디지털산업정보학회 논문지 제6권 제3호-2010년 9월

추천시스템을 위한 연관군집 최적화 기반 협력적 필터링 방법

이 현 진* · 지 태 창**

An Collaborative Filtering Method based on Associative Cluster Optimization for Recommendation System

Lee, Hyun Jin · Jee, Tae Chang

〈Abstract〉

A marketing model is changed from a customer acquisition to customer retention and it is being moved to a way that enhances the quality of customer interaction to add value to our customers. Such personalization is emerging from this background. The Web site is accelerate the adoption of a personalization, and in contrast to the rapid growth of data, quantitative analytical experience is required. For the automated analysis of large amounts of data and the results must be passed in real time of personalization has been interested in technical problems.

A recommendation algorithm is an algorithm for the implementation of personalization, which predict whether the customer preferences and purchasing using the database with new customers interested or likely to purchase. As recommended number of users increases, the algorithm increases recommendation time is the problem.

In this paper, to solve this problem, a recommendation system based on clustering and dimensionality reduction is proposed. First, clusters customers with such an orientation, then shrink the dimensions of the relationship between customers to low dimensional space. Because finding neighbors for recommendations is performed at low dimensional space, the computation time is greatly reduced.

Key Words : Recommender system, collaborative filtering, Multidimensional Scaling, Personalization

Ⅰ. 서론

인터넷의 활성화로 인하여 다양한 정보가 인터넷 상에서 사용자에게 제공되고 있다. 추천시스템은 다양한 정보 중에서 사용자가 원하는 정보를 획득하도록 도와주는 방법이다. 추천시스템은 인터넷 기반의 고객 개인에 대한 일대일 마케팅을 할 수 있도록 하는 e-CRM의 한 분야이며, Amazon, CDnow 등 전자상거래 사이트에 적용되어 활용되고 있다[1]. 또한 유비쿼터스 컴퓨팅 환경하에서 컨텍스트(context) 정보를 기반으로한 추천 시스

* 한국사이버대학교 컴퓨터정보통신학과 부교수(교신저자)
** 연세대학교 컴퓨터과학과 박사과정

템에 관한 연구도 진행되고 있다[2].

개인화 서비스는 고객들이 필요로 하는 제품을 명시적으로 묻지 않고 제공하는 서비스이다. 개인에 대한 정보를 바탕으로 서비스가 제공되기 때문에 서비스 제공자와 개인간의 정보교류가 원활할 때 효과적으로 이루어진다. 개인화 추천시스템은 학습과정과 정보필터링 과정으로 구성된다. 학습과정은 사용자 행위에 따라 사용자의 성향을 학습하는 것이며, 정보 필터링은 사용자에 따른 추천 정보를 나타내는 것이다[3].

추천시스템의 정보필터링 구현 기법은 규칙기반 필터링(rule-based filtering), 내용기반 필터링(content-based filtering), 협력적 필터링(collaborative filtering) 방식이 있다. 규칙기반 필터링 방식은 사용자의 과거행위나 명시적 정보에 의해 미리 생성된 규칙을 적용하는 방법이다. 내용기반 필터링 방식은 다른 사람의 평가와 무관하게 미리 평가한 아이템을 기반으로 추천하는 방법이기 때문에 다른 사용자의 선호도에 영향을 받지 않는다. 협력적 필터링 방식은 유사한 취향의 고객을 찾아서 그 고객이 좋아하는 아이템을 추천하는 방식이다[4-5]. 오늘날 추천시스템에서 가장 많이 사용되는 방식은 협력적 필터링 방식이다[6]. 하지만 이러한 협력적 필터링 방식은 새로운 입력에 대한 정보 부족으로 추천 정확도가 저하되는 희박성(sparsity)문제, 이용자수의 증가에 따라 추천시간이 증가하는 확장성(scalability)문제가 발생한다. 또한 선호도 예측 과정에서 사용자의 선호도 계산시간의 부담의 문제도 존재한다.

본 논문에서는 이용자수의 증가에 따라 추천시간이 증가하는 것을 방지하기 위하여 K-means 군집화 알고리즘을 적용하여 유사한 선호도를 가지는 고객들로 탐색공간을 줄이는 방법을 사용하였다. 또한 고객에게 아이템을 추천할 때에는 해당 고객이 속한 군집내의 고객들에 대하여 저차원사상을 통해 거리개념으로 표현한 선호도를 기반으로 추천하는 방법을 제안하였다.

본 논문의 구성은 다음과 같다. 2장에서 협력적 필터링에 관계된 관련연구들을 살펴본다. 3장에서는 제안하는 방법의 구성과 사용된 방법들을 분석하다. 4장에서는 실험환경과 실험 결과를 분석하고 5장에서 결론을 맺는다.

Ⅱ. 관련연구

협력적 필터링 방식은 개인화 추천 시스템에서 가장 많이 사용되는 방식이다. 하지만 이 방식은 초기 평가치 문제, 희박성 문제, 확장성문제들이 존재한다. 따라서 이를 해결하기 위하여 내용기반 필터링, 최근접 이웃 알고리즘, 베이시안 분류기, 군집화 방법 등 다양한 방법을 결합하는 연구들이 진행되고 있다[7].

