군집 분석을 통한 Collaborative Filtering 기반의 추천시스템의 성능개선

우희성*°, 서용무**
*고려대학교 대학원 경영학과, **고려대학교 경영대학 경영학과

Performance Improvement Using Clustering in Collaborative Filtering Recommendation Systems

Woo, Hee Sung°, Suh Yong-Moo
Korea University
E-mail: wooheesung@hotmail.com, ymsuh@korea.ac.kr

요 약

추천시스템을 설계하는 방법에는 크게 Content-Based Filtering 기법과 Collaborative Filtering 기법이 있다. 이 중 Collaborative Filtering 기법은 사용자가 아직 평가하지 못한 상품에 대한 예측값을 계산할 때, 나와 유사한 상품선호를 갖고 있는 사람들이 그 상품에 대해 평가한 점수를 활용하는 방법이다. 하지만 순수한 Collaborative Filtering 방법은 일반적으로 알려진 Data Sparsity의 문제, First Rater의 문제뿐만 아니라 예측값의 부정확성과 기하급수적 계산량의 증가로 실제 구현이 어렵다는 문제점을 가지고 있다. 본 연구에서는 이러한 'Collaborative filtering' 시스템의 문제들 중 예측의 부정확성과 실제 구현의 어려움을 해결할 수 있는 방법으로 군집분석을 적용해 보았다. 특히 본 연구에서는 군집을 나눌 때, 실제 추천이 이루어지는 상품 도메인이 아닌, 그 상품도메인과 비슷한 선호의 기준을 가지고 선택하게 되는 '선택의 상관관계'가 높은 '이웃 상품도메인'에서 사용자들의 군집을 나누고 이를 실제 추천이 이루어지는 상품도메인에 적용하는 방식을 사용하였다.

1. 서론

대형 온라인 상점은 수백만 개의 상품을 판매한다. 이렇게 많은 상품 중에서 자신이 원하는 상품을 찾는 것은 소비자에게 쉽지 않은 일이다. 이러한 도전에 대응해서 개발된 추천시스템은 사용자의 선호를 파악해 사용자가 좋아할 것이라 예상되는 상품을 추천해 주는 시스템이다 [1]. 오늘날

추천시스템은 다양한 형태로 존재하며 실제로 많은 전자상거래 사이트에서 활용되고 있다. 이의 대표적인 예로는 Amazon.com, CDNow.com, eBay.com, Levis.com, Moviefinder.com, Reel.com 등이 있다.

추천시스템을 설계하는 방법은 크게 Information Filtering과 Collaborative Filtering의 두 가지가 있다[2]. Information Filtering(IR)은 Content-based filtering(CBF) 이라고도 하며 사용자의 선호 프로파일을 생성, 활용함으로써 평가되지 않은 상품에 대해서도 그 상품이 가지고 있는 속성이 사용자의 프로파일과 일치하는 정도가 높다면 그 상품을 추천해 주는 방식을 말한다[2]. 하지만 이 방법은 프로파일을 작성할 속성을 추출하기 힘든 성격의 상품들 예컨데, 영화나 유머 등과 같은 상품에 대해서는 적용하기가 어렵다는 단점이 있다. 바로 이러한 단점을 보완하기 위해 등장한 것이 Collaborative Filtering(CF)이다. 이는 사용자의 선호 프로파일을 생성하는 것이 아니라, 사용자와 유사한 선호를 갖는 사람을 찾는 방식이다[2]. 곧, 소비자와 선호가 유사한 다른 소비자를 찾아 내어 그가 좋아한 상품을 소비자도 좋아할 것이라는 가정하에 상품을 추천하게 된다.

하지만 순수한 Collaborative Filtering 방식은 몇 가지 점에서 문제점을 갖고 있다. 먼저, Data Sparsity, First-Rater의 문제점은 널리 알려진 문제점이다[3]. Data Sparsity의 문제점은 사용자들간의 선호의 유사도를 계산할 때 반드시 존재해야 하는 '공통으로 평가한 상품의 수'가 많지 않다는 것이다. A와 B라는 사람의 선호의 유사도를 구하기 위해서는 A와 B가 과거에 공통으로 평가한 상품이 존재해야 하는데 현실의 세계에서는 무수히 많은 상품들 중에서 공통으로 평가한 상품의 수가 많지 않기 때문에 유사도를 계산하는 것이 쉽지 않다는 것이다. First-rater의 문제는 아무도 평가하지 않은 상품에 대해서는 예측 자체가 불가능 하기 때문에 처음으로 아이템을 평가하는 사람 곧, First-Rater가 항상 요구된다는 것이다. 또한 Collaborative Filtering 역시 예측 값의 정확도가 떨어지는 문제점을 갖고 있으며 대용량의 데이터베이스가 형성되는 오늘날의 전자상거래 환경에서 사용자와 아이템의 수가 증가할수록 계산의 양이 무한히 증가하므로 현실적으로 구현하기 힘들다는 한계점도 있다[4].

