추천시스템을 위한 k-means 기법과 베이시안 네트워크를 이용한 가중치 선호도 군집 방법

박화범* · 조영성** · 고형화***

Clustering Method of Weighted Preference Using K-means Algorithm and Bayesian Network for Recommender System

Wha-Beum Park* · Young-Sung Cho** · Hyung-Hwa Ko***

Abstract

Real time accessiblity and agility in Ubiquitous-commerce is required under ubiquitous computing environment. The Research has been actively processed in e-commerce so as to improve the accuracy of recommendation. Existing Collaborative filtering (CF) can not reflect contents of the items and has the problem of the process of selection in the neighborhood user group and the problems of sparsity and scalability as well. Although a system has been practically used to improve these defects, it still does not reflect attributes of the item. In this paper, to solve this problem, We can use a implicit method which is used by customer's data and purchase history data. We propose a new clustering method of weighted preference for customer using k-means clustering and Bayesian network in order to improve the accuracy of recommendation. To verify improved performance of the proposed system, we make experiments with dataset collected in a cosmetic internet shopping mall.

Keywords : RFM Method, K-means Algorithm, Bayesian Network, Collaborative Filtering, Recommender System

논문접수일 : 2013년 08월 27일 논문수정일 : 2013년 09월 26일 논문게재확정일 : 2013년 09월 27일

* 남서울대학교 교양학부, e-mail : Koreasnooker@nsu.ac.kr
** 동양미래대학교, e-mail : youngscho@empas.com
*** 광운대학교 전자통신공학과, e-mail : hhkoh@kw.ac.kr

1. 서 론

유비쿼터스(Ubiquitous) 컴퓨팅 환경 하에서 스마트폰과 아이패드 같은 지능형 단말기를 이용한 모바일 인터넷이 생활의 일부가 되면서 정보의 양도 급속도로 팽창하여 대량의 데이터(Big data) 속에서 정보를 찾아내는 기술이 부각되고 있다. 추천시스템은 고객을 대신하여 적합한 아이템을 빠른 시간 내에 추천하고, 그 추천된 내용이 또한 정확하다면 고객은 만족감을 얻을 수 있다.

고객 프로파일을 생성하는 명시적인 방법인 기존의 협력 필터링은 희박성 문제와 아이템의 속성을 반영하지 못하는 문제 그리고 확장성의 문제 등 여전히 많은 문제가 존재한다. 또한 상품에 대한 고객들의 구매 패턴은 시간에 따라 다르게 설정되어야 하며, 상품에 대한 가격 등과 같이 항목마다 다른 가중치를 설정하여야 하는 경우가 일반적이기 때문에 실제적으로 선호도계산에 있어서 모든 항목에 대하여 같은 가중치를 부여하는 선호도 계산법은 매우 현실적이지 못한 방법이라 생각된다. 또한 시간 변화에 따른 고객 선호도의 변화는 유연하게 선호도 계산이 이루어짐으로써 적절한 정보가 이루어져야 한다.

고객에 맞는 선호도 기반의 군집 방법을 이용한 개인화 추천 방법에 대한 연구[Cho et al., 2002; Cho et al., 2011; Cho et al., 2012; Cho et al., 2012]가 활발히 진행되고 있다. 본 논문은 이러한 관련 연구로써 지능형 단말기를 사용하는 고객이 고객 프로파일을 생성하지 않고 고객정보와 구매이력 데이터를 이용한 묵시적인 방법으로 고객 선호도 계산을 위해 고객점수 가중치 기반의 k-means 기법과 베이시안 네트워크를 이용한 가중치 선호도를 이용하여 새로운 군집방법을 제안한다.

본 논문의 구성은 다음과 같다. 제 2장은 관련 연구를 다루었으며, 제 3장에서는 제안 추천 시스템 설명하였다. 제 4장에서는 실험 및 성능 평가를 실행하였고, 제 5장에서는 본 논문의 결론과 향후 연구에 대하여 기술하였다.

2. 관련 연구

2.1 RFM(Recency Frequency Monetary)

RFM은 최근성, 빈도성, 총구매액으로 구성된다. RFM은 구매 가능성이 높을 고객을 선정하기 위한 데이터 분석 방법이다. 세 가지 요소를 기준으로 고객 각각에 대해 점수를 부여하고 세 가지 기준의 가중치를 주어 RFM 점수를 계산하게 된다. 이 RFM 점수를 고객 가치를 평가하는 지표로 삼는 방식이 RFM에 의한 고객점수 부여 방법이다 [Cho et al., 2012].

다음은 RFM 점수 산출식의 예를 나타낸 것이다. RFM 점수의 가중치(A, B, C)는 경영 상태나 경영 전략에 따라 변경이 가능하다.

$$\text{RFM점수} = (A \times R + B \times F + C \times M) \quad (1)$$

RFM 점수의 합계는 최고점수는 100점, 최하점수는 0점이다. RFM 점수를 위해서 사용되는 R, F, M 요소는 예측력이 강한 변수이다. RFM은 구매 가능성이 높은 고객을 선정하기 위한 데이터 분석 방법이다.

본 논문에서는 고객의 구매 가능성 세분화를 위해 RFM 점수기반 고객과 취급되는 제품을 등급화하여 분석한다. RFM이 적용된 시스템의 이점은 여러 가지 정보를 손쉽게 알 수 있다는 것이다. 예를 들면, 고객정보의 고객 분류코드와 RFM 점수를 비교하여 현재 아이템의 인기도나, 선호도, 관심 및 추천 아이템 등을 쉽게 파악할 수 있다 [Cho et al., 2002].