미네소타 대학의 GroupLens 프로젝트에서는 피어슨 상관계수(Pearson correlation coefficient)를 이용하여 사용자간의 유사성을 구하고, 하나의 아이템에 대한 사용자의 선호도를 계산하기 위해서 다른 모든 사용자와의 유사도를 계산하여 이를 바탕으로 다시 선호도 값을 계산하였다. 이러한 방법은 사용자가 많은 시스템에 실시간으로 적용하기에는 많은 연산시간을 요구하고, 예측 정확도 측면에서 비효율적이다. 또한 처음 방문한 고객에게는 기본 프로파일 정보이외에 어떤 선호 경향도 파악할 수 없기 때문에 정확도 측면에서 문제가 있다[8].

Good등[9]은 다수의 평가 에이전트(agent)를 활용하여 사용자가 평가하지 않은 많은 항목들에 대해 자동으로 평가를 수행하도록 하여 희소성문제를 해결하는 연구를 수행하였다. Herlocker등[10]은 정확도를 향상시키기 위해서 공통으로 평가한 항목의 수를 유사도에 반영시키는 유사도 계산방식을 제안하였다. Breese는 피어슨 상관계수와 벡터 유사도와 같은 메모리 기반 기법을 사용하고, 각각에 기본값을 사용하여 정확도와 수렴도를 향상시키는 방법을 제안하였다[11]. Sarwar등은 SVD를 활용하여 사용자-항목 평가 행렬의 차원을 축소하여 예측 하는 방법과 개별고객들간의 유사도를 계산하는 대신 아이템들간의 유사도 계산하여 계산량을 줄이는 방법을 수행

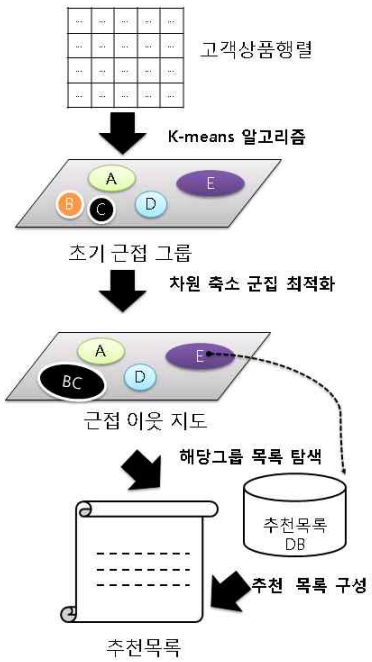
하였다[12-13]. Linden등은 고객차원의 축소, 상품차원 축소방법을 수행하였다[14]. Li, Xue등은 고객들을 군집화후 군집들간의 유사도를 계산하여 계산량을 줄이는 연구를 수행하였다[15-17]. 군집화를 수행하여 군집들간의 유사도를 계산하는 것으로 계산량이 감소될 수 있지만 여전히 각각의 아이템들을 고려하려면 계산량은 여전히 많게 된다. 따라서 본 연구에서는 군집화를 수행하여 협력적 필터링을 구현하는데 있어서 고객들의 군집을 보다 효율적으로 구성하고, 속도를 향상시키는 방법을 제안하고자 한다.

Ⅲ. 제안하는 시스템

3.1 제안하는 시스템의 구성

추천시스템의 성능향상을 위해서는 고객의 수와 거래 데이터의 개수가 늘어남에 따라 목표고객의 최근접 이웃을 찾기 위한 연산이 늘어나는 것을 고려해야 한다. 실시간으로 동작되는 시스템에 추천을 위한 프로세스를 추가할 경우 계산시간은 중요하다. 본 논문에서는 실시간 시스템에 적용할 수 있는 추천 시스템을 제안하고자 한다.

제안하는 협력적 필터링에 기반을 둔 추천 시스템의 구성은 다음과 같다. 첫 번째 단계에서는 고객과 상품간의 평가치 매트릭스를 구성한다. 두 번째 단계에서는 유사한 선호도를 갖는 고객들을 군집화하여 최적화된 근접 이웃지도를 구성한다. 군집화 방법은 자기조직화 신경망(Self Organizing Map)을 이용하는 방법과 완전연결(Complete Linkage)방법에 비해 유한시간 내에 결과를 낼 수 있고, 특징벡터가 커져도 계산시간이 점진적으로 증가하는 분할 군집화(partitional clustering) 방법인 K-means를 기반으로 차원 축소와 군집 병합 방법을 결합한 알고리즘을 제안한다. 첫 번째 단계와 두 번째 단계는 기 축적된 데이터를 이용하여 고객에 대한 특성을 찾아서 새로운 데이터에 대한 추천 목록을 구성할 수 있는 메타데이터를 구축하는 단계이다. 마지막 단계는 새로운 데이터에 대하여 이전 단계에서 생성된 메타데이터를 이용하여 선호도를 예측하여 추천 목록을 구성하는 단계이다. 선호도는 k-NN을 사용하여 새로운 고객의 해당 제품에 대한 선호도를 예측한다.