이러한 Data Sparsity, First-Rater의 문제점을 해결하기 위해 평가되지 않은 결측값을 예측값으로 채우는 연구[2], 예측의 정확도를 높이기 위해서 사용자의 인구통계학적 정보들을 활용하거나[6, 7], 아이템의 숨겨진 속성을 찾아 활용하거나[8], 에이전트를 사용하는 연구[9] 등이 있었다. 그리고 현실적으로 구현 가능한 시스템의 설계를 위해 데이터 축소의 다양한 기법들도 적용되었다[4, 5].

본 연구는 군집분석(Clustering)의 방법을 Collaborative Filtering 기법에 적용해 봄으로써, 군집을 나누지 않고 전체를 대상으로 유사한 선호 집단을 파악하는 Collaborative Filtering 시스템에 비해 과연 예측의 정확도가 높아지는 지를 살펴보았다.

2. 연구내용 및 방법

본 연구에서는 기본적으로 Collaborative filtering 방법을 사용하되, 앞에서 지적한 collaborative filtering의 여러 가지 문제점들의 해결 방안으로 군집분석을 사용하였다 [4, 5]. 본 연구에서의 중요한 점은 군집을 나눌 때 실제적으로 추천이 이루어지는 영역에서의 속성을 기준으로 한 것이 아니라 '타 상품 도메인에서의 선호'라는 새로운 속성을 기반으로 군집을 나누었다는 것이다. 구체적으로 말하자면, 본 연구에서 추천의 도메인으로 선택한 것은 '영화'이며, 군집을 나눌 때 사용한 도메인은 '음악' 장르이다. 영화에 대한 추천시스템을 구현할 때 음악이라는 장르를

활용한 이유는 두 가지 도메인 모두 사람들의 문화적 기호를 반영한다는 점에서 유사성이 있기 때문이다. '영화'에 대해 선호가 비슷한 사람끼리 몇 개의 군집을 나눌 수 있어서 이 군집 별로 Collaborative Filtering의 알고리즘을 사용한다면 나와 선호가 유사한 보다 양질의 Neighbor를 파악할 수 있기 때문에 추천시스템의 정확도를 높일 수 있다[7].

하지만 '영화'라는 도메인에서 군집을 나눌 수 있는 속성을 파악하기는 쉽지 않다. 사용자에 대한 정보가 없이 그들의 평가점수만 주어져 있다고 가정할 때 군집을 나눌 수 있는 속성을 발견하는 것은 더욱 어려워진다. 그러나, 만약 음악이라는 상품 도메인에서 사용자의 선호를 파악할 수 있는 정보를 가지고 있고 이를 기준으로 군집을 나눌 수 있다면 우리는 이 정보를 활용할 수 있다. 상식적으로 생각해도, 서정적이고 잔잔한 경향의 음악을 좋아하는 사람이 좋아하는 영화와, 헤비메탈과 락을 좋아하는 사람이 좋아하는 영화는 분명한 차이가 있을 것으로 쉽게 예상할 수 있으며 이것이 본 연구에서 영화에 대한 추천시스템을 설계할 때 음악이라는 장르의 선호를 가지고 군집을 나눈 이유이다.

실험을 위하여 구축한 웹 사이트[1]에서 온라인 설문조사를 실시하여 데이터를 수집하였으며, 설문조사는 개별 영화를 보여주고 점수를 받는 문항 28개 항목과 음악 장르 8개에 대한 선호점수를 받는 8개 문항, 영화에 관한 관심정도를 묻는 2문항을 포함해 38개의 문항으로 구성되었다. 군집분석의 방법은 K-Means method를 활용하였고, 사용자들간의 Similarity Function, Prediction Function은 MS-SQL2000 Server Function으로 구현하였다.

