ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2014 June: 20(2): 73~92 http://dx.doi.org/10.13088/jiis.2014.20.2.073

적응형 군집화 기반 확장 용이한 협업 필터링 기법

이오준
단국대학교 소프트웨어학과
(*concerto_grs@naver.com*)

홍민성
단국대학교 컴퓨터학과
(*hold_time@naver.com*)

이원진
단국대학교 미디어콘텐츠연구원
(*god7300@dankook.ac.kr*)

이재동
단국대학교 소프트웨어학과
(*letsdoit@dankook.ac.kr*)

기존 협업 필터링 기법은 사용자들의 아이템에 대한 선호도를 기반으로 유사 아이템 집합 또는 유사 사용자 집합을 구성하고, 이를 이용해 예측된 사용자의 특정 아이템에 대한 선호도를 기반으로 추천을 수행한다. 이로 인해, 사용자 선호도 정보가 부족하게 되면, 유사 아이템 사용자 집합의 신뢰도가 낮아지고, 추천 서비스의 신뢰도 또한 따라서 낮아진다. 또한, 서비스의 규모가 커질수록, 유사 아이템, 사용자 집합의 생성에 걸리는 시간은 기하급수적으로 증가하고 추천 서비스의 응답시간 또한 그에 따라 증가하게 된다. 위와 같은 문제점을 해결하기 위해 본 논문에서는 적응형 군집화 기법을 제안하고 이를 적용한 협업 필터링 기법을 제안하고 있다. 이 기법은 크게 네 가지 방법으로 이루어진다. 첫째, 사용자와 아이템의 특성 벡터를 기반으로 사용자와 아이템 각각을 군집화 하여, 기존 협업 필터링 기법에서 유사 아이템, 사용자 집합을 생성하는데 소요되는 시간을 절약하며, 사용자 선호도 정보만을 이용한 부분 집합 생성보다 추천의 신뢰도를 높이고, 초기 평가 문제와 초기 이용자 문제를 일부 해소한다. 둘째, 미리 구성된 사용자와 아이템의 군집을 기반으로 군집간의 선호도를 이용해 추천을 수행한다. 사용자가 속한 군집의 선호도가 높은 순서대로 아이템 군집을 조회하여 사용자에게 제공할 아이템 목록을 구성하여, 추천 시스템의 부하 대부분을 모델 생성 단계에서 부담하고 실제 수행 시 부하를 최소화한다. 셋째, 누락된 사용자 선호도 정보를 사용자와 아이템 군집을 이용하여 예측함으로써 협업 필터링 추천 기법의 사용자 선호도 정보 희박성으로 인한 문제를 해소한다. 넷째, 사용자와 아이템의 특성 벡터를 사용자의 피드백에 따라 학습시켜 아이템과 사용자의 정성적 특성 정량화의 어려움을 해결한다. 본 연구의 검증은 기존에 제안되었던 하이브리드 필터링 기법들과의 성능 비교를 통해 이루어졌으며, 평가 방법으로는 평균 절대 오차와 응답 시간을 이용하였다.

주제어 : 추천 시스템, 적응형 시스템, 협업 필터링, 하이브리드 필터링, 군집화

논문접수일 : 2014년 4월 5일 논문수정일 : 2014년 5월 14일 게재확정일 : 2014년 5월 16일
투고유형 : 국문일반 교신저자 : 이재동

1. 개요

인터넷의 진화는 결과적으로 사용자가 접하는 상품과 콘텐츠의 폭발적인 증가를 가져왔다. 방대한 양의 상품과 콘텐츠로 인해, 사용자들은 오히려 원하는 상품이나 아이템을 찾거나 구매하기가 더욱 어려워졌다. 이에 따라 사용자들에게 적절한 상품을 제공하고 사용자들의 의사 결정을 지원하는 추천 시스템의 중요성은 점점 더 높아지고 있다. 사용자의 입맛에 맞는 상품을 제공

* 본 논문은 문화체육관광부에서 지원하는 2013년도 콘텐츠산업기술지원사업 (P2013040019)의 연구 수행으로 인한 결과물임을 밝힙니다.

하는 개인화된 맞춤형 추천 시스템은 사용자의 만족도와 고객 충성도를 향상 시킬 수 있을 뿐만 아니라, 전자 소매상들의 이윤을 크게 증가시킨다(Adomavicius and Tuzhilin, 2005). 아마존(Linden et. al., 2003), 구글(Das et. al., 2003), 넷플릭스(Bennet and Lanning, 2007) , 티보(Ali and Stam, 2004) 그리고 야후(Park and Pennock, 2007)와 같은 인터넷 시장의 선도 기업들은 개인화된 추천이 가능한 맞춤형 상품 추천 엔진을 이미 운용하고 있다.

추천 시스템을 구축하기 위한 여러 기법 중 널리 사용되는 기법으로는 협업 필터링 추천 기법이 있다. 협업 필터링 추천 기법은 다른 사용자의 선호도 정보를 기반으로 아이템을 추천하는 기법이다. 따라서 해당 특정 도메인에 대한 지식을 필요로 하지 않으며, 사용자와 아이템에 대한 광범위한 데이터를 필요로 하지도 않는다. 또한, 사용자 프로파일과 아이템 특성 정보의 정확도의 한계로 인해 발생하는 문제로부터 자유롭다. 하지만, 초기 평가 문제, 초기 사용자 문제, 확장성 문제, 행렬 희박성 문제와 같은 근본적인 한계점 또한 가지고 있다. 이 문제들을 해결하기 위해 아래와 같은 다양한 기법들이 제안되었으나, 이 문제들을 모두 해결하지는 못했다.

Shepitsen, A.(Shepitsen et. al., R, 2008)와 Sanghwa Kim(Kim et. al., 2012)는 초기 평가, 초기 사용자 문제를 개선하기 위해 부가적인 정보를 활용하는 방법을 제안하였다. Shepitsen, A.는 소셜 태그 정보를 바탕으로 사용자를 계층적으로 군집화하는 방법을 제안하였다. 이 방법은 확장성 문제 또한 어느 정도 해소할 수 있는 방법이지만 특정한 형태의 서비스에서만 적용할 수 있다는 문제가 있다. Kim, Sanghwa은 콘텐츠 특성 정보를 사용자 선호도 예측에 활용하는 방법을 제안하였으나, 콘텐츠 특정 정보 정확도의 한계를 해결하지 못하였고 초기 사용자 문제에 대해서는 해결책을 제시하지 못했다.

Park, Kyusik(Park et. al., 2010)과 Shen, Yan (Shen et. al., 2012), Kim, Su-Yeon(Kim et. al., 2012)은 초기 평가, 초기 사용자 문제의 개선을 위해 내용 기반 필터링 추천 기법을 병용하는 방법을 제안하였다. Park, Kyusik은 내용 기반 필터링 추천 기법과 협업 필터링 추천 기법을 조합한 실험을 통해 가중치 값을 결정하여 추천의 성능을 향상시키는 방법을 제안하였다. 또 Shen, Yan은 내용 기반 필터링 추천 기법과 협업 필터링 추천 기법을 조합하고 강화 학습 알고리즘을 이용하여 사용자에게 최적화된 가중치를 적용하는 방법을 제안하였다. Kim, Su-Yeon은 계층분석과정을 이용해 사용자 프로파일을 분석하여 가중치를 설정하고, 이 가중치를 이용해 내용 기반 필터링과 협업 필터링의 선호도 예측 결과를 조합하는 방법을 제안하였다. 이 방법들은 초기 평가, 초기 사용자 문제는 어느 정도 해소할 수 있으나, 시간 복잡도를 높여 확장성 문제를 심화시킨다. 또한, 내용 기반 필터링 추천 기법의 한계점인 정성적 특성의 정량화에 대한 해결책을 제시하지 못한다.

Kim, Yong(Kim and Moon, 2006)는 규칙 기반 필터링 추천 기법과 협업 필터링 추천 기법을 병용하는 방법을 제안하였다. 이 방법은 초기 평가, 초기 사용자 문제를 어느 정도 개선할 수 있다. 하지만, 협업 필터링의 결과를 규칙 기반 필터링을 이용해 보정하는 정도에 그쳐, 서비스 초기 사용자의 이용내역이 부족할 경우의 추천 시스템 성능 저하에 대처하지 못한다는 문제점을 갖는다.

Ji, Hao(Ji et. al., 2013)는 행렬 희박성 문제를

개선하기 위해 사용자 선호도 예측 시, 사용자 기반 협업 필터링과 아이템 기반 협업 필터링을 병용하는 방법을 제안하고 실험을 통해 두 예측 값을 결합하기 위한 가중치를 결정하였다. 이 방법은 사용자 선호도 예측의 신뢰도를 향상 시킬 수 있지만, 가중치 값이 추천 모델의 상태를 반영하지 못한다는 문제점을 갖으며, 시간복잡도를 높여 확장성 문제를 더욱 심화시킨다.