<그림 1> 제안하는 시스템의 구성

본 논문에서는 선호도 예측과정에서 이웃선정 과정과 유사도 가중치 계산과정에 계산상의 부담은 최소화 하면서 유사성을 향상하기 위하여 다차원 척도법(MDS: Multi-Dimensional Scaling)에 기반하여 근접 이웃을 형성하였다. 다차원 척도법은 다차원의 정보를 2차원 평면상으로 축소하여 표현하면서도 다차원상의 관계는 유지하고 있기 때문에 계산량을 줄일 수 있다. 뿐만 아니라 그들의 관계를 거리개념으로 표현할 수 있기 때문에 유클리디안 거리를 이용하여 관계성 분석이 가능하다. 따라서 차원축소에 의한 최적 군집화 지도를 이용하면 새로운 고객에 대하여 아이템을 추천하는데 있어서 모든

아이템의 차원에 기반하여 탐색하는 것이 아니라 축소된 차원에 기반하여 탐색이 가능하기 때문에 탐색 속도를 향상 시킬 수 있다.

3.2 초기 그룹 형성과 평가 척도

유사한 고객의 그룹 형성을 위하여 K-means 군집화 알고리즘에 의해 초기 그룹을 형성한다. K-means는 n개의 입력 데이터를 K개의 군집으로 분할하는 방법이다 [18-19]. 군집화를 수행하는데 있어서 유사성 계산척도에 따라 그 성능의 차이를 보인다. 많이 사용되는 유사도 척도는 유클리디안 거리(Euclidean distance), 맨하탄 거리(Manhattan distance), 피어슨 상관 계수(Pearson correlation coefficient), 코사인 상관 계수(Cosine correlation coefficient) 등이며 이를 계산하는 공식은 <표1>과 같다. 여기서 M 은 패턴의 개수 이고, $j \in (1 \ldots M)$이다. N은 입력데이터의 개수이고, $i \in (1 \ldots N)$ 이다. K 는 군집의 개수이고, $k \in (1 \ldots K)$이다. c_{kj}는 k번째 군집의 j번째 특징값 이고, x_{ij}는 i번째 입력데이터의 j번째 특징을 의미한다. 본 논문에서 사용한 MovieLense 데이터에서 패턴의 개수 M 은 사용자가 평가한 영화의 수이고, 입력 데이터의 개수 N은 영화를 평가한 사용자의 수를 의미한다. c_{kj}는 k번째 군집의 j번째 영화의 특징값이고, x_{ij}는 i번째 사용자의 j번째 영화에 대한 특징값이다.

K-means 알고리즘에서 사용되는 유사성 척도 중에서 유클리디안 거리, 맨하탄 거리는 희박성 데이터에 대해서는 좋은 성능을 보이지 못하고 있다. 따라서, 데이터가 희박하게 분포하는 다차원 데이터에 대해서는 유클리디안 거리와 맨하탄 거리를 사용하기 어렵다. 피어슨 상관계수와 코사인 상관 계수는 희박한 데이터에 대해 좋은 성능을 보이고 있어서 다차원 데이터에 대한 K-means 알고리즘의 유사성 척도로 사용할 수 있다. 계산량 관점에서 살펴보면, 피어슨 상관 계수는 코사인 상관 계수에 비해서 더 많은 계산이 필요하기 때문에 속도가 느린 단점이 있다. 아이템들이 많은 추천시스템에 피어슨 상관계수를 유사도 척도로 사용하기에는 계산상의 부담이 크고, 온라인 시스템의 경우 그 차이가 커지게 된다. 따라서 본 논문에서는 희박성 문제에 덜 민감하고 계산상의 부담이 적은 코사인 상관계수를 척도로 하여 군집을 형성하였다. 군집을 형성하는데 사용한 K-means 알고리즘은 <표2>와 같다.

<표 1> 유사도 척도

방법	수식
유클리디안 거리	$d = \sqrt{\sum_{j=1}^{M}(x_{ij} - c_{kj})^2}$
맨하탄 거리	$d = \mid \sum_{j=1}^{M}(x_{ij} - c_{kj}) \mid$
피어슨 상관계수	$d = \dfrac{\sum_{j=1}^{M} x_{ij}c_{kj} - \dfrac{\sum_{j=1}^{M} x_{ij}\sum_{j=1}^{M} c_{kj}}{M}}{\sqrt{\left(\sum_{j=1}^{M} x_{ij}^2 - \dfrac{\left(\sum_{j=1}^{M} x_{ij}\right)^2}{M}\right)\left(\sum_{j=1}^{M} c_{kj}^2 - \dfrac{\left(\sum_{j=1}^{M} c_{kj}\right)^2}{M}\right)}}$
코사인 상관계수	$d = \dfrac{\sum_{j=1}^{M} x_{ij}c_{kj}}{\sqrt{\sum_{j=1}^{M} x_{ij}^2 \sum_{j=1}^{M} c_{kj}^2}}$

<표 2> K-means 알고리즘

1단계: K개의 초기 중심값을 $C_i (i \in (1 \ldots K))$를 초기화한다.

