications of the ACM*, Vol. 40, No. 3(1997), 77-87.

[13] Kuo, Y.-F., and L.-S. Chen, "Personalization Technology Application to Internet Content Provider," *Expert Systems with Applications*, Vol. 21, No. 4(2001), 203-215.

[14] Li, Q., and B.M. Kim, "Clustering Approach for Hybrid Recommender System," *Proceedings of IEEE/WIC International Conference on Web Intelligence*(2003), 33-39.

[15] Mild, A., and M. Natter, "Collaborative Filtering or Regression Models for Internet Recommendation Systems?" *Journal of Targeting, Measurement and Analysis of Marketing*, Vol.10, No. 4(2002), 304-313.

[16] Mobasher, B., H. Dai, T. Luo, and M. Nakagawa, "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization," *Data Mining and Knowledge Discovery*, Vol. 6, No.1(2002), 61-82.

[17] Nunes, P.F., and A. Kambil, "Personalization? No Thanks," *Harvard Business Review*, Vol. 79, No. 4(2001).

[18] Roh, T.H., K.J. Oh, and I. Han, "The Collaborative Filtering Recommendation Based on SOM Cluster-Indexing CBR," *Expert Systems with Applications*, Vol. 25, No. 3(2003), 413-423.

[19] Sarwar, B.M., G. Karypis, J.A. Konstan, and J. Riedl, "Analysis of Recommendation Algorithms for e-Commerce," *Proceedings of the ACM E-Commerce 2000 Conference* (2000), 158-167.

[20] Sarwar, B.M., G. Karypis, J.A. Konstan, and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," *Proceedings of the 10th International World Wide Web Confernece*(2001), 285-295.

[21] Schafer, J.B., J.A. Konstan, and J. Riedl, "E-Commerce Recommendation Applications," *Journal of Data Mining and Knowledge Discovery*, Vol. 5, No. 1(2001), 115-152.

[22] Schafer, J.B., J.A. Konstan, and J. Riedl, "Recommender Systems in e-Commerce," *Proceedings of the ACM Confernece on Electronic Commerce*(1999).

[23] Shimodaira, H., "A Method for Selecting Similar Learning Data in the Prediction of Time Series Using Neural Networks," *Expert Systems with Applications*, Vol. 10, No. 34(1996), 429-434.

[24] Shin, K.S., and I. Han, "A Case-Based Approach Using Inductive Indexing for Corporate Bond Rating," *Decision Support Systems*, Vol. 32, No. 1(2001), 41-52.

[25] Weng, S.-S., and M.-J. Liu, "Feature-Based Recommendations for One-to-One Marketing," *Expert Systems with Applications*, Vol. 26, No. 4(2004), 493-508.

[26] Yuan, S., and W. Chang, "Mixed Initiative Synthesized Learning Approach for Web Based CRM," *Expert Systems with Applications*, Vol. 20, No. 2(2001), 187-200.

Abstract

A Web Personalized Recommender SystemsUsing Clustering-based CBR

Taeho Hong* · Hee-Jung Lee** · Bomil Suh***

Recently, many researches on recommendation systems and collaborative filtering have been proceeding in both research and practice. However, although product items may have multi-valued attributes, previous studies did not reflect the multi-valued attributes. To overcome this limitation, this paper proposes new methodology for recommendation system. The proposed methodology uses multi-valued attributes based on clustering technique for items and applies the collaborative filtering to provide accurate recommendations. In the proposed methodology, both user clustering-based CBR and item attribute clustering-based CBR technique have been applied to the collaborative filtering to consider correlation of item to item as well as correlation of user to user. By using multi-valued attribute-based clustering technique for items, characteristics of items are identified clearly. Extensive experiments have been performed with MovieLens data to validate the proposed methodology. The results of the experiment show that the proposed methodology outperforms the benchmarked methodologies: Case Based Reasoning Collaborative Filtering (CBR_CF) and User Clustering Case Based Reasoning Collaborative Filtering (UC_CBR_CF).

Key words : 추천시스템, 협업 필터링, 개인화, 클러스터링, 사례기반추론

* Pusan National University, College of Business, Division of Business Administration
** Pusan National University, College of Business, Division of Business Administration
*** Sookmyung Women's University, College of Economics & Business Administration, Division of Business Administration