Braak(Braak et. al., 2009)은 사용자 프로파일을 군집화하여 추천의 성능을 향상시키는 방법을 제안하였고, Xue(Xue et. al., 2005)는 아이템 군집화 방법을, George, Thomas(George and Merugu, 2005)는 사용자와 아이템 모두를 군집화 하는 방법을 제안하였다. 하지만 위 방법들은 확장성 문제를 개선할 수 있지만, 사용자 프로파일 정확도의 한계로 인한 문제를 해결하지 못하고 있다.

본 연구에서는 위와 같은 문제점들을 개선하기 위해 적응형 군집화 기법과 이를 이용한 확장 용이한 협업 필터링 기법을 제안한다. 이 기법은 크게 네 부분으로 이뤄진다. 첫째, 사용자 특성과 아이템 특성을 바탕으로 사용자와 아이템을 군집화하고 군집 간 선호도를 구하며 둘째, 사용자 선호도 정보와 군집, 군집간 선호도를 이용해 추천을 수행하고 셋째, 이 군집을 바탕으로 사용자의 선호도를 예측하며 넷째, 추천에 대한 사용자 피드백을 사용자와 아이템의 특성 벡터에 반영하여, 높은 추천 신뢰도를 가지며 서비스 규모의 변화에도 일정한 성능을 보일 수 있는 협업 필터링 기법을 제안한다.

이를 위해 2장에서는 협업 필터링 추천 기법에 관한 기존 연구들을 살펴보고 3장에서는 본 연구에서 제안하는 적응형 군집화 기반 협업 필터링 기법에 대해 상세히 기술한다. 4장에서는 제안된 기법을 바탕으로 구현된 시스템에 대해 검증하고 평가하며 5장에서는 결론 및 향후 연구 방향을 제시하고자 한다.

2. 관련 연구

추천 시스템은 “사람들이 추천을 제공하며 시스템이 통합하여 적당한 사람에게 보여준다.”고 정의할 수 있다(Renick and Varian, 1997). 추천 시스템은 사용자에게 사용자가 선호하는 아이템을 제공해주는 시스템으로써, 인터넷의 음악, 영화 사이트의 콘텐츠 추천, 인터넷 쇼핑몰에서의 상품 추천 등에 널리 이용되고 있다. 인터넷이 발달하면서 접할 수 있는 정보 및 아이템의 수와 종류도 증가하고 있어 수많은 아이템 중에 사용자가 선호할 만한 아이템을 자동으로 추천해주는 시스템도 함께 발전한 것이다.

판매자나 구매자 모두 추천 시스템으로 인하여 부가가치를 얻을 수 있다. 판매자의 입장에서는 대량 맞춤을 통하여 일대일 마케팅이 가능하고 웹 개인화 서비스가 가능하여 고객의 충성도를 높일 수 있고 또한 고객의 입장에서는 현재 시장 정보 과부화 현상을 완화하여 자신이 원하는 콘텐츠나 정보를 더욱 쉽게 획득할 수 있다.

추천 시스템의 초기 방식은 사용자의 사용 순위 중에서 상위에 위치한 아이템 혹은 조회수가 높은 아이템을 추천해주는 방식이었다. 그러나 이러한 방식은 개개인의 취향이나 특성을 고려하지 않았기 때문에 추천에 대한 사용자들의 만족도가 높지 않았다. 이러한 문제점을 개선하고자 다양한 형태의 추천 시스템들이 연구되었으며, 대표적으로는 협업 필터링 추천 기법이 있다.

협업 필터링 추천 기법(Connor and Herlocker, 1999; Groh and Ehming, 2007; Herlocker and Konstan, 2002; Huang et. al., 2004)은 사용자들의 선호도 정보를 기반으로 새로운 사용자가 관심을 가질 것으로 예측되는 항목을 추천해주는 기법이다. 규칙 기반 필터링 추천 기법이나 내용 기반 필터링 추천 기법 등의 방법이 항목 자체의 속성 정보를 이용해서 사용자에게 아이템을 추천하는 것과는 달리 협업 필터링 추천 기법은 아이템에 대한 다른 사용자들의 선호를 기반으로 한다. 이전의 추천 시스템들이 대부분 텍스트 기반의 자료를 대상으로 하였으나 협업 필터링 추천 기법은 다양한 멀티미디어 아이템을 추천하기 위한 많은 시도들에 이용되었다. 이는 아이템의 특성을 기반으로 하는 여타 추천 기법들을 멀티미디어에 적용하기에는 멀티미디어 콘텐츠의 정성적인 특성을 정량적으로 수치화하거나, 추출하는데 어려움이 있기 때문이다.

이러한 협업 필터링 추천 기법의 장점은 다음과 같다. 첫째, 협업 필터링 추천 기법을 사용해서 항목의 선호도를 예측할 때 예측 대상이 되는 아이템이 다양한 분야에 속해있다고 가정하면, 사용자 A가 선호하는 분야와 같은 분야의 아이템을 선호했던 이웃이 새로운 분야를 선호하였을 경우 특정 사용자에게 새로운 분야의 아이템을 추천할 수 있다. 둘째, 협업 필터링 추천 기법은 동질적인 선호도를 가진 사용자 집단에서 비교적 정확한 예측이 가능하다. 셋째, 협업 필터링 추천 기법은 많은 계산량을 요구하지 않는 중소규모의 환경에서 실시간 추천을 위해서 고안되었기 때문에 개인화된 추천을 위한 계산은 복잡하지 않고 속도도 빠르다.

그러나 협업 필터링 추천 기법은 초기 평가, 초기 사용자 문제, 확장성 문제, 행렬의 희박성 문제 등의 제약점을 갖고 있다(Konstan et. al., 1997; Sarwar et. al., 2001; Huang et. al., 2004). 첫째, 초기 평가, 초기 사용자 문제는 평가를 전혀 하지 않은 새로운 사용자에게 아이템을 추천하거나 전혀 평가되지 않은 아이템을 추천하는 것이 어렵다는 것이다. 대안으로 충분한 선호도 정보가 쌓일 때까지 사용자 프로파일의 속성 정보나 아이템의 특성 정보를 이용한 내용 기반 추천 기법을 병행하는 방법들이 연구 되어왔다(Melville et. al., 2002). 둘째로, 확장성 문제는 협업 필터링 추천 기법에서 사용자와 아이템의 수가 증가할수록 유사 사용자 혹은 유사 아이템 집합을 생성하는데 필요한 연산 비용도 비례하여 증가한다는 것이다. 이는 실시간 추천을 어렵게 만드는 요인이 되며, 이를 개선하기 위한 방법 중 하나가 군집화 기법을 적용하는 것이다(Melville et. al, 2002). 셋째로, 행렬의 희박성 문제는 일반적으로 고객들은 상품 평가를 잘 하지 않는 경향이 있으며, 선호도 행렬이 희박할 경우 추천의 신뢰도를 떨어뜨릴 수 있다는 것이다. 특히 행렬이 희박할 경우, 몇몇 사용자들에 의해서만 높게 평가된 아이템들에 대해서는 유사한 아이템을 찾기 힘들기 때문에 추천이 어려울 수 있다(Pazzani, 1999). 희박성 문제를 개선하기 위해 Pazzani, M.은 음식점 추천 시스템에서 성별, 나이, 지역 코드, 학교 직장 등의 인적 정보를 사용하였으며, Huang, Z.(Huang et. al., 2004)는 사용자간 트랜지티브 연관성을 사용자 탐색 방법에 이용하는 방법을 제안하였다. 또한, Ji, Hao(Ji et. al., 2013)는 사용자 기반 선호도 예측 기법과 아이템 기반 선호도 예측 기법을 병용하여 선호도 예측의 정확도를 높이는 방법을 제안하였다.

3. 적응형 군집화 기반 협업 필터링 기법

본 연구에서는 내용 기반 필터링 추천 기법의 방법론을 도입하고 모델 상황에 능동적으로 적응하는 “적응형 군집화 기법”을 적용한 “적응형 군집화 기반 협업 필터링 기법”을 제안한다. 이 기법은 추천 서비스의 신뢰도를 높이고 시간 복잡도를 낮춰, 규모 변화에 대한 추천 신뢰도와 응답시간의 변화의 폭을 줄이고 서비스 초기 상황에서의 추천 신뢰도를 개선한 확장 용이한 협업 필터링 기법이다.