2단계: M개의 입력데이터 $X_j (j \in \{1 \ldots M\})$와 중심 C_i 사이의 거리를 계산하고, 가장 최단거리의 중심에 할당한다.

$$u_{ij} = \begin{cases} 1 \text{ if } \dfrac{x_j c_i}{\sqrt{x_j^2 c_i^2}} \leq \dfrac{x_j c_k}{\sqrt{x_j^2 c_k^2}} \\ 0 \; otherwise \end{cases} \quad (1)$$

이때 $k \in \{1 \ldots K, k \neq i\}, j \in \{1 \ldots M\}$

3단계: 각 군집의 중심을 재계산한다.

$$c_i = \frac{1}{N_i}\sum_k x_k \quad (2)$$

여기서 N_i는 i군집에 할당된 입력데이터의 수

4단계: 입력데이터의 소속 군집에 변화가 없을 때 까지 2-3단계를 반복한다.

3.3 군집 최적화

본 논문에서는 유사성향의 그룹의 유사성을 극대화 시키고, 온라인에 적용시킬 수 있도록 계산량을 줄 일 수 있는 군집 최적화 방법을 적용하였다. 다차원 척도법을 통하여 다차원의 정보를 2차원평면으로 사상하여 근접 이웃 지도를 생성하였다. 이렇게 생성된 평면하에서 군집최적화를 수행하고, 이렇게 형성된 근접 이웃 지도를 기반으로 추천목록을 형성한다.

다차원 척도법은 다차원 공간상에서 자극 좌표 또는 가중치를 유도하기 위하여 유클리디안과 가중치 유클리디안 모형을 이용한다. 데이터들의 근접성 자료를 처리하여 다차원 공간상의 데이터들을 저차원 공간에 위치적으로 표시하는 일련의 통계 기법들을 말한다. 근접이웃 지도를 형성하기 위하여 적용된 다차원 척도법은 <표3>과 같다.

<표 3> 다차원척도법 알고리즘

1단계:	n개의 데이터를 p차원 공간의 임의의 지점에 할당한다.
2단계:	각 점의 쌍들의 거리를 계산해서 $n \times n$ 거리 매트릭스 $[d_{ij}]$를 구한다. $$d_{ij} = \left(\sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert^{r} \right)^{\frac{1}{r}}, r \geq 1 \quad (3)$$
3단계:	$[d_{ij}]$ 매트릭스와 $[\delta_{ij}]$ 매트릭스를 스트레스 함수 S를 통해 비교하게 된다. 이 스트레스 값이 작을수록 두 매트릭스의 연관도는 큰 것이다. $$S = \left\{ \frac{\sum_i \sum_j (d_{ij} - f(\delta_{ij}))^2}{\sum_i \sum_j d_{ij}^2} \right\}^{1/2} \quad (4)$$ 이때, d_{ij}는 저차원으로 축소후 계산된 거리이며, δ_{ij}는 고차원인 입력 데이터 거리, $f(\delta_{ij})$는 입력 데이터를 크기순서로 변환하는 함수이다.
4단계:	스트레스 함수의 값이 최소화 되도록 점의 좌표를 조정한다.
5단계:	스트레스 함수의 값이 최소가 될 때 까지 2-4단계를 반복한다.

다차원 척도법은 다차원 공간상에서 계산된 데이터의 분산을 저차원 공간상에 가장 잘 유지시킬 수 있는 방법이다. 다차원 공간상 데이터의 선형 부분공간 (linear subspace)에 존재하는 진정한 구조를 발견하는 것을 보장하는 것이 밝혀져 있다[20-21]. 따라서 다차원 공간의 특성을 저차원으로 축소하여도 그 관계를 유지하면서 계산량을 감소시킬 수 있다.

다차원 척도 방법에 기반한 근접 이웃 지도는 2차원 거리형태로 표현되기 때문에 시각적으로 표현 할 수 없는 다차원 정보를 인간이 이해 할 수 있는 2차원으로 표현할 수 있다. 뿐만 아니라 2차원으로 표현가능하기 때문에 관계성을 2차원 거리 형태로 판단 할 수 있다. 따라서 추천을 위하여 근접이웃을 판단하는데 있어서 다차원의 데이터에 분석이 아닌 간단한 2차원 거리 형태로 판단할 수 있 수 있어 추천시간에 대한 계산량을 감소 시킬 수 있다.

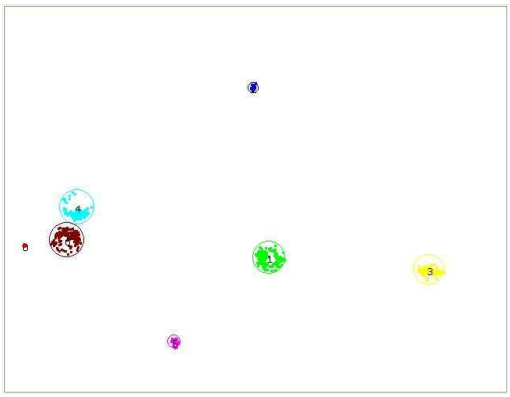
<표 4> 군집들간의 상관관계의 예

	0	1	2	3	4	5	6
0	-	1.26	1.33	1.42	1.09	1.22	1.05
1	1.26	-	1.20	1.08	0.98	1.10	1.08
2	1.33	1.20	-	1.26	1.13	1.34	1.14
3	1.42	1.08	1.26	-	1.32	1.22	1.30
4	1.09	0.98	1.13	1.32	-	1.14	0.61
5	1.22	1.10	1.34	1.22	1.14	-	0.93
6	1.05	1.08	1.14	1.30	0.61	0.93	-

<표 4>는 3.2절의 초기 군집화 알고리즘에 의해서 생성된 7개의 군집들간의 코사인 상관계수를 구한 결과에 아크코사인(ArcCosine)함수를 적용한 결과이다. 코사인 상관계수는 작은 값일수록 상관 관계가 낮은 것이기 때문에 다차원 척도법에 바로 적용하기 어렵다. 따라서, 아크코사인(ArcCosine) 함수를 이용하여 0에 가까울수록 상관도가 더 높아지도록 변환하였다.