3. 추천시스템 설계의 세 가지 접근 방법

[1] http://www.mymovie.pe.kr

3.1. Content-Based Filtering

Content-Based Filtering(CBF) 기법에 의한 추천시스템은 사용자가 좋아한 아이템들에 대한 문서적 정보를 분석하고 이것을 통해 사용자의 선호 프로파일을 생성한다. 그리고 새로운 아이템에 대한 추천은 이렇게 형성된 사용자 선호 프로파일을 기초로 해서 이루어진다[8]. CBF 추천시스템에서 사용자는 아이템의 매력도에 따라 수치적 점수로 아이템을 평가하게 된다. CBF 추천시스템은 이러한 평가된 아이템들의 컨텐츠(키워드, 문장)를 분석해서 사용자의 선호를 표현하는 프로파일을 생성한다. 그런후에 사용자가 아직 경험하지 못한 아이템들에 대해 그 컨텐츠와 사용자의 프로파일을 비교하여 이것이 사용자에게 과연 얼마만큼 흥미로운 아이템인지를 평가하게 된다. CBF 추천시스템은 시스템이 분석하여 '이해할 수 있는' 아이템들에 대해서만 사용자에게 추천할 수 있기 때문에 대개 문서정보로 되어있는 아이템들을 추천하는 데 사용된다. Tak W. Yan and Hector Garcia-Molina (1995) 는 인터넷 뉴스기사에 텍스트 필터링을 하는 SIFT라는 간단한 CBF시스템을 통해 효율적 필터링 알고리즘을 연구했으며[10],NewsWeeder[11], NewsDude[12] 등은 CBF 기반의 추천 시스템의 좋은 예라 할 수 있다. 한편, 아이템과 사용자 프로파일을 표현하는 것은 다음과 같은 벡터공간을 통해 이루어진다.

$$\{(w_{i1},\ldots, w_{in}), r_i\} \quad \text{(식1)}$$

(식1) 에서 w_{in} 은 단어에 대한 가중치, r_i 는 아이템에 대한 평가점수이다. 각각의 단어는 미리 정의된 사전(Dictionary)으로부터 오며 각 단어에 대한 가중치는 일반적으로 'Term Frequency -

Inverse Document Frequency'를 통해 주어진다. 곧 아이템에서 추출되는 특정 단어의 빈도가 높을수록, 그리고 그 특정 단어가 나오는 전체 문서의 수가 적을수록 그 단어에 대한 가중치가 높아지도록 하는 방식이다. 한편, 사용자 프로파일은 다음 식과 같이 표현된다.

$$profile_j = \sum_{i=1}^{N} r_i' \cdot w_{ij} \quad \text{(식2)}$$

여기에서 w_{ij}는 단어 j에 대한 가중치를 표현하며 r_i'는 가중된 평가값을 의미한다.

3.2. Collaborative Filtering

Collaborative Filtering(CF) 추천시스템은 과거 사용자의 평가정보를 바탕으로 선호가 비슷한 사람들을 파악한다. 그래서 이러한 유사한 선호를 가진 사람들의 의견을 이용해서 미지의 상품에 대해 추천하는 방식을 취한다. 이 방식은 CBF 방식과 같이 아이템의 구조적 컨텐츠를 고려하는 것이 아니라 아이템에 질적인 속성까지 포함하여 아이템 전반에 대한 평가에 기초해서 추천이 이루어진다. CF 초기의 연구이며 가장 잘 알려진 시스템으로는 유즈넷 뉴스의 추천에 사용된 GroupLens 시스템 [13,14] 과 음악앨범의 추천에 사용된 Ringo[15] 등이 있다. CF시스템에서 핵심이 되는 것은 사용자들간의 선호의 유사도를 측정하는 방법이다. 선호의 유사도 측정방법에는 Pearson Correlation Coefficient, Vector Similarity, Probabilistic Distance Measure, Nearest Neighbors, Bayes's Rule, Mean Squared Difference 등[16] 이있으며 이 중에서도 가장 일반적인 방법으로 Pearson Correlation Coefficient가 사용되고 있다. (식3)과 (식4)는 Pearson Correlation Coefficient의 방법을 사용했을 때 각각 사용자간 선호의 유사도 측정과 예측값을 계산하는 수식[6, 8]이다.

$$S_{a,u} = \frac{\sum_{i=1}^{m}(r_{a,i} - \overline{r_a})(r_{u,i} - \overline{r_u})}{\sqrt{\sum_{i=1}^{m}(r_{a,j} - \overline{r_a})^2 \sum_{i=1}^{m}(r_{u,j} - \overline{r_u})^2}} \quad \text{(식3)}$$

(식3)에서 $S_{a,u}$는 a와 u의 유사도(Similarity)를, $r_{a,j}$는 아이템 j에 대한 사용자 a의 평가점수를, $\overline{r_a}$는 사용자 a가 평가한 아이템의 평균치를 의미한다.

$$P_{a,i} = \overline{r_a} + \frac{\sum_{u=1}^{m}(r_{u,i} - \overline{r_u}) \cdot S_{a,u}}{\sum_{u=1}^{n} S_{a,u}} \quad \text{(식4)}$$

(식4)에서 $P_{a,i}$는 i 아이템에 대한 사용자 a의 예측값을, $r_{u,i}$는 i 아이템에 대한 사용자 u의 평가점수, $S_{a,u}$는 a(나)와 u(이웃)의 선호의 유사도를 의미한다.