이 기법은 적응형 군집화 기법을 이용하여 추천 모델을 생성하는 과정과 생성된 추천 모델을 이용해 추천을 수행하는 부분으로 나뉜다. 적응형 군집화 기법은 높은 시간 복잡도로 인해 상황의 변화를 실시간으로 반영하기 어려운 군집화 기법을 개선한 것이다. 이는 두 가지 방법으로 이루어진다. 첫째는 군집화 과정에서 사용자의 선호도를 반영하는 것이며, 둘째는 사용자의 피드백에 따라 각 아이템과 사용자의 군집을 재배치하는 것이다.

이를 위해, 본 연구에서는 네 가지 방법을 사용한다. 첫째, 사용자와 아이템의 특성 벡터를 기반으로 사용자와 아이템 각각을 군집화 한다. 이는 기존 협업 필터링 추천 기법에서 사용자나 아이템의 부분 집합을 생성하는데 소요되는 시간을 절약하며 사용자 선호도 정보만을 이용한 부분 집합 생성보다 추천의 신뢰도를 높일 수 있고 초기 평가 문제와 초기 사용자 문제를 일부 해소할 수 있다. 둘째, 미리 구성된 사용자와 아이템의 군집을 기반으로 군집간의 선호도를 통해 추천을 수행한다. 사용자가 속한 군집의 선호도가 높은 순서대로 아이템 군집을 조회하여 사용자에게 제공할 아이템 목록을 구성하고, 순위를 매긴다. 이를 통해, 추천 시스템의 부하 대부분을 모델 생성 단계에서 부담하고 실제 수행 시 부하를 최소화한다. 셋째, 누락된 사용자 선호도 정보를 사용자와 아이템 군집을 이용하여 예측함으로써 협업 필터링 추천 기법의 사용자 선호도 정보 희소성으로 인한 문제를 해소한다. 또한 이는 선호도 예측 과정의 시간 복잡도를 크게 낮춰 수행 시간 중에 누락된 선호도를 예측하는 것을 가능하게 한다. 넷째, 사용자와 아이템의 특성 벡터를 사용자의 피드백에 따라 학습시켜 나감으로써 아이템과 사용자의 정성적 특성의 정량화의 어려움을 해결한다. <Table 1>은 적응형

〈Table 1〉 Algorithm of Adaptive Collaborative Filtering

Algorithm of Adaptive Collaborative Filtering
Creation of Recommendation Model (1) Cluster users and items by feature vectors of users and items. (2) Estimate Inter-Cluster Preference of user cluster to item-cluster according to preference of each user of each user-cluster. Missing User Preference Point Prediction (1) Create User Preference Matrix. (2) Predict Missing User Preference Points. Execution of Recommendation (1) Choose k item-clusters which have high inter-cluster preference of user-cluster including user requesting service. (2) Recommend items which have high preference of user in selected item-clusters. Applying User Feedback (1) Apply User Preference in user feedback to user preference. (2) Make user and item feature vector trained by using user feedback.

군집화 기반 협업 필터링 기법의 전체 알고리즘의 요약이다.

아래에서부터는 적응형 군집화 기반 협업 필터링 기법의 세부적인 알고리즘을 자세히 기술한다.

3.1. 적응형 군집화 기반 협업 필터링 추천 모델

이 절에서는 적응형 군집화 기반 협업 필터링 기법의 추천 모델을 생성하는 방법에 대해 기술한다. 이 모델은 추천 시스템이 사용자에게 아이템을 추천하는 기준이 되며, 사용자의 피드백에 따라 지속적으로 변화하며 일정한 주기에 따라 재 생성된다. <Table 2>는 적응형 군집화 기반 협업 필터링 추천 모델 생성 알고리즘을 상세히 기술한 것이다.

〈Table 2〉 Algorithm of Creating Recommendation Model

Algorithm of Creating Recommendation Model

1. Clustering Users and Items
(1) Decision of Number of Cluster, k
For → i=0 → I = 10+log(*Number of User or Item*)
Make users clustered into user-clusters ($Model_i$).
Calculate BIC(Bayesian Information Criteria) of Model according to i.
i is Number of Cluster, k when BIC of $Model_i$ is larger than $Model_{i-1}$.
End For
(2) Execution of Clustering
k_u : Number of User Cluster
k_c : Number of Item Cluster

Make users clustered into k_u user-clusters.
Make items clustered into k_c item-clusters.
2. Estimation of Inter-Cluster Preference
$CP_{i,j}$: Inter-Cluster Preference of User-Cluster, i to Item-Cluster j

For I = 1 → i=k_u (*Number of User Cluster*)
For j = 1 → j = k_c (*Number of Item Cluster*)

$$CP_{i,j} = \sum_{l}^{N_{uc_l}} \sum_{n}^{N_{cc_j}} W_{UC_i,u_l} \times W_{CC_j,c_n} \times R_{u_L c_n}$$

<Formula 1>

End For
End For

적응형 군집화 기반 협업 필터링 추천 모델은 크게 사용자와 아이템의 군집 그리고 사용자 군집과 아이템 군집 간의 선호도 두 가지 부분으로 이루어지며, 따라서 그 세부 알고리즘 또한 군집 생성 과정과 군집간 선호도를 구하는 두 부분으로 나뉜다.

3.1.1 적응형 군집화 기법

군집화는 군집의 수를 결정하고 군집화를 수행하는 크게 두 과정을 나뉜다. 군집의 수는 군집화를 수 차례 반복하며 모델의 상황을 반영한 최적의 값으로 결정된다. 결정 기준은 베이지안 정보 기준(Bayesian Information Criteria)(Kass and Wasserman, 1995; Schwarz, 1987)을 이용한다. 이를 구하는 방법은 <Formula 2>와 같다.

$$(\text{베이지안 정보 기준})(M_k) = BIC(M_k) = \hat{l}(X) - \frac{n_p}{2} \log N$$

〈Formula 2〉

여기서, M_k는 군집의 수가 k인 군집 모델을 나타내며, X는 모델이 갖는 전체 원소를 가리킨다. 또, N은 모델이 갖는 전체 원소의 수를, n_p는 각 원소가 갖는 파라미터의 수를, $\hat{l}$은 현재 모델의 log-우도 값을 나타낸다. log-우도는 주어진 원소들을 가장 잘 표현하는 모델에서 최댓값을 갖게

되는데, 이 값은 집단의 개수에 비례하며 N개의 집단으로 구성된 모델이 N개의 원소를 나타내는 경우 log-우도는 최댓값 0을 갖게 된다. 따라서 과적합화를 막기 위해 <Formula 2>의 두 번째 항과 같은 벌칙항이 필요하다. 이 항은 모델의 복잡도를 나타내는 항으로 모델의 복잡도가 증가함에 따라 그 값이 증가한다. 따라, 베이지안 정보 기준은 우도와 모델 복잡도를 고려하여 최적의 모델을 찾게 된다. log-우도는 <Formula 3>과 같이 구할 수 있다(Heo and Woo, 2008).

$$\hat{l}(C_i) = \sum_{j \in C_i} \log \hat{p}(x_j)$$

$$= \sum_{j \in C_i} \left[\log \frac{N_i}{N} - \frac{d}{2} \log 2\pi - \frac{1}{2} \log |\widehat{\Sigma_i}| - \frac{1}{2}\, trace\,(x_{j-\mu_{(j)}})^T \widehat{\Sigma^{-1}} (x_{j-\mu_{(j)}}) \right]$$

$$= N_i \log N_i - N_i \log N - \frac{dN_i}{2} \log 2\pi - \frac{N_i}{2} \log |\widehat{\Sigma_i}| - \frac{(N_i - 1)d}{2}$$

〈Formula 3〉

여기서, N_i는 i번째 군집에 속하는 원소의 개수를 나타내며, C_i는 i번째 군집을 나타내고, μ_j와 Σ_j는 각각 원소 j가 속한 군집의 평균과 분산을 나타낸다.

군집화 과정은 기대치최대화 알고리즘을 기반으로 가우시안-베이지안 확률 모델과 대상 원소 간 유사도를 이용한다. 기대치최대화 알고리즘은 원소가 속한 군집을 지역적으로 판단하는 k-평균 알고리즘과 달리 전역적으로 판단하기 때문에 원소가 속한 군집을 더욱 정확히 판별할 수 있다. 이는 인접한 군집들에 대한 원소의 군집 판별에서 더욱 명확히 드러난다. <Table 3>은 사용자와 아이템 군집화 과정의 세부 알고리즘이다.