<그림 2>는 <표 4>의 군집들 간의 상관 관계를 <표 3>의 다차원 척도법 알고리즘에 적용한 결과이다. <표 4>와 같이 0-4-6 군집들이 서로 가까이 모여 있는 것을 확인할 수 있다. 이 들 군집에 속한 데이터들은 다른 군집보다 더 연관성이 높다고 볼 수 있기 때문에 이렇게

충분히 가까이 있는 군집들은 서로 결합하는 것이 더 좋은 결과를 보일 수 있다.



<그림 2> 초기 군집 지도의 예

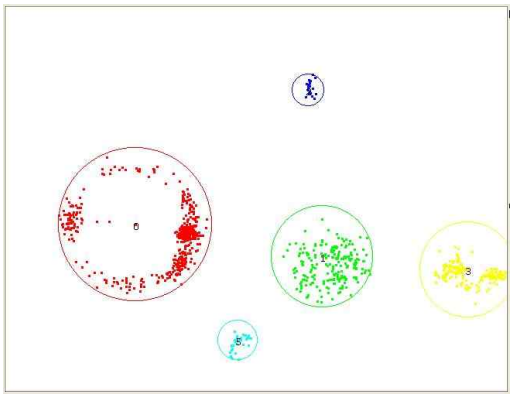
근접 이웃 지도상에서 군집을 최적화 할 때는 계층적 군집화(Hierarchical Clustering)를 수행한다. 계층적 군집화는 군집간의 거리 정보를 트리(tree) 형태로 구성한 덴드로그램(dendrogram)을 이용하여 리프(leaf)부터 루트(root) 방향으로 가장 가까운 군집들을 묶어서 새로운 군집을 형성하여 가는 방법이다. 이 방법은 군집의 개수를 미리 알 필요가 없다는 장점이 있다. 하지만 덴드로그램을 구성 한 후 군집을 형성하기 위하여 어디까지 융합(merge)할 것인지를 결정한다는 문제점이 존재한다.

계층적 군집화의 종료 조건을 결정하는 방법은 군집의 수를 이용하는 방법과 임계치를 이용하는 방법이 있다. 군집의 수를 이용하는 방법은 먼저 최종적으로 결정하고자 하는 군집의 개수를 선정 한 후 융합하면서 남아있는 군집의 수를 목표 군집의 수와 비교하는 방법이다. 임계치를 이용하는 방법은 군집들간의 융합을 결정하기 위한 최대 거리를 선정한 후 덴드로그램의 최소 거리가 임계치에 도달할 때까지 군집화를 시행하는 방법이다.

본 논문에서는 임계치를 이용하는 방법을 사용하여 계층적 군집화를 수행하였다. 임계치는 다음과 같은 방법으로 계산한다.

$$T = \frac{1}{\langle \sqrt{N} \rangle} \times \alpha \quad (5)$$

여기서, N은 초기 데이터의 개수이고, α는 보정치이다. 임계치는 초기 데이터를 2차원 평면상에 사상할 때 이상적으로 분포할 때의 거리를 의미한다.



<그림 3> 최적화된 군집 지도의 예

<그림 3>은 <그림 2>의 가까이 있는 군집들을 제안하는 최적화 방법으로 결합한 결과이다. <그림3>에서 보는 바와 같이 0-4-6 군집이 결합되어 하나의 군집을 형성하고 있다.

3.4 추천 목록 구성

최적화된 이웃선정 그룹을 기반으로 추천대상이 되는 고객이 평가하지 않은 아이템의 선호도를 예측하여 추천 목록을 구성한다. 근접 이웃 구성시 분류 기법(Classification)중의 하나인 k-NN(k-Nearest Neightbors)을 적용하였다. 선호도를 예측하고자 하는 고객을 기준으로 해당 고객과의 거리가 가장 가까운 이웃들을 탐색한다. 이때 추천하고자 하는 아이템을 평가한 고객들을 대상으로 유사도가 가장 높은 상위 k명의 고객들을 최근접 이웃으로 구성한다. 유사도를 계산할 때는 최적화된

군집 지도상에서 계산하기 때문에 유클리디언 거리를 사용한다. 이렇게 형성된 이웃을 기반으로 선호도 $R_{A,i}$는 고객 A에 대하여 아이템 i에 대한 선호도로 고객 A와 유사한 그룹내의 선호도들을 가중 평균하여 구하며 이는 식 (6)으로 예측한다.

$$R_{A,i} = \overline{R_A} + \frac{\sum_{j=1}^{k} \omega(A,j)(R_{j,i} - \overline{R_j})}{\sum_{j=1}^{k} |\omega(A,J)|} \quad (6)$$

$\overline{R_j}$는 고객 A의 유사그룹소속 고객 j의 이용 가능한 선호도들의 평균값이다. 추천하고자 하는 고객에 가장 가까운 이웃을 N개를 찾아서 해당고객들이 추천하고자 하는 상품에 대한 평균선호도로 추천 목록을 구성한다.