CF 기반 추천시스템은 컨텐츠를 일일이 구분하여 선호프로파일을 생성하기 어려운 아이템에 대해서도 사용할 수 있는 큰 장점을 갖고 있다[7]. 하지만 이 방법은 Data Sparcity, First Rater의 문제를 갖고 있을 뿐만 아니라 사용자 수, 평가항목 수의 증가로 기하급수적으로 늘어나는 데이터베이스를 감당할 수 없는 단점도 갖고 있다.

3.3 Hybrid Model

CBF 추천시스템, CF 추천시스템 개별적으로는 모두 한계점을 갖고 있다. CBF 시스템은 대체적으로 예측률이 낮은 단점을 갖고 있다. 그리고 CF 시스템은 위에서 언급된 것과 같이 Data Sparsity, First Rater의 고전적인 문제를 가지며 최근에는 대용량의 데이터베이스로 인해 시스템적 제약을 받는 한계점을 갖고 있다. 그래서 추천시스템을 연구한 많은 연구자들은 이 두 가지 방법을 함께 적용하는 연구를 해왔다.

이 두 가지 방법을 접목한 초기의 대표적 연구로는 스탠포드 대학의 FAB 시스템이 있다 [3]. 이 방법은 사용자의 선호 프로파일을 유지하는 Selection Agent를 이용해 유사한 선호를 가진 사용자를 파악하고 자신의 이웃이 좋아한 Article을 나에게 추천하게 해 주는 방식으로 설계되었다. Claypool, Gokhale and Miranda(1999)는 CBF 방식과 CF 방식을 각각 사용하되 아이템 추천을 위해서 이들의 예측값을 가중평균하는 방법을 사용하였다[17]. 또한 Basu, Hirsh and Cohen(1998)은 추천시스템 설계의 접근법을 Inductive Learning의 관점에서 접근하여 정확한 예측값을 산출하는 것이 아니라 어떤 아이템에 대해 {liked, disliked}로 분류하는 Classification의 문제로 접근하였다[18]. 한편 Polcicova and Navrat (2000)은 영화도메인에서 Data Sparsity의 문제를 해결하기 위해서 Missing Value들에 대해서만 CBF의 방법을 적용하여 빈 값을 채우는 방법의 연구를 통해 예측의 정확성을 높였다 [2].

위의 방법들과는 달리 CF 시스템의 전반적인 프레임웍 내에서 추가적인 사용자의 정보를 이용하여 예측값의 정확도를 높이고자 한 연구들도 있었다. Kyenah Yu, Sukmin Choi(2000)는 Data Sparsity의 문제를 해결하고 예측값의 정확도를 높이기 위해서 각각의 평가아이템이 속한 카테고리 정보를 활용하고자 하였다. 즉, 각각의 카테고리도 하나의 상품아이템으로 보고, 하나의 아이템을 평가했을 때 그 아이템에 대한 일차적인 선호뿐 아니라 평가된 아이템이 속한 카테고리에 대한 내재적인 선호정보를 이용하려고 하였다[8]. Young-Suk Ryu, Taehun Kim(2001)은 CF 방법의 프레임웍에 사용자가 보여준 암묵적 선호정보를 활용하여 사용자를 군집으로 나누고 또한 '나이'라는 인구통계학적 정보를 함께 사용하여 예측의 정확도를 높이려고 하였다[7].

특히 최근의 몇몇 연구결과들은 클러스터링의 방법이 예측의 정확도를 높여줄 뿐만 아니라 대용량의 데이터로 인한 시스템적 자원의 한계에 봉착해 있는 추천시스템의 문제해결방법으로 사용될 수 있음을 보여주고 있다. 대표적으로 Badrul M. Sarwar, George Karypis(2002)은 CF 시스템을 설계할 때의 현실적인 문제를 Data Sparsity, Scalability의 두 가지로 크게 정의하면서 데이터 축소의 방법으로 클러스터링의 방법을 제시하고 있다[4].

4. 연구모델

4.1 데이터 수집과 처리

웹사이트를 구축하여 2003년 8월20일부터 2003년 9월 20일까지 한 달에 걸쳐 온라인 설문을 받았고 이와 병행하여 오프라인으로도 설문조사가 이루어졌다. 일곱 개의 장르로 영화를 분류하여 각 장르별 각각 4편의 영화를 선정하여 총 28개의 영화에 대한 평가를 받았다. 평가점수는 중간값 4점부터 '최고의 영화' 7점, '최악의 영화' 1점에 이르기까지 7단계로 구별하였다. 음악장르에 대한 선호를 파악하기 위해서는 개별 음반에 대해 질문하지 않고 대중음악, R&B/Soul, 힙합/래게, 가스펠, 락, 재즈, 클래식, 뉴에이지의 8개의 음악장르 자체에 대한 선호점수를 역시 7점 척도

로 입력받았다.