〈Table 3〉 Algorithm of Clustering User and Item

Algorithm of Clustering User and Item

C_i: Center of Cluster i
μ_i: Average of Cluster i
Σ_i: Deviation of Cluster i

While, $\mu(Inter\ Cluster\ Deviation) < \alpha$
For $j = 1 \rightarrow j = (Number\ of\ Element)$
For $i = 1 \rightarrow i = k(Number\ of\ Cluster)$

$$P(C_i|x_j) = P(x_j|C_i)\,P(C_i)/\ P(x_j) \cong \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i-\mu_i)^2}{2\sigma^2}} P(C_i)$$

<Formula 4>

For $x_l \in C_i$,

$$L_{C_i x_j} = P(C_i|x_j) \sum_{1}^{N_{C_i}} \frac{R_j \cdot R_l}{\| R_j \| \| R_l \|}$$

$$= P(C_i|x_j) \sum_{1}^{N_{C_i}} \frac{\sum_{k=1}^{R_j \cap R_l} R_{j,i} \times R_{l,i}}{\sqrt{\sum_{k=1}^{R_j \cap R_l} (R_{j,i})^2} \times \sqrt{\sum_{k=1}^{R_j \cap R_l} (R_{l,i})^2}}$$

<Formula 5>

End For
$x_j \in C_i$, when $L_{c_i x_j}$ has maximum value.
End For
Following the changed model, calculate each value of c_i, μ_i, σ_i, and $P(C_i)$ of each clusters.
End For
End While

여기서, μ_i는 i번째 군집 C_i의 평균을, c_i는 군집의 중심을, σ_i는 군집의 표준 편차를, $P(C_i)$는 원소가 해당 군집에 속할 확률을 나타낸다. R_j는 원소 j의 선호도 벡터를 나타내며, $R_{j,i}$는 원소 j의 i에 대한 선호도를 나타낸다. <Formula 4>는 가우시안-베이지안 확률 모델을 통해 원소 x_j가 군집 C_i에 속할 확률을 추청하는 과정이며, 이 수식의 결과 값인 $P(C_i|x_j)$는 위 확률의 근사치이다. 또한 <Formula 5>는 위 알고리즘에서 사용하

는 기대치최대화 알고리즘의 최대 우도추정치를 구하는 과정으로 $P(C_i|x_j)_j$에 원소 x와 군집 C_i 내부의 원소들 간의 유사도의 합을 곱하여 최대 우도추정치로 사용한다. 이러한 방법을 사용하는 것은 정확도를 사용자 프로파일링 과정이나 아이템 특성 추출 과정에 의존하는 특성 벡터에 비해 사용자의 서비스 이용에 따라 지속적으로 정보를 축적하고 변화하는 선호도 정보가 원소의 특성을 더 정확히 표현할 수 있기 때문이다 (Baeza-Yates and Ribeiro-Neto, 1999).

3.1.2 군집간 선호도 추정

군집간 선호도 추정은 본 논문에서 제안하는 기법을 확장 용이성을 보장하는데 주요한 역할을 하는 부분이다. 이 방법을 통해 특정 사용자 군집에 대해 추천할 아이템 군집의 순위를 얻을 수 있다. 따라서, 사용자로부터 서비스 요청이 들어올 경우, 사용자의 군집에 따라 추천할 아이템의 군집이 순위화 되어 있기 때문에 단순히 군집간 선호도가 높은 군집부터 예측된 사용자의 선호도가 낮은 아이템을 필터링하는 것만으로 아이템 추천을 수행할 수 있다. 이 방법은 유사 아이템, 사용자 집합을 생성하지 않아 시간 복잡도를 크게 낮출 수 있고, 사용자에 따른 아이템 제공 범위를 한정하여, 대규모 서비스에서도 응답시간을 유지하고 동일한 추천 신뢰도를 유지할 수 있도록 한다.

이 과정의 알고리즘은 <Formula 1>과 같다. 여기서 W_{UC_i,u_l}은 사용자 u_l의 사용자 군집 UC_i에 대한 우도를 나타내며, W_{CC_j,c_n}은 아이템 c_n의 아이템 군집 CC_j에 대한 우도를 나타내고, R_{u_l,c_n}은 사용자 u_l의 아이템 c_n에 대한 선호도를 나타낸다. $CP_{i,j}$는 i번째 사용자 군집의 j번째 아이템 군집에 대한 선호도를 나타낸다. 사용자 군집의 아이템 군집에 대한 선호도는 사용자 군집의 원소의 사용자 군집에 대한 우도와 아이템 군집의 원소의 아이템 군집에 대한 우도, 사용자 군집의 원소와 아이템 군집의 원소 간의 선호도를 곱한 값의 총합으로 구한다.

3.1.3 사용자 피드백 반영

이 절에서는 본 논문에서 제안하는 기법의 사용자 피드백 반영 알고리즘에 대해 상세히 기술한다. 사용자 피드백의 모델에 대한 반영은 모델을 재구축하지 않더라도 사용자의 변화를 추천 결과에 반영할 수 있도록 하며, 정확도를 보장할 수 없는 사용자와 아이템의 특성벡터를 학습을 통해 보완해갈 수 있도록 한다. 이는 사용자의 피드백을 전처리하여 선호도 정보를 얻는 과정, 사용자 선호도 정보를 선호도 행렬에 입력하는 과정, 사용자 선호도 정보를 이용해 사용자 특성 벡터와 아이템 특성 벡터를 학습시키는 과정, 변경된 특성 벡터를 이용해 아이템과 사용자의 군집을 재배치 하는 과정으로 이뤄진다. 이러한 방법을 사용하기 위해서는 사용자 특성 벡터와 아

〈Table 4〉 Example of Feature Vector of Item and User

Example of Feature Vector of Item	
Physical	Float
Emotion	Float
Society	Float
Intelligence	Float
Mental	Float

Example of Feature Vector of User	
Prefer Physical	Float
Prefer Emotion	Float
Prefer Society	Float
Prefer Intelligence	Float
Prefer Mental	Float

이템 특성 벡터의 각 항목이 1대 1로 매칭될 필요가 있다. <Table 4>는 이 연구에서 검증을 위해 구현한 추천 시스템에서 사용한 사용자와 아이템의 특성 벡터의 예시이다.

사용자의 특성 벡터는 아이템의 각 요소에 대한 사용자의 선호도를 나타내고, 아이템의 특성 벡터는 아이템이 가진 각 요소를 지수의 형태로 수치화한 것이다. 본 연구에서는 실험을 위해 문화재, 관광지, 레저 아이템을 이용하고 있으며, 아이템의 특성을 신체적, 정서적, 사회적, 지적, 정신적 다섯 가지 요소로 정의하고 사용자의 특성 벡터를 아이템 특성 벡터의 각 요소에 대한 선호도로 정의하고 있다. 사용자 피드백 반영 방법의 자세한 알고리즘은 <Table 5>와 같다.

〈Table 5〉 Algorithm of Applying User Feedback

Algorithm of Applying User Feedback

$F_{a,m}$:Feedback of user α about item m
$\sigma(F_a)$:Average of Feedback of user a
$\overrightarrow{UV_a}$:Feature vector of user a
$\overrightarrow{CV_m}$:Feature vector of item m

$$\mathrm{preR_{a,m}} = \frac{F_{a,m} - \overrightarrow{F_a}}{\sigma(F_a)}$$

<Formula 6>

If $\mathrm{preR_{a,m}} = 10$,

$$\mathrm{preR_{a,m}} = 10$$

<Formula 7>

End If

If $\mathrm{preR_{a,m}} < -10$,

$$\mathrm{preR_{a,m}} < -10$$

<Formula 8>

End If

$$\mathrm{R_{a,m}} = \mathrm{pre\,R_{a,m}}/20 + 1$$

<Formula 9>

$$\overrightarrow{\mathrm{UV_a}} = \mathrm{R_{a,m}} \times \mathrm{CV_m} + \left(1 - \mathrm{R_{a,m}}\right) \times \mathrm{pre}\overrightarrow{\mathrm{UV_a}}$$

$$\overrightarrow{\mathrm{CV_m}} = \mathrm{R_{a,m}} \times \mathrm{UV_a} + \left(1 - \mathrm{R_{a,m}}\right) \times \mathrm{pre}\overrightarrow{\mathrm{CV_m}}$$

<Formula 10>

이때, <Formula 6-8>은 사용자의 피드백을 전처리하기 위한 과정이다. 사용자 마다 점수를 부여하는 기준점과 선호도의 정도에 따라 부가하는 점수의 크기가 다르기 때문에 이를 정규화 하기 위해, 과거 피드백의 평균과 표준편차를 이용한다. 또한 $\mathrm{preR_{a,m}}$의 절댓값이 10보다 클 경우 확률적으로 의미가 없으므로 이 값들에는 -10또는 10의 값을 부여한다. 마지막으로 <Formula 9>의 과정을 거쳐 이 값의 표현 범위를 0부터 1로 변환하여 <Formula 9>에 용이하게 적용할 수 있도록 한다. <Formula 10>은 $\mathrm{R_{a,m}}$의 값에 따라, 상대 원소의 특성 벡터를 자신의 특성 벡터에 적용하여 특성벡터를 학습시키는 과정이다. 이러한 과정을 거쳐 내용 기반 필터링의 개념을 사용한 추천 시스템에서 문제가 되는 정성적 특성 정량화의 어려움을 해소한다.