Ⅳ. 실험 및 결과

4.1 실험 환경 및 데이터

실험환경은 Intel Core2Duo E6550, 2GB RAM 시스템 상에서 수행하였다. 구현은 C#언어로 하였으며 닷넷 프레임워크 2.0 기반하에 수행 하였다. 실험데이터는 GroupLens에서 공개하는 100K MovieLens 자료를 이용하여 분석하였다. 100K MovieLens 데이터는 943명의 사용자가 1682편의 영화에 대해 자신의 선호도를 1~5의 점수로 평가된 수치이다. 각 사용자가 최소 20편의 영화에 대해 평가하여 총 100,000개의 데이터로 구성되어 있다 [22-23]. 사용자의 인구 통계 정보로는 나이, 성별, 직업, 우편번호등이 있다. 영화는 19개의 장르로 구분되어 있고, 중복장르 선택을 허용하였다. 본 연구에서는 사용자의 영화에 대한 평가 결과를 이용하여 실험하였다.

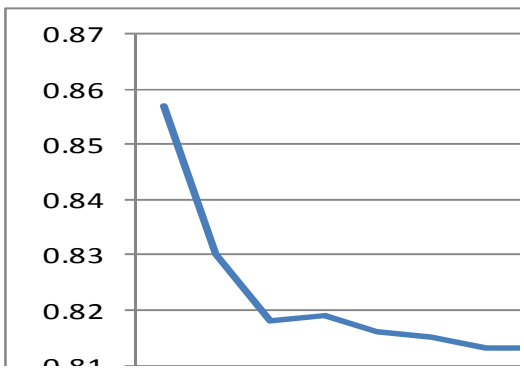
실험을 위하여 학습데이터와 평가데이터로 나누어 u1, u2, u3, u4, u5의 5개 데이터 집합을 생성하였다. 데이터를 랜덤하게 추출하여 80%는 학습데이터로 20%는 평가 데이터로 사용하였다. 각각 공통 원소를 갖지 않는 집합으로서 5-fold 교차검증(cross validation)이 가능하도록 구성하였다. 계층적 군집화의 종료조건인 임계치를 계산할 때 필요한 α값은 여러번 실험을 통하여 우수한 성능을 보인 0.8로 적용하였다.

4.2 실험결과 및 분석

본 논문에서는 추천시스템의 성능을 평가하는 방법은 여러지표 중에서 예측값과 실제값의 차이를 표시하는 평균절대오차(Mean Absolute Error: MAE)를 적용하였다. 평균절대오차는는 실제 선호도 값과 예측된 선호도 값의 차이로 정의되며, 이는 예측의 정확성을 판단하는데 사용된다.

$$|E| = \frac{\sum_{i=0}^{N} |\epsilon_i|}{N} \quad (7)$$

여기서 N은 예측회수이고, ε_i는 실제값과 예측값의 차이이고, i는 각 예측단계를 나타낸다. 평균절대오차는 예측된 선호도들이 실제 고객의 선호도들과 평균적으로 얼마나 유사한지를 나타내는 지표이다.



<그림 4> k-NN의 k 변화에 따른 MAE 값의 변화

K-Means의 초기 K 값은 고객의 수를 기준으로 결정

하였다. 고객의 수가 943명이므로 Log 함수를 취하여 7을 초기 군집수로 결정하였다. k-NN의 최적 k 값을 결정하여 위하여 실험을 진행하였다. k를 5부터 시작하여 5씩 증가시키면서 100까지 실험을 진행하였고, 그 결과는 <그림 4>와 같다. k의 수가 60일 때 가장 좋은 결과를 보였고, 따라서 향후 실험은 모두 k의 수를 60으로 결정하였다.

제안하는 이웃군집 최적화 방법의 성능 비교를 위하여 다음과 같은 종류에 대한 성능을 비교하였다. k-NN은 전체 데이터에 대해 k-NN을 통해 이웃을 선정해서 근접 이웃의 결과를 추천하는 방법이다. 전체 데이터에 대해 적용하기 때문에 코사인 상관 계수로 이웃 간의 거리를 계산하였다.

K-Means + k-NN 은 전체 데이터를 먼저 K-Means를 이용하여 군집화 하여 이웃 그룸을 형성 한 뒤 입력값이 속한 군집에 대해서만 k-NN을 적용한 방법이다. K-Means는 전체 데이터에 대해 적용하기 때문에 코사인 상관 계수를 이용하여 군집을 형성하였고, k-NN을 이용한 이웃 결정시에도 코사인 상관 계수를 이용하여 계산하였다.

K-Means + MDS + k-NN 방법은 K-Means를 이용하여 군집화한 결과를 다차원 척도 법으로 2차원으로 사상한 후 k-NN을 적용한 방법이다. K-Means를 계산할 때에는 코사인 상관 계수를 이용하였다. MDS를 적용하면 2차원에 데이터가 존재하기 때문에 k-NN을 계산할 때에는 유클리디언 거리를 사용하여 계산하였다.