총 608명이 설문조사에 참여하였다. 하지만 신뢰성 있는 데이터 분석을 위해 15개 이상의 영화에 대해 평가한 사람만을 대상으로 표본을 축소하였고 그 결과 481개의 데이터셋을 모을 수 있었다.

4.2 '음악' 도메인에서 사용자 군집을 파악

본 연구에서는 K-means 방법을 이용하여 군집을 파악하였다. 군집의 수 K를 결정하는 것은 원칙적으로 자료탐색을 통해 알아내야 하지만 많은 데이터를 갖는 자료의 경우 이 방법이 현실적으로는 쉽지 않기 때문에 이해와 관리 가능한 수준에서 결정하는 것이 보통이다[19]. 본 연구에서는 2개에서 5개까지로 군집의 수를 적용시켜 보면서 자료의 특성을 분석하여 특성있게 군집이 분류되는 수준인 3개의 분류로 확정하게 되었다. 최종적으로 각 군집에 속하는 레코드 수는 첫째로 첫번째 군집이 144개, 두번째 군집이 198개, 세번째 군집이 139개이다.

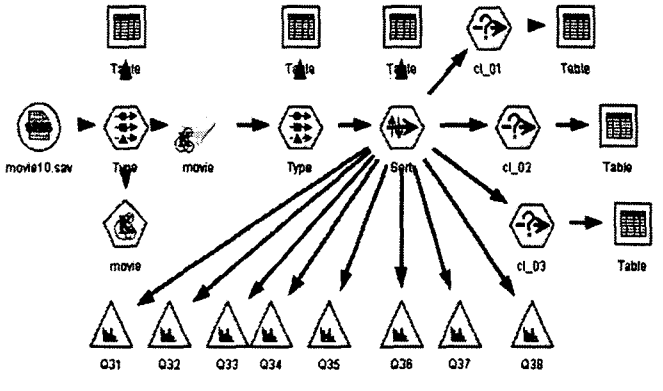

그림1: 클레멘타인7.0에서의 군집분석 스트림

<그림1>에서는 소스데이터 movie10.sav를 Type노드를 통해 음악장르에 관한 선호를 묻는 8개의 문항의 데이터타입을 연속형으로 설정해 주고 K-Means 모델을 돌린 모습이다. 사전에 지정한 대로 3개의 군집으로 나뉘었으며 각각의 군집에 대한 특성을 파악하기 위해 Distribution 노드를 통해 8개의 문항별 군집의 분포를 파악하였다.

군집1은 전반적으로 음악을 즐기지 않는 부류이지만 클래식과 가스펠 음악에 대해서는 상대적으로 높은 선호수준을 보였다[그림2].

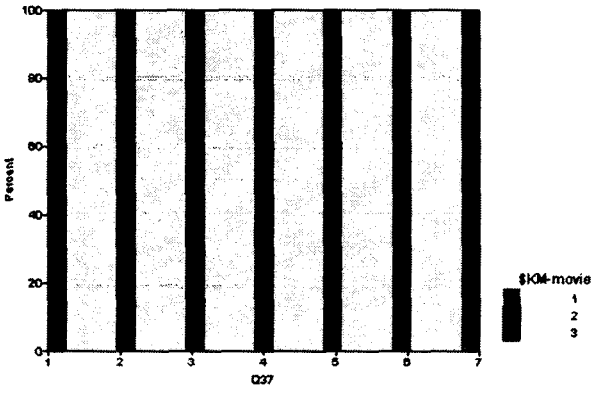

그림2: 가스펠 음악에 대한 정규화된 선호분포

군집2는 거의 대부분의 장르를 좋아하되 특히 락, 클래식, 재즈, 뉴에이지 음악에 대해 높은 선호수준을 보였다[그림3].

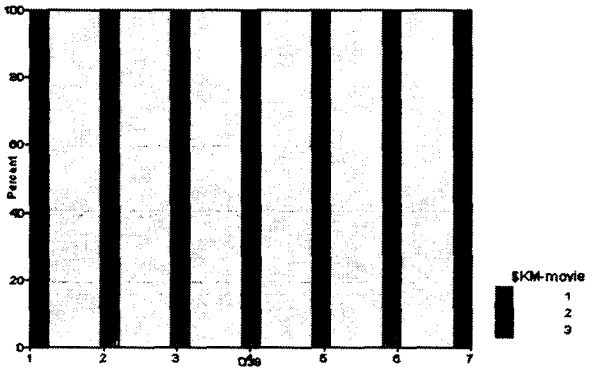

그림3: 뉴에이지 음악에 대한 정규화된 선호분포

군집3은 대체로 모든 장르에 걸쳐 평균적인 음악선호를 보이되 힙합/래게 등의 빠른 음악을 상대

적으로 좋아하는 것으로 나타났다[그림4].