또한, 사용자의 피드백이 입력된 경우, <Formula 5>를 이용하여 사용자와 사용자가 평가한 아이템의 각 사용자, 아이템 군집에 대한 최대 우도추정치를 구하고, 이 값이 최대가 되는 군집으로 사용자와 해당 아이템을 재배치한다. 이는 시간 복잡도가 매우 높은 군집화 과정을 다시 거치지 않고도 사용자의 피드백을 추천 모델에 반영하기 위함이다. 또한, 이를 통해 군집화 기법을 이용함으로써 발생하는 모델 유연성 저하를 해소하고 추천 모델을 변화에 능동적으로 적응할 수 있도록 한다.

3.1.4 군집 기반 사용자 선호도 예측

군집 기반 선호도 예측 알고리즘은 기존 협업 필터링 기법의 아이템 기반 선호도 예측 알고리즘과 사용자 기반 선호도 예측 알고리즘을 모두 이용하면서, 본 논문에서 제안하는 기법에 적합하도록 변형시킨 알고리즘이다. 이 알고리즘에

서는 사용자와 사용자간의 관계와 아이템과 아이템간의 관계 모두를 동시에 사용하여, 예측의 정확도를 개선한다. 또한 사용자와 아이템의 유사 집합을 구성하는 대신 각 사용자와 아이템이 속한 군집을 이웃 집단으로 이용하여 기존 예측 알고리즘의 시간 복잡도를 개선한다. 이 방법은 예측 정확도를 향상시킬 뿐만 아니라, 선호도 행렬 희박성 문제에 대한 강건성 또한 높일 수 있으며, 대규모 서비스 환경에서도 실시간으로 누락된 사용자 선호도를 예측할 수 있다. <Table 6>은 군집 기반 사용자 선호도 예측 알고리즘의 요약이다.

〈Table 6〉 Algorithm of Preference Prediction in Adaptive Collaborative Filtering Model

Algorithm of Preference Prediction in Adaptive Collaborative Filtering Model
1) Calculate similarity between each items and make item-similarity matrix, and Calculate similarity between each users and make user-similarity matrix. 2) Missing rating points are predicted by combination predicted preference values each based on item-cluster and user-cluster.

첫 번째 단계에서 아이템들 간의 아이템 유사도 가중치는 <Formula 11>을 통해서, 사용자들 간의 사용자 유사도 가중치는 <Formula 12>을 통해서 계산된다. 아이템 유사도 표와 사용자 유사도 표는 각 두 아이템의 유사도 가중치와 각 두 사용자의 유사도 가중치를 바탕으로 각각 구성된다.

$$w_{i,j} = \frac{\sum_{u \epsilon U} (R_{u,i} - \overline{R_i})(R_{u,j} - \overline{R_j})}{\sqrt{\sum_{u \epsilon U} (R_{u,i} - \overline{R_i})^2} \sqrt{\sum_{u \epsilon U} (R_{u,j} - \overline{R_j})^2}}$$

〈Formula 11〉

U가 아이템 i와 j를 모두를 평가한 적이 있는 사용자들의 집합일 때, $R_{u,i}$는 사용자 u의 아이템 i에 대한 선호도이고, $R_{u,j}$는 사용자 u의 아이템 j에 대한 선호도이다. 또한, $\overline{R_i}$는 사용자 집합 U의 아이템 i에 대한 선호도의 평균이며 $\overline{R_j}$는 사용자 집합 U의 아이템 j에 대한 선호도의 평균이다.

$$w_{a,u} = \frac{\sum_{i \epsilon I} (R_{a,i} - \overline{R_a})(R_{u,i} - \overline{R_u})}{\sqrt{\sum_{i \epsilon I} (R_{a,i} - \overline{R_a})^2} \sqrt{\sum_{i \epsilon I} (R_{u,i} - \overline{R_u})^2}}$$

〈Formula 12〉

I가 사용자 a와 u 모두에 의해 평가된 적이 있는 아이템들의 집합일 때, $R_{a,i}$는 사용자 a의 아이템 i에 대한 선호도이고, $R_{u,i}$는 사용자 u의 아이템 i에 대한 선호도이다. 또한, $\overline{R_a}$는 아이템 집합 I에 대한 사용자 a의 선호도의 평균이며 $\overline{R_j}$는 아이템 집합 I에 대한 아이템 u의 선호도의 평균이다.

두 번째 단계에서는 사용자 α의 아이템 m에 대한 평가를 사용자 α가 속한 사용자 군집의 사용자 가중치의 평균과 아이템 m이 속한 아이템 군집의 가중치의 평균의 합으로 예측한다. 그 과정은 <Formula 13>과 같다.

$$p_{a,m}(hybrid) = \alpha \times p_{a,m}(u) + (1-\alpha) \times p_{a,m}(i)$$

$$= \frac{\sigma(CC_m)}{\sigma(UC_a) + \sigma(CC_m)} \times \left[\overline{R_a} - \frac{\sum_{n \in UC_a} (R_{n,m} - \overline{R_a}) \times uw_{a,n}}{\sum_{n \in UC_a} uw_{a,n}} \right]$$

$$+ \frac{\sigma(UC_a)}{\sigma(UC_a) + \sigma(CC_m)} \times \frac{\sum_{l \in CC_m} R_{a,l}\, cw_{m,l}}{\sum_{l \in CC_m} cw_{m,l}}$$

〈Formula 13〉

사용자 α의 아이템 m에 대한 선호도를 예측할 때, 사용자 α가 속한 군집을 UC_a, 아이템 m이 속한 군집을 CC_m이라 하면, $p_{a,m}(hybrid)$는 사용자 α의 아이템 m에 대한 선호도가 된다. $p_{a,m}(u)$는 사용자 가중치의 평균을 의미하며, $p_{a,m}(i)$은 아이템 가중치의 평균을 의미하고, $\sigma(UC_a)$는 사용자 a가 속한 군집의 표준 편차를, $\sigma(CC_m)$는 아이템 m이 속한 군집의 표준편차를, $uw_{a,n}$과 $cw_{m,l}$은 각 사용자 a와 n의, 아이템 m과 l의 유사도 가중치를 말한다. a는 모델의 상황에 따라 능동적으로 변화하는 값으로 각 사용자 군집과 유사도 군집의 데이터 신뢰성에 따라, 1부터 0의 값을 갖는다. 이 방법은 각 유사 사용자 집합과 유사 아이템 집합의 선호도 예측의 근거로써의 신뢰도를 표준 편차를 이용해 추정하여 두 예측 결과값을 조합하는데 반영함으로써 예측의 정확도를 높일 수 있게 한다.

3.3 적응형 군집화 기반 협업 필터링 추천

이 절에서는 적응형 군집화 기반 협업 필터링 추천 모델을 이용해 추천 서비스를 수행하기 위한 알고리즘을 상세히 기술한다. 이 알고리즘은 사용자와 아이템의 군집 그리고 사용자 군집의 아이템 군집에 대한 선호도, 사용자의 아이템에 대한 선호도를 이용해 이뤄진다. 사용자가 추천 서비스를 요청할 경우, 선호도가 높은 아이템 군집에서부터 사용자의 선호도가 사용자의 선호도 평균 이상인 아이템을 추천 리스트에 추가하고 추천 리스트를 선호도 순으로 정렬하여 사용자에게 제공하게 된다. 이러한 방법은 일반적인 협업 필터링 추천 기법보다 수행시간의 연산량을 줄여, 대규모 서비스에서 서비스 품질을 유지하면서 실시간 서비스를 가능하게 한다. 이 알고리즘의 상세한 수행 절차는 <Table 7>과 같다.

〈Table 7〉 Algorithm of Executing Recommendation

Algorithm of Executing Recommendation

UC_a : User-cluster including user a
CC_{arr} : Array of item-clusters
CP : Inter-cluster Preference
c_j : jth element of $CC_{arr}[i]$
R_{a,c_j} :Preference of user α about item c_j
p_{a,c_j} : Predicted value of preference of user α about item c_j
RC_{arr} : Array of items to recommend

Sort $CC_{arr}[i]$ in descending order according to CP.
For $i = 1 \rightarrow i = (Number\ of\ Item\ Cluster)$
For $j = 1 \rightarrow j = \mathrm{N}_{\mathrm{cc}_{\mathrm{arr[i]}}}(Number\ of\ Element\ in\ Item - Cluster, CC_{arr}[i])$
If $R_{a,c_j} = \emptyset$,

$$R_{a,c_j} = p_{a,c_j}(hybrid)$$ <Formula 14>

End If
If $\overline{R_a} < R_{a,c_j}$,
Add item, c_j to RC_{arr}.
End If
If RC_{arr} is including more than k items,
End For
End If
End For
End For
Sort RC_{arr} in descending order according to preference of user a about each item.