K-Means + MDS + Optimization + k-NN -방법은 KMeans를 이용하여 군집화한 결과를 다차원척도법을 적용하여 2차원으로 사상한 후 최적한 적용 후 k-NN을 적용시킨 방법이다. K-Means를 계산할 대에는 코사인 상관 계수를 이용하였다. Optimization과 k-NN은 모두 MDS에 의한 2차원 평명 상에서 이루어지기 때문에 유클리디안 거리를 사용하여 계산하였다.

5가지 실험 집단에 대해 4가지 실험 방법을 모두 10번씩 실험하여 평균치를 계산하였다. 평균절대오차와 수행시간을 분석한 결과는 다음과 같다.

<표 5>의 수행 성능을 보면, 4가지 방법 모두 유사한 경향을 보이지만, 먼저 군집화를 통하여 유사그룹을 형성한 방법이 전체 데이터에 대해 검색하는 것보다 약 1% 정도 우수한 성능을 보였다. 또한 제안하는 방법은 다차원 척도법에 의해 2차원으로 사상을 하지만 원래의 정보를 유지하기 때문에 큰 성능저하는 발생되지 않았고, 일부의 경우에는 더 좋은 결과를 보이기도 하였다.

<표 5> 평균절대오차 비교

	k-NN	K-Means + k-NN	K-Means + MDS + kNN	K-Means + MDS + Opt. + kNN
u1	0.817	0.806	0.808	0.806
u2	0.817	0.802	0.806	0.803
u3	0.807	0.802	0.801	0.799
u4	0.817	0.806	0.809	0.807
u5	0.822	0.809	0.813	0.81

수행시간은 <표 6>과 같다. K-NN은 전체 다차원 데이터에 대해 코사인 상관 계수 계산이 이루어지기 때문에 수행시간이 가장 많이 소요된다. 하지만, 먼저 군집화를 수행하면 이웃을 선정할 때 군집에 속한 데이터로 범위를 줄여서 탐색을 하기 때문에 k-NN에 비해서 약 30%정도의 시간을 단축시킬 수 있었다.

<표 6> 수행시간비교

	k-NN	K-Means + k-NN	K-Means + MDS + kNN	K-Means + MDS + Opt. + kNN
u1	6분34초	4분41초	14초	15초
u2	5분54초	4분32초	12초	13초
u3	5분50초	4분53초	11초	12초
u4	6분19초	4분15초	9초	10초
u5	5분44초	3분27초	13초	13초

군집화 결과에 다차원 척도법을 적용하면 다차원 데이터가 2차원으로 사상되어 최근접 이웃을 찾는 계산식이 코사인 상관관계에서 유클리디안 거리로 변경되기 때문에 탐색 시간이 큰 폭으로 감소한다. 실험 결과 K-Means +MDS+k-NN가 K-Means + k-NN 보다 약 95% 정도 시간이 감소한 것을 확인할 수 있다. 실험을 통하여 제안하는 방법은 성능의 저하는 없으면서 수행시간을 크게 단축시키는 것을 확인할 수 있었다.

Ⅴ. 결론

성공적인 마케팅은 고객이 구매할 수 있도록 유도하는 능력뿐만 아니라 오랜 기간 동안 고객과의 관계를 유지하는 능력도 필요하다. 이러한 상황에서 고객 개인화(personalization)는 마케팅에 있어서 매우 중요한 문제이다. 고객 추천 시스템은 고객 개인화를 달성하는데 있어서 매우 중요한 역할을 담당하고 있다. 추천 시스템의 핵심이 되는 협력적 필터링에 대한 연구는 보다 나은 마케팅 수행을 위한 중요한 요소가 될 것이다.

본 논문에서는 협력적 필터링 기법을 구현하는데 있어서 다차원 축소에 기반한 근접 이웃을 구성하는 방법을 제안하였다. 제안하는 방법은 군집화를 통하여 추천 시스템의 성능을 향상 시킬 수 있었고, 다차원 척도에 기반한 차원축소 방법을 통하여 추천시 소용되는 시간을 감소시킬 수 있었다. 따라서, 온라인 개인 추천 시스템에 적당한 방법이라고 할 수 있다.

향후 연구과제는 다음과 같다. 우선, 현재는 단편적인 아이템에 대해서만 사용이 되고 있다. 하지만, 인터넷 홈쇼핑 등 다양한 아이템이 복합되어 있는 곳에서는 어떻게 문제를 해결할 것인가에 대한 연구가 필요하다.

두번째로 고객과 관련된 데이터는 상품에 대한 선호도 또는 구매 이력 데이터만 포함하지는 않는다. 인구 통계 데이터(나이, 성별, 직업, 우편번호 등), 웹 로그 등의 데이터도 고객을 개인화 하는데 있어서 중요한 정보이다. 따라서 이러한 정보들과 연동시키기 위한 연구가 필요하다.

참고문헌

[1] 이재식 · 박석두, “장르별 협업필터링을 이용한 영화 추천 시스템의 성능 향상,” 한국지능정보시스템학회논문지, Vol. 13, No. 4, 2007, pp. 65~78.

[2] 권준희 · 김성림 “유비쿼터스 환경에서 상황 데이터 기반 모바일 콘텐츠 서비스를 위한 추천 기법,” 디지털산업정보학회 논문지, Vol. 6, No. 2, 2010, pp. 1-9.