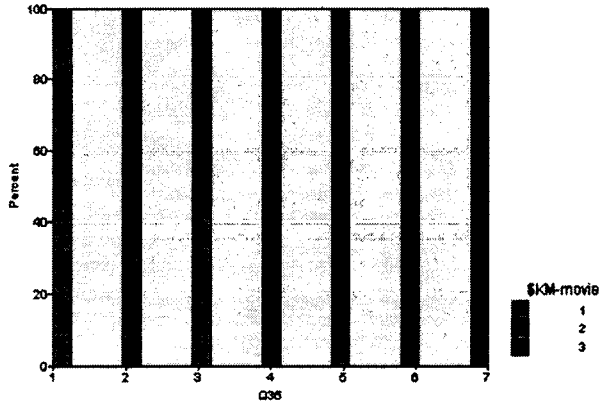

그림4: 힙합, 래게 음악에 대한 정규화된 선호분포

4.3 군집별로 'Collaborative filtering'의 적용과 예측률의 평가

추천시스템의 정확도를 측정하는 데에는 Coverage, Mean Absolute Error(MAE), Accuracy, Precision, Recall, F-measure 등의 다양한 기준이 있다[2]. 본 실험에서는 가장 간단하고 널리 사용되는 MAE를 이용하여 예측의 정확도를 평가하였다. 이 방법은 사용자가 실제 평가한 값과 예측값과의 차이를 비교하여 오차를 계산하고 이 오차의 평균을 계산하는 통계적인 측정방법이다. 따라서 MAE의 값이 작을수록 더 정확한 예측이 이루어졌다고 할 수 있다[7].

$$\left|\overline{E}\right| = \frac{\sum_{i=1}^{N} \left|\varepsilon_i\right|}{N} \qquad \text{(식5)}$$

(식5)에서 $\left|\overline{E}\right|$ 는 오차의 평균을 즉 MAE를, N 은 총 평가아이템 수를, $\left|\varepsilon_i\right|$ 는 각각의 아이템에 대한 오차의 절대값을 의미한다.

실험을 위해서 세 개로 나누어진 각각의 클러스터에서 10명씩의 사용자를 뽑아 이들이 평가한 영화 중에 3개씩을 임의로 선정하였다. 따라서 평가가 이루어진 아이템의 수는 각 군집별로 30개씩 총 90개의 아이템이다.

userid	question	rating	cluster1	comparison	difference1	difference2
18	q10	4	4.189333	4.348609	0.189333	0.348609
18	q15	6	6.257462	6.331724	0.257462	0.331724
18	q2	7	6.414653	5.978673	0.585347	1.021327
36	q17	4	3.913551	4.211457	0.086449	0.211457
36	q2	4	3.90616	3.923237	0.09384	0.076763
36	q23	5	5.401336	5.480362	0.401336	0.480362
39	q2	6	5.537693	5.332785	0.462307	0.667215
39	q23	6	6.321533	6.158889	0.321533	0.158889
39	q24	4	3.841184	4.176363	0.158816	0.176363
46	q17	5	4.190422	4.734223	0.809578	0.265777
46	q2	2	1.606253	4.160268	0.393747	2.160268
46	q21	7	6.92536	6.226683	0.07464	0.773317
55	q1	6	5.760901	4.901826	0.239099	1.098174
55	q15	4	5.580868	5.60381	1.580868	1.60381
55	q16	2	3.694949	3.737653	1.694949	1.737653
62	q13	5	5.077529	2.453466	0.077529	2.546534
62	q14	5	5.887905	4.34266	0.887905	0.65734
62	q15	3	3.424498	1.30985	0.424498	1.69015
64	q1	4	3.561964	3.595898	0.438036	0.404102
64	q10	6	7.570515	4.56097	1.570515	1.43903
64	q11	4	3.610659	4.307367	0.389341	0.307367
66	q13	6	5.213541	5.139853	0.786459	0.860147
66	q15	6	5.522417	5.471236	0.477583	0.528764
66	q17	3	3.540031	3.953993	0.540031	0.953993
					15.69019	24.09317

그림5: 클러스터 1의 실험결과

<그림5>는 SQL2000 Server에서 구현한 Prediction Function의 결과값을 엑셀로 정리한 표의 일부이다. rating은 실제 사용자가 영화에 대해 평가한 값이며 Cluster1 칼럼은 첫번째 군집에서 Prediction Function을 돌렸을 때 나온 예측값, comparison 칼럼은 군집을 나누지 않은 상태에서 Prediction Function을 돌렸을 때의 예측값이다. difference1, difference2 칼럼은 각각 군집을 나누었을 때와 나누지 않았을 때 실제값과 예측값의 차이 즉, 오차의 절대값이다.