4. 실험 및 검증

본 연구에 대한 검증은 다음 두 가지 기준을 통해 이루어진다. 첫째는 추천 시스템의 성능 평가를 위해 가장 많이 사용되는 평균 절대 오차(*MAE*)(Herlocker et. al., 2004; Goldberg et. al., 2001)방법이다. 이 측정법은 예측한 순위와 실제 순위의 차이를 통해 추천 시스템의 신뢰도를 측

정한다. 둘째는 응답시간이다. 이는 추천 시스템의 시스템 부하를 보여준다. 평균 절대 오차(*MAE*, Mean Absolute Error)를 구하는 방법은 <Formula 15>와 같다.

$$MAE = \frac{\sum_{i=1}^{N} |p_i - r_i|}{N}$$

〈Formula 15〉

여기서 N은 예측한 아이템의 개수, p_i는 i번째 아이템의 예측한 순위, r_i는 i번째 아이템의 실제 순위이다.

실험은 다음과 같은 환경에서 진행되었다. 서버는 Windows 7기반의 Apache Tomcat 7.0을 사용하였으며, 데이터베이스는 MySQL5.5를 사용하였다. 서버 측 통합개발환경은 VisualStudio 2010를 사용하였으며, 사용된 서버 측 언어는 VC++이다. 클라이언트는 안드로이드 응용프로그램으로 구현되었고, 클라이언트 측 통합개발환경은 Eclipse Indigo, 개발언어는 Android SDK를 이용한 JAVA이다.

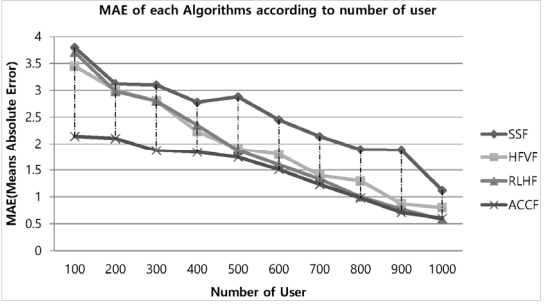
실험 데이터는 다음과 같다. 아이템의 경우, 종로구 인근의 300여 개의 문화재, 관광지, 레저 아이템을 이용하였고, 사용자 프로파일의 경우, 실험에 참여한 사용자들이 직접 입력하였으며, 아이템의 특성은 아이템에 관한 공개된 정보를 이용하였다. 또한 실험을 위해 구현한 시스템은 실시간 비교사적 학습을 바탕으로 하여 별도의 훈련 집합을 필요로 하지 않으나, 최초의 모델 생성을 위하여 단국대학교 학생 및 교직원으로 구성된 50명의 사용자 집단으로 훈련 집합을 구성하였다. 훈련 집단은 20대부터 60대까지 각 연령대에 균등한 연령 분포를 갖도록 구성하였다.

아이템 제공 및 추천 방법은 다음과 같다. 아이템은 사용자가 지정한 거리 내의 인근 아이템을 사용자의 프로파일과 아이템의 특성을 바탕으로 최대 30개까지 추천한다. 이 과정은 예측된 사용자의 선호도를 바탕으로 사용자가 속한 군집의 군집간 선호도가 높은 아이템 군집의 아이템을 우선적으로 추천하는 형태로 이뤄진다. 사용자는 아이템의 이용이 끝나면 0점에서 10점까지의 11단계의 점수를 입력하여, 이용한 아이템에 대해 평가한다.

실험은 Park, Kyusik(Park et. al., 2010)이 제안한 SSF(Single-Scaled Hybrid Filtering)기법과, Kim, Yong(Kim and Moon, 2006)이 제안한 HFUF(Hybrid Recommendation System Based on Usage frequency), Shen, Yan(Shen et. al., 2012)이 제안한 RLHF(Reinforcement Learning Algorithm Based Hybrid Filtering)기법을 본 연구에서 제안한 추천 기법과 같은 환경에서 구현하여 평균 절대 오차(*MAE*)와 응답 시간을 비교하는 방식으로 진행되었다.

각 실험의 상세한 수행 과정은 다음과 같다. 첫 번째 실험의 경우, 단국대학교 학생과 교직원 및 인근 주민을 대상으로 150명의 20대부터 60대까지 균등한 연령 분포의 표본 사용자 집단으로 실험 집합을 구성하였다. 사용자는 사용자 자신에게 추천된 모든 아이템에 대한 선호도 정보를 입력하고, 사용자가 평가를 바탕으로 추천된 아이템들의 실제 순위를 구하고 이를 예측된 아이템의 순위와 비교하여 평균 절대 오차(*MAE*)를 구하는 방식으로 진행되었다. 두 번째 실험의 경우, 첫 번째 실험에 참여한 모든 사용자를 대상으로 서비스 응답 시간을 서버의 로그를 바탕으로 측정하고 그 평균을 구하였다. 모든 실험은 동일한 서버와 클라이언트 기기, 네트워크 환경에서 진행되었다. <Figure 1>과 <Figure 2>는 각

기법을 적용한 추천 시스템에서 사용자 수의 변화에 따른 평균 절대 오차(*MAE*)와 평균 응답 시간의 변화를 나타낸 그래프이며 <Figure 2>의 측정단위는 1/1000초(*ms*)이다. 또한 <Table 8>과 <Table 9>는 각각 평균 절대 오차(*MAE*)와 응답 시간의 평균과 표준편차, 범위를 나타낸 표이다. 본 연구에서 제안하는 적응형 군집화 기반 협업 필터링 기법은 ACCF(Adaptive Clustering based Collaborative Filtering)라 표기한다.

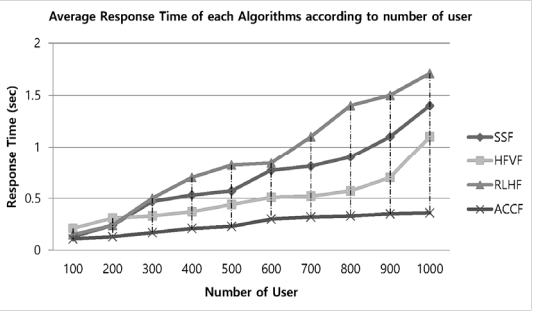


〈Figure 1〉 MAE of each Algorithms according to number of user

〈Table 8〉 Average, Standard Deviation and Range of MAE for each Algorithms

	SSF	HFVF	RLHF	ACCF
Average	2.516	1.956	1.899	1.472
Standard Deviation	0.778	0.907	1.038	0.562
Range	2.678	2.650	3.132	1.540

<Figure 1>과 <Table 8>은 적응형 군집화 기반 협업 필터링 기법이 서비스 초기에 다른 추천 기법들에 비해 우수한 성능을 보이며 사용자 수의 변화에 대해 비교적 일정한 추천 신뢰도를 보이는 것을 보여준다. RLHF기법과의 비교에서 이용자의 수가 500명 이상인 경우, 거의 비슷한 성능을 보이고 있지만 사용자가 1000명 이상인 경우에는 다른 추천 기법들에 비해 약간의 신뢰도 저하를 보일 것으로 추정된다. 또한, ACCF의 경우 평균 절대 오차(*MAE*)의 범위가 1.54로 RLHF의 3.13에 비해 50.79% 개선되었으며, 평균 절대 오차(*MAE*)의 평균 또한 ACF는 1.472, RLHF는 1.899로 22.48% 개선되었음을 볼 수 있다. 또한 표준 편차를 통한 비교에서도 ACF가 다른 기법들에 비해 안정적인 추천 신뢰도를 보이고 있음을 알 수 있다.



〈Figure 2〉 Response Time of each Algorithms according to number of user

〈Table 9〉 Average, Standard Deviation and Range of Response Time for each Algorithms

	SSF	HFVF	RLHF	ACCF
Average	0.672	0.485	0.894	0.251
Standard Deviation	0.359	0.267	0.531	0.093
Range	1.131	0.980	1.587	0.253

<Figure 2>와 <Table 9>에서와 같이 적응형 군집화 기반 협업 필터링 기법이 다른 추천 기법들에 비해 짧은 응답시간을 보인다. 비교 대상들 중 가장 짧은 응답시간을 보이는 HFVF와 비교할 경우, 사용자가 700명 미만인 상황에서 0.4초 미만의 차이를 보인다. 하지만, 사용자가 700명 이상의 경우에서는 HFVF의 경우 기하급수적 증

가를 보이지만 ACCF는 선형적 증가를 보이며 오히려 증가율의 감소를 보인다. 이는 본 논문에서 제안하는 기법이 기존 협업 필터링 기법의 시간 복잡도를 개선하였음을 보인다. 또한 <Table 9>에서 확인할 수 있는 바와 같이, 응답시간의 평균은 HFVF에 비해 48.25% 개선되었고 범위 또한 74.18% 개선되었고, 표준편차의 경우에도 확연히 개선되었음을 볼 수 있다.