[3] 조동주 · 정경용 · 임기욱 · 이정현, “개인화 추천 시스템에서 FP-Tree를 이용한 연관 군집 방법," 한국콘텐츠학회논문지, Vol. 7, No. 10, 2007, pp. 19-26.

[4] 부종수 · 홍종규 · 박원익 · 김룡 · 김영국, “추천시스템의 성능 향상을 위한 시간스키마 적용 2단계 클러스터링 기법,” 전자거래학회지, Vol. 10, No. 2, 2005, pp. 109-132.

[5] 김성림 · 권준희, “상황인식 정보 검색 기법을 이용한 하이브리드 협업 필터링 기법,” 디지털산업정보학회 논문지, Vol. 6, No. 1, 2010, pp. 143-149.

[6] 이석준, “근접 이웃 선정 협력적 필터링 추천 시스템에서 이웃 선정 방법에 관한 연구,” 한국데이터정보과학회지, Vol. 20, No, 5, 2009, pp. 809-818.

[7] 윤수진, 윤희병, “개인화 추천 시스템의 성능향상 적용 알고리즘 분석,” 한국퍼지 지능시스템 학회 춘계 학술 발표 논문집, Vol. 15, No. 1, 2005, pp. 181-183.

[8] P. Resnick, N. Iacovou, M. Suchak, P. Bertorm, J. Riedl, “GroupLens: An open architecture for collaborative filtering of netnews,” Proceedings of ACM Conference on Computer Supported Cooperative Work, 1994, pp. 175-186.

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

[10] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, John Riedl"An algorithmic framework for performing collaborative filtering," In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999, pp. 230-237.

[11] John S. Breese, David Heckerman, Carl Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," Proceedins of the 14th Conference of Uncertainty in Artificial Intelligence, 1998.

[12] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl "Application of dimensionality reduction in recommender system-a case study," ACM WebKDD Workshop, 2000.

[13] B. Sarwar, "Sparsity, Scalability, and Distribution in Recommender Systems," Ph. D. Diss., Dept. of Computer and Information Sciences, Univ. of Minesota, 2001.

[14] G. Linden, B. Smith, J. York, "Amazon. com Recommendations: Item-to-item Collaborative Filtering," IEEE Internet Computing, Vol. 7, No. 3, 2003, pp. 76~80.

[15] Gui-Rong Xue, Chenxi Lin, Qiang Yang, WenSi Xi, Hua-Jun Zeng, Yong Yu, Zheng Che, "Scalable Collaborative Filtering Using Cluster-based Smoothing," Proceedings of the 2005 ACM SIGIR Conference, 2005, pp. 114-121.

[16] P. Li, S. Yamada, "A Movie Recommender System Based on Inductive Learning," IEEE Conf. on Cybernetics and Intelligent Systems, 2004, pp. 318-323.

[17] Gui-Rong Xue, Chenxi Lin, Qiang Yang, Wensi Xi, Hua-Jun Zeng, Yong Yu, and Zheng Chen August, "Scalable Collaborative Filtering using Cluster based Smoothing," Proceedings of the 2005 ACM SIGIR Conference, 2005, pp. 114-121.

[18] E. Gose, R. Johnsonbugh and S. Jost, "Pattern Recognition and Image Analysis," Prentice Hall, 1996.

[19] J. He, A. H. Tan, C. L. Tan, and S. Y. Sung, "On quantitative evaluation of clustering systems," In Weili We, Hui Xiong, and Shashi Shekhar, editors, Information Retrieval and Clustering. Kluwer Academic Publishers, 2003.

[20] C. G. Li, J. Guo, G. Chen, X. F. Nie and Z. Yang, "A Version of ISOMAP with Explicit Mapping," In Proc. of Fifth International Conference on Machine Learning and Cybernetics, 2006, pp. 3201-3206.

[21] J. B. Tenenbaum, V. de Silva and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," SCIENCE, Vol. 290, 2000, pp. 2319-2323.

[22] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, John Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News," Communications of the ACM, Vol. 40, No. 3, 1997, pp. 77-87.

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

■ 저자소개 ■

이 현 진
Lee, Hyun Jin

2003년 3월~현재
한국사이버대학교
컴퓨터정보통신학과 부교수
2002년 8월 연세대학교 컴퓨터과학과
(공학박사)
1998년 8월 연세대학교 컴퓨터과학과
(공학석사)
1996년 8월 순천향대학교 전산학과 (공학사)

관심분야 : 기계학습, 데이터마이닝, 이러닝
E-mail : hjlee@mail.kcu.ac

지 태 창
Jee, Tae Chang

2004년 8월~현재
연세대학교 컴퓨터과학과
(박사과정)
1999년 3월~현재
LG CNS 과장
1999년 2월 연세대학교 컴퓨터과학과
(공학석사)
1997년 2월 연세대학교 컴퓨터과학과(공학사)
관심분야 : 군집화, 기계학습, 데이터마이닝
E-mail : garura@csai.yonsei.ac.kr

논문접수일 : 2010년 7월 23일
수 정 일 : 2010년 8월 15일(1차), 8월 28일(2차)
게재확정일 : 2010년 9월 5일