<표1>은 MAE를 측정기준으로 한 각 클러스터별 결과값이다.

표1: 군집별 MAE 수치

	MAE 1	MAE 2
Cluster1	0.50973	0.80125
Cluster2	0.534547	0.707416
Cluster3	0.43773	0.614796

이 표는 동일한 사용자와 평가항목에 대하여 군집을 나누었을 때와 군집을 나누지 않았을 때 예측의 정확도를 비교해 놓은 표이다. MAE1은 군집을 나누었을 때이며 MAE2는 군집을 나누지 않았을 때의 결과이다. 위 표를 기준으로 볼 때 군집을 나누었을 때가 군집을 나누지 않았을 때에 비해 추천의 정확도가 일관되게 높음을 알 수 있다.

4.4 '도메인에 대한 관심정도'에 따라 그룹 분류

설문문항 중 한달 동안 영화를 보는 횟수를 파악하는 문항이 있었다. 이 중에서 극단적인 그룹을 파악하기 위해 한 달에 영화(비디오 포함)를 2회 미만으로 보는 그룹과 7회 이상으로 보는 그룹으로 나누었다. 이 실험의 전제는 다음과 같다.

H: 영화에 관심이 많은 사람들로 구성된 집단에서 CF 시스템의 정확도가 높게 나타난다.

7회 이상으로 영화를 보는 사람은 전체 481명 중에 37명이었고 이들과의 동일한 사람 수에서 비교하기 위해 2회 미만의 사람들 중에 37명을 뽑아 또 하나의 집단을 구성하였다. 전 실험과 마찬가지로 이들 중에서도 10명씩을 임의로 뽑아 각각이 평가한 영화 중에 3개씩을 임의로 각각 30개씩에 대해 실험을 실시하였다.

표2: 도메인에 대한 관심에 따른 MAE 값

	관심높음	관심낮음
MAE	0.92674	0.72731

<표2>에서 보는 바와 같이, 영화에 대해 관심이 많은 집단의 MAE값은 0.92674로 오히려 영화에 대해 관심이 적은 집단의 0.72731 값에 비해 낮게 나타났다. 이처럼 선호가 유사한 사람을 찾는데 있어서는 그 집단이 더 그 도메인에 대해 관심을 갖고 있는 집단이던 그렇지 않든지 간에 유의한 차이가 없으며 오히려 본 실험에서는 예상과 달리 도메인에 관심이 없는 집단에서 실시한 예측률의 결과가 높게 나타났다.

5. 결론

CF 기반의 추천시스템은 영화와 음악, 미술작품 등과 같이 아이템의 질적인 요인이 선호의 중요한 기준이 되는 도메인에 있어서 유용하게 사용될 수 있는 방법이다. 하지만 '사용자의 수', '평가 아이템의 수'가 기하급수적으로 증가하게 되는 최근의 환경에서는 계산량의 급증으로 현실적인 적용이 어려운 한계점이 있다. 또한 시스템 초기에는 사용자가 너무 적어 Data Sparsity, First Rater의 문제가 발생하고 예측의 정확도도 떨어지는 단점이 있다.

본 연구에서는 대용량화되고 있는 데이터를 축소하고 예측의 정확도를 제고하기 위해 군집분석의 방법을 사용해 보았다. 실험의 결과는 또 다른 속성을 도입하여 군집을 나누는 방법을 사용하였을 때 예측의 정확도가 일관되게 높아짐을 보여주었다. 그러나 가능한 사용자에게 정보를 적게 요구하는 것을 미덕으로 삼는 CF 시스템에서는 군집을 나눌 수 있는 속성을 찾는 것이 쉽지 않다. 따라서 본 실험과 같이 다른 도메인에서 파악된 고객의 선호에 대한 추가적 정보를 이용하여 시스템의 성능을 높일 수 있다는 것은 매우 고무적인 결과라 할 수 있다. 또 다른 실험결과로 영화에 대한 관심이 많은 사람과 관심이 아주 작은 사람들끼리 따로 군집을 분류하여 실험해본 것은 기대와 달리 오히려 영화에 대한 관심이 작은 집단의 예측률이 높게 나왔다.