위의 두 실험은 본 논문에서 제안하고 있는 적응형 군집화 기반 협업 필터링 기법이 여타 추천 기법들에 비해 응답 시간 면에서 확연히 개선되었음을 보이며, 응답시간의 개선을 위해 군집화 방법을 이용하였음에도 기존 협업 필터링 기법들과 비교해 추천 신뢰도가 거의 저하되지 않고, 오히려 서비스 초기에는 월등한 추천 신뢰도를 보임을 알 수 있다. 이는 적응형 군집화 기반 협업 필터링 기법이 기존 협업 필터링 기법들에 비해 대규모 서비스에 적합하다는 것을 뜻한다 할 수 있으며, 서비스 초기 상황에서도 다른 추천 기법들보다 우수한 추천 신뢰도를 보여주고 있어 확장성 있는 추천 시스템 구현에 적합하다는 것을 알 수 있다. 이는 본 기법이 기존 협업 필터링의 고질적인 문제인 초기 사용자, 초기 평가 문제, 확장성 문제를 해소했다 볼 수 있으며, 협업 필터링 기법의 상용화 시의 문제점을 대부분 해소했다 할 수 있다.

본 논문에서 제안하는 기법이 기존 협업 필터링 기법의 초기 사용자, 초기 평가 문제를 개선하고 있지만, 다른 기법과의 상대적인 비교가 아닌 절대적인 수치로 평가할 경우 개선의 여지가 남아있다. 사용자가 100명부터 500명인 상황에서 평균 절대 오차(*MAE*)의 평균은 1.938로 서비스 초기 상황이라는 점을 감안하지 않는다면 추천의 신뢰도가 높다고 볼 수는 없는 수치이다. 또한 사용자가 1000명인 시점에서는 RLHF의 평균 절대 오차(*MAE*) 0.02 더 낮으며, 평균 절대 오차(*MAE*)의 감소율을 볼 때, 사용자가 1000명 이상인 상황에서는 추천 시스템의 신뢰도가 다른 추천 기법들과 비슷할 것이라 볼 수 있다.

5. 결론

본 연구에서는 서비스의 규모 변화에도 안정적인 성능을 보이는 상용화에 적합한 추천 시스템을 구현하기 위하여, 적응형 군집화 기반 협업 필터링 추천 기법을 제안하고 구현하였으며, 이를 실제 문화재, 관광지, 레저 추천 시스템에 적용하여 그 성능을 평가함으로서 제안하는 기법의 유효성을 검증하였다.

제안하는 기법은 협업 필터링 추천 기법에서 나타나는 초기 사용자, 초기 평가 문제, 확장성 문제, 선호도 행렬 희박성 문제를 해결하기 위하여, 적응형 군집화 기법을 도입하였다. 그 방법은 다음과 같이 크게 네 가지로 나뉘어진다. 첫째, 사용자와 아이템의 특성 벡터를 기반으로 사용자와 아이템 각각을 군집화 하여 협업 필터링 추천 기법에서 추천 시 사용자나 아이템의 부분 집합을 생성하는데 소요되는 시간을 절약하며 사용자 선호도 정보만을 이용한 부분 집합 생성보다 추천의 신뢰도를 높이고 초기 평가 문제와 초기 이용자 문제를 일부 해소한다. 둘째, 미리 구성된 사용자와 아이템의 군집을 기반으로 군집간의 선호도를 통해 추천을 수행한다. 사용자가 속한 군집의 선호도가 높은 순서대로 아이템 군집을 조회하여 사용자에게 제공할 아이템 목록을 구성하고, 순위를 매긴다. 이를 통해, 추천 시스템의 부하 대부분을 모델 생성 단계에서 부

담하고 실제 수행 시 부하를 최소화한다. 셋째, 누락된 사용자 선호도 정보를 사용자와 아이템 군집을 이용하여 예측함으로써 협업 필터링 추천 기법의 사용자 선호도 정보 희소성으로 인한 문제를 해소한다. 또한 이는 선호도 예측 과정의 시간 복잡도를 낮춰 수행 시간 중에 누락된 선호도를 예측하는 것을 가능하게 한다. 넷째, 사용자와 아이템의 특성 벡터를 사용자의 피드백에 따라 학습시켜 나감으로써 아이템과 사용자의 정성적 특성의 정량화의 어려움을 해결한다.

본 연구의 검증은 Park, Kyusik이 제안한 SSF 기법과, Kim, Yong이 제안한 HFUF기법, Shen, Yan이 제안한 RLHF기법을 본 연구에서 제안한 추천 기법과 같은 환경에서 구현하여 평균 절대 오차(*MAE*)와 응답 시간을 비교하는 방식으로 진행되었다. 두 실험은 본 논문에서 제안하고 있는 적응형 군집화 기반 협업 필터링 기법이 여타 추천 기법들에 비해 응답 시간 면에서 평균 48.25% 개선되었음을 보이며, 응답시간의 개선을 위해 군집화 방법을 이용하였음에도 기존 협업 필터링 기법들과 비교해 추천 신뢰도가 평균 50.79% 개선되었음을 보인다. 이는 본 기법이 기존 협업 필터링 기법들에 비해 확장성 있는 추천 시스템의 구현에 적합함을 증명한다.

본 논문의 개선의 여지 또한 존재한다. 서비스 초기의 추천 신뢰도는 개선되었지만, 절대적인 수치로 보면 신뢰도가 크게 높다고 보기 어려우며, 서비스의 규모가 증가할수록 추천의 신뢰도가 올라가는 여타 협업 필터링 기반 추천 시스템들에 비해, 신뢰도의 증가율이 낮다는 것이다. 이는 군집화 방법을 사용하여 추천의 대상이 되는 아이템의 범위를 한정하기 때문으로 보이며, 이를 해결하기 위해 사용자의 성향을 더욱 효과적으로 반영할 수 있는 군집화 방법이 연구되어야 할 것이다. 또한 협업 필터링의 고질적인 문제인 초기 평가, 초기 이용자 문제에 대한 완전한 해결책을 제시하지는 못하고 있다. 향후 연구는 규칙 기반 전문가 시스템을 활용하여 초기 평가, 초기 이용자 문제에 대한 더욱 확실한 해결책을 제시하는 방향으로 나아갈 것이다.

참고문헌 (References)

Adomavicius, G., and A. Tuzhilin, "Towards the next generation of recommender system : A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, (2005), 634~749.

Ali, K., and W. V. Stam, "Tivo: Making show recommendations using a distributed collaborative filtering architecture," *Conference on Knowledge Discovery and Data Mining,* (2004), 394~401.

Baeza-Yates, R., and B. Ribeiro-Neto, *Modern information retrieval,* ACM press, New York, 1999.

Bennet, J., and S. Lanning, "The netflix prize," *KDD Cup and Workshop*, www.netflixprize.com, (2007).

Braak, P. T., N. Abdullah, and Y. Xu, "Improving the Performance of Collaborative Filtering Recommender Systems through User Profile Clustering," *Web Intelligence and Intelligent Agent Technologies*, 2009. WI-IAT '09. IEEE/ WIC/ACM International Joint Conferences on Vol.3, (2009), 147~150.

Connor, M. O., and J. Herlocker, "Clustering Items for Collaborative Filtering," *Proceedings of the ACM SIGIR Workshop on Recommender Systems*, Berkeley, CA, (1999).

Das, A., M. Datar, A. Garg, and S. Rajaram, "Google news personalization: Scalable online collaborative filtering," *World Wide Web Conference*, (2003), 271~280.

George, T., and S. Merugu, "A scalable collaborative filtering framework based on co-clustering," In: *Data Mining, Fifth IEEE International Conference on*. IEEE, (2005).

Goldberg, K., T. Roeder, D. Gupta, and C. Perkins, "Eigentaste: a constant time collaborative filtering algorithm," *Information Retrieval*, Vol.4, No.2(2001), 133~151.

Groh, G., and C. Ehming, "Recommendations in Taste Related Domains: Collaborative filtering vs. Social filtering," *Proceedings of GROUP'07*, (2007), 127~136.

Heo, G. Y., and W. W. Woo, "Extensions of X-means with Efficient Learning the Number of Clusters," *The journal of the Korea Institute of Maritime Information & Communication Sciences*, Vol.12, No.4(2008), 772~780.

Herlocker, J. L., J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, Vol.22, No.1(2004), 5~53.

Herlocker, J. L., J. A. Konstan, and J. T. Riedl, "An Empirical Analysis of Design Choices in Neighborhood-based Collaborative Filtering Systems," *Information Retrieval*, Vol.5, (2002), 287~310.

Huang, Z., H. Chen, and D. Zeng, "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering," *ACM Trans. Information Systems*, Vol.22, No.1(2004), 116~142.

Ji, H., J. Li, C. Ren, and M. He, "Hybrid collaborative filtering model for improved recommendation," *Service Operations and Logistics, and Informatics (SOLI)*, 2013 IEEE International Conference, (2013), 142~145.

Kass, R. E., and L. Wasserman, "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, Vol.90, No.431(1995), 928~934.

Kim, S. H., B. H. Oh, M. J. Kim, and J. H. Yang, "A Movie Recommendation Algorithm Combining Collaborative Filtering and Content Information," *Journal of KIISE*. Vol.39, No.4(2012), 261~268.

Kim, S. Y., S. H. Lee, and H. S. Hwang, "Design and Implementation of the recommendation system that is personalized using the hybrid filter and AHP," *Journal of Korean Industrial Information Systems Society*, Vol.17, No.7 (2012), 111~118.

Kim, Y., and S. B. Moon, "A Study on Hybrid Recommendation System Based on Usage frequency for Multimedia Contents," *Journal of the Korean society for information management*, Vol.23 No.3(2006), 91~125.

Konstan, J., D. B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Ridl, "GroupLens: Applying collaborative filtering to Usenet news," *Communication of ACM*, Vol.40, No.3(1997), 77~87.

Linden, G., B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing*, (2003), 76 ~80.

Melville, P., R. J. Mooney, and R. Nagarajan, "Content-Boosted Collaborative Filtering for

Improved Recommendations," *Proceedings of the Eighteenth National Conference on Artificial Intelligence, Edmonton, Canada*, (2002), 187~192.

Park, K. S., J. M. Choi, and D. H. Lee, "A Single-Scaled Hybrid Filtering Method for IPTV Program Recommendation," *INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING*, Vol.4, No.4(2010), 161~168.

Park, S., and D. Pennock, "Applying collaborative filtering techniques to movie search for better ranking and browsing," *Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2007), 550~559.

Pazzani, M., "A Framework for Collaborative, Content-Based and Demographic Filtering," *Artificial Intelligence Review*, Vol.13(1999), 398~408.

Renick, P., and H. R. Varian, "Recommender System," Communication of the ACM, Vol.40, No.3(1997), 56~58

Sarwar, B., G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithm," *Proceedings of the 10th international conference on World Wide Web*, (2001), 285~295.

Schwarz, G., "Estimating the Dimension of a Model," *The Annals of Statistics*, Vol.6, No.2(1987), 461~464.

Shen, Y., H. C. Shin, D. G. Kim, Y. H. Hong, and P. K. Rhee, "Reinforcement Learning Algorithm Based Hybrid Filtering Image Recommender System," *The journal of the Institute of Internet Broadcasting and Communication*, Vol.12, No.3(2012), 75~81.

Shepitsen, A., J. Gemmell, B. Mobasher, and R. Burke, "Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering," *Proceeding of the 2008 ACM conference onRecommender systems*, Lausanne, Switzerland, (2008), 259~266.

Xue, G. R., C. Lin, Q. Yang, W. Xi, H. J. Zeng, Y. Yu, and Z. Chen, "Scalable collaborative filtering using cluster-based smoothing," *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval,* (2005), 114-121.

Abstract

Scalable Collaborative Filtering Technique based on Adaptive Clustering

O-Joun Lee* · Min-Sung Hong** · Won-Jin Lee*** · Jae-Dong Lee****

An Adaptive Clustering-based Collaborative Filtering Technique was proposed to solve the fundamental problems of collaborative filtering, such as cold-start problems, scalability problems and data sparsity problems. Previous collaborative filtering techniques were carried out according to the recommendations based on the predicted preference of the user to a particular item using a similar item subset and a similar user subset composed based on the preference of users to items. For this reason, if the density of the user preference matrix is low, the reliability of the recommendation system will decrease rapidly. Therefore, the difficulty of creating a similar item subset and similar user subset will be increased. In addition, as the scale of service increases, the time needed to create a similar item subset and similar user subset increases geometrically, and the response time of the recommendation system is then increased. To solve these problems, this paper suggests a collaborative filtering technique that adapts a condition actively to the model and adopts the concepts of a context-based filtering technique. This technique consists of four major methodologies. First, items are made, the users are clustered according their feature vectors, and an inter-cluster preference between each item cluster and user cluster is then assumed. According to this method, the run-time for creating a similar item subset or user subset can be economized, the reliability of a recommendation system can be made higher than that using only the user preference information for creating a similar item subset or similar user subset, and the cold start problem can be partially solved. Second, recommendations are made using the prior composed item and user clusters and inter-cluster preference between each item cluster and user cluster. In this phase, a list of items is made for users by examining the item clusters in the order of the size of the inter-cluster preference of the user cluster, in which the user belongs, and selecting and ranking the items according to the predicted or recorded user

* Dept. of Software Science, Dankook University
** Dept. of Computer, Dankook University
*** Institute of Media Contents, Dankook University
**** Corresponding author: Jae-Dong Lee
Dept. of Software Science, Dankook University
152, Jukjeon-ro, Giheung-gu, Yongin-si, Gyeonggi-do, Korea
Tel: +82-031-8005-3254, Fax: +82-031-8021-7180, E-mail : letsdoit@dankook.ac.kr

preference information. Using this method, the creation of a recommendation model phase bears the highest load of the recommendation system, and it minimizes the load of the recommendation system in run-time. Therefore, the scalability problem and large scale recommendation system can be performed with collaborative filtering, which is highly reliable. Third, the missing user preference information is predicted using the item and user clusters. Using this method, the problem caused by the low density of the user preference matrix can be mitigated. Existing studies on this used an item-based prediction or user-based prediction. In this paper, Hao Ji's idea, which uses both an item-based prediction and user-based prediction, was improved. The reliability of the recommendation service can be improved by combining the predictive values of both techniques by applying the condition of the recommendation model. By predicting the user preference based on the item or user clusters, the time required to predict the user preference can be reduced, and missing user preference in run-time can be predicted. Fourth, the item and user feature vector can be made to learn the following input of the user feedback. This phase applied normalized user feedback to the item and user feature vector. This method can mitigate the problems caused by the use of the concepts of context-based filtering, such as the item and user feature vector based on the user profile and item properties. The problems with using the item and user feature vector are due to the limitation of quantifying the qualitative features of the items and users. Therefore, the elements of the user and item feature vectors are made to match one to one, and if user feedback to a particular item is obtained, it will be applied to the feature vector using the opposite one. Verification of this method was accomplished by comparing the performance with existing hybrid filtering techniques. Two methods were used for verification: MAE(Mean Absolute Error) and response time. Using MAE, this technique was confirmed to improve the reliability of the recommendation system. Using the response time, this technique was found to be suitable for a large scaled recommendation system. This paper suggested an Adaptive Clustering-based Collaborative Filtering Technique with high reliability and low time complexity, but it had some limitations. This technique focused on reducing the time complexity. Hence, an improvement in reliability was not expected. The next topic will be to improve this technique by rule-based filtering.

Key Words : Recommendation System, Adaptive System, Collaborative Filtering, Hybrid Filtering, Clustering

Received: April 5, 2014 Revised: May 14, 2014 Accepted: May 16, 2014

저 자 소 개

이 오 준

현재 단국대학교 소프트웨어학과 학사 과정에 재학 중이다. 연구 관심 분야는 적응형 시스템, 개인화 맞춤형 시스템, 추천 시스템, 데이터 마이닝, 비즈니스 인텔리젼스 등이다.

홍 민 성

단국대학교 컴퓨터공학과 학사 과정 후, 단국대학교 컴퓨터학과 석사 과정에 재학 중이다. 연구 관심 분야는 개인화 맞춤형 시스템, 추천 시스템, 모바일 센서 네트워크, 상황 인식, 데이터 마이닝 등이다.

이 원 진

경일대학교 컴퓨터공학부 학사, 경북대학교 컴퓨터공학과 석사, 금오공과대학교 전자통신공학 박사 과정 후, 2007년부터 2009년까지 경일대학교 컴퓨터공학부 전임강사를 역임하고, 현재 단국대학교 미디어콘텐츠연구원 초빙교수로 재직중이며, 연구 관심 분야는 유비쿼터스 컴퓨팅 보안, 홈네트워크 보안, 센서네트워크 보안 등이다.

이 재 동

인하대학교 전산학 학사, Cleveland State University 컴퓨터과학 석사, Kent State University 컴퓨터과학 박사 과정 후, 현재 단국대학교 소프트웨어학과 교수, 국제문화교류처 처장으로 재직 중이며, 연구 관심 분야는 모바일 소프트웨어, 소셜 네트워크, 융합 시스템, 그리고 모바일 서비스 등이다.