하지만 이 실험은 초기표본의 크기가 608개로 작고 실험여건의 제약으로 군집을 통해 데이터를 줄였을 때 계산의 속도가 얼만큼 단축되었는지 확인하는 것이 쉽지 않았다. 그리고 군집을 나눌 때

사용한 속성을 음악이라는 도메인에서 파악한 것은 음악과 영화의 선호가 어느정도 상관관계가 있을 것이라는 가정에 기초하고 있다. 그러나 이 관계는 명확한 논리적 인과관계가 뒷받침 된 것은 아니다. 향후 연구에서는 사용자의 군집을 나눌 때 사용한 속성을 여러 개로 늘려보고 도메인간의 선호의 상관관계에 대한 보다 정교한 분석이 필요할 것이다. 또한, 군집분석을 통해 기대되는 효과는 예측률의 제고뿐 아니라 계산량의 감소로 인한 속도의 향상이기 때문에 이것을 평가할 수 있는 방법들을 향후 모색해야 할 것이다.

[참고문헌]

[1] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. "Analysis of Recommender Algorithms for E-Commerce,"In Proceedings of the ACM E-Commerce 2000 Conference, pp. 158-167, 2000

[2] G. Polčicová, P. Návrat. "Combining Content-based and Collaborative Filtering," In Proceedings of The International Workshop on Application of Advanced Information Technologies to Medicine, pp.103-116, 2000

[3] Marko Balabanovi´c and Yoav Shoham. "Fab: Content-Based, Collaborative Recommendation," Communications of The ACM, Vol. 40, No. 3, 1997

[4] B.M. Sarwar, G. Karypis, J. Konstan, and J. Riedl. "Recommender Systems for Large-Scale E-Commerce: Scalable Neighborhood Formation Using Clustering," Proceedings of the Fifth International Conference on Computer and Information Technology, 2002

[5] Ungar, L. H., and Foster, D.P. "Clustering Methods for Collaborative Filtering," In Workshop on Recommender Systems at the 15th National Conference on Artificial Intelligence, 1998

[6] Paolo Buono, Maria Francesca Costabile, Stefano Guida, Antonio Piccinno and Giuseppe Tesoro. "Integrating User Data and Collaborative Filtering in a Web Recommendation System," Proceedings of the Eight International Conference on User Modeling, 2001

[7] Young-Suk Ryu, Taek-Hun Kim, Ji-sun Park, Seok-In Park and Sung-Bong Yang. "Using Content Information for Finding Neighbors in the Collaborative Filtering Framework,"; Proceedings of International Conference in Electronic Commerce, 2001

[8] Kyeonah Yu, Sukmin Choi, Juntae Kim. "Improving the Performance of Collaborative Recommendation By Using Multi-Level Similarity Computation," Proceedings of the IASTED International Conference on Artificial Intelligence and Soft Computing, 2000

[9] Good, N., Schafer, J.B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., and Riedl, J. "Combining Collaborative Filtering with Personal Agents for Better Recommendations," Proceedings of the 1999 Conference of the American Association of Artifical Intelligence. pp 439-446, 1999

[10] Tak W. Yan and Hector Garcia-Molina. SIFT --- A tool for wide-area information dissemination. In Proceesings of the 1995 UNSENIX Technical Conference, pages 177--186, 1995

[11] Lang, K.: NewsWeeder: Learning to filter Netnews. In Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA, 1995

[12] Billsus, D., Pazzani, M.J.: A Hybrid User Model for News Story Classification. In: J. Kay(ed.) User Modeling '99-Proceedings of the 7th International Conference. Springer-Verlag, Wien-New York 99-108, 1999

[13] Resnick, P., Iacovou, N., Sushak, M., Bergstrom, P., and Riedl, J. "GroupLens: An open architecture for collaborative filtering of netnews" Proceedings of the 1994 Computer Supported Collaborative Work

Conference, 1994

[14] http://www.grouplens.org

[15] U.Shardanand and P. Maes. "Social Information Filtering: Algorithms for Automating "Word of Mouth" In Proceedings of the 1995 ACM Conference on Human Factors in Computing Systems, str.210-217, New York, 1995.ACM, 1994

[16] Paulson, Patrick and Aimilia TzanavariU.(2003) "Combining Collaborative and Content-Based Filtering Using Conceptual Graphs". Book chapter in: J.Lawry, J.G.Shanahan and A.Ralescu (eds.). Modeling with Words: Learning, Fusion, and Reasoning within a Formal Linguistic Representation Framework, LNAI 2873, Springer-Verlag Berlin Heidelberg 2003, pp. 168-185.

[17]Claypool, M., Gokhale, A.k Miranda, T. "Combining Content-Based and Collaborative Filters in an Online Newspaper," In Proceedings of the ACM SIGIR Workshop on Recommender Systems, 1999

[18] Basu, C., Hirsh, H.,Cohen, W. "Recommendation as Classification: Using Social and Content-Based Information in Recommendation," In Proceedings of the 15th National Conference on Artificial Intelligence, 1998

[19] 허명회, 이홍구. 데이터마이닝 모델링과 사례, SPSS 아카데미, pp63-74, 2003