2.2 협력 필터링

협력 필터링은 고객들의 선호도 정보를 바탕으

로 유사한 성향을 가지는 다른 고객에 의해 높은 선호를 보인 구매 아이템 등을 고객에게 추천하는 방식이다. 고객의 구매데이터를 기반으로 고객간의 유사도(Similarity)를 계산하고 그로부터 구매하지 않은 아이템에 대한 선호도를 예측하는 시스템이다.

협력 필터링은 아이템에 대한 다른 고객들의 선호도를 기반으로 하기 때문에 협력적이라는 용어를 사용하게 된다. 협력 필터링 시스템은 시스템이 묵시적인 자료를 사용하는지 명시적인 자료를 사용하는 것에 따라 구분을 한다.

또한 추천 정보를 제시하는 방법으로 협력 필터링, 인구통계학적 필터링, 규칙 기반 필터링, 내용 기반 필터링 등이 사용되고 있다. 기존의 협력 필터링의 희박성과 확장성의 문제점을 개선하려는 연구가 진행되어 왔으며 실제로 많은 성과가 있었다. 그러나 명시적인 자료를 기반으로 하기 때문에 여전히 희박성이 존재하고 아이템의 속성 대한 선호도를 반영되지 않는 문제점이 남아 있다[Kim et al., 2007].

본 논문에서는 이러한 문제점 해결 방안으로 고객에게 번거로운 질의 응답 과정이 없이 묵시적인 방법으로 고객정보와 아이템정보, 구매이력정보를 이용한다.

2.3 k-means 기법

클러스터링의 대상이 되는 객체(Object)들은 각 객체의 특성을 나타내는 속성을 가지고 있다. 객체들은 클러스터링을 통해서 특정 군집에 속하게 되며, 각 군집은 소속 객체들의 속성 정보를 소유한다.

객체에 대한 클러스터링 결과를 분석하면 각 군집에 분포된 객체들의 분포도에 대한 정보를 얻을 수 있다. 가장 많이 사용되는 클러스터링 기법으로 k-means 기법이 있다. 가장 가까운 중심점을 갖는 군집에 각 항목을 할당하는 과정을 반복하여 k개의 군집으로 항목들을 나누는 것이다.

거리 기반 클러스터링 방법으로 고객의 선호도를 다차원 공간상의 점으로 표시하고, 거리를 계산함으로써 전체 고객들의 집합을 k개의 군집으로 나눈다. 고객 a와 k 사이의 거리는 식(2)와 같이 계산하고, 식에서 a_i는 고객 a의 속성(차원) i에 대한 선호도 값을 의미한다.

$$d_{a,k} = \sqrt{\sum_i (a_i - k_i)^2} \qquad (2)$$

본 논문에서는 고객정보변수로 구성된 인구통계학적 변수(나이, 성별, 직업, 피부타입 등)와 고객점수가 적용된 고객 데이터베이스와 구매이력 데이터베이스를 이용하여 클러스터링(Clustering)을 위해 k-means 기법을 적용한다.

적용된 k-means 기법은 계산 속도가 빠르고 대량의 자료에서 군집을 발견하는데 상당히 효과적인 것으로 알려져 있다[Lee et al., 2004].

2.4 베이시안 네트워크(Bayesian Network)

베이시안 네트워크란 사전에 일어난 일을 바탕으로 사후의 확률을 추론하는 방법으로, 고객이 구매했던 이력을 분석하여 다음에 무엇을 구매할지 예측하는 방법이다.

베이시안 네트워크는 의사결정 트리(Decision tree)나 신경망(Neural Network)와 비교해 볼 때 대용량 데이터베이스에서 높은 정확성과 속도를 나타내고 주어진 클래스의 속성값이 다른 속성에 영향을 주지 않는다. 베이시안 네트워크는 $\{x_1, x_2, \cdots, x_n\}$라는 n차원의 특징벡터 X가 존재하며, $\{C_1, C_2, \cdots, C_n\}$라는 m개의 클래스가 존재한다고 가정한다. 임의의 데이터 X가 가장 높은 사후확률을 가지는 클래스에 속할 것이

라는 예측은 베이시안 네트워크를 사용하여 식 (3)과 같이 계산할 수 있다[Kim et al., 2007].

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)},\ 1 \le i \le m \quad (3)$$

$P(X)$는 모든 클래스에 대해 일정한 값을 가지므로, 오직 $P(X|C_i)P(C_i)$만을 최대화하도록 고려한다. 만약 클래스의 사전확률을 알 수 없다면 $P(X|C_i)$만을 고려할 수 있다. $P(X|C_i)$는 베이시안 네트워크의 독립가정에 의해 식 (4)와 같이 계산되며, 그 결과 X는 가장 큰 사후확률을 가지는 클래스로 분류할 수 있다.

$$\begin{aligned} P(X|C_i) &= P(x_1, x_2, ..., x_n, C_i)\, P(C_i) \\ &= P(x_1|C_i)\, P(x_2|C_i) \ldots P(x_n|C_i)\, P(C_i) \\ &= P(C_i) \prod_{k=1}^{n} P(x_k|C_i) \end{aligned} \quad (4)$$

고객이 구매했던 이력을 분석하여 선호도 테이블을 생성할 수 있다. 여기서 고객은 사이트에 접속한 로그인 사용자를 의미한다. 예를 들어 고객별 아이템에 대한 선호도에서 고객별 아이템 구매이력 건수 테이블이 있다고 가정하자.

고객{u_1, u_2, u_3, u_4}가 아이템 카테고리 {C_1, C_2, C_3, C_4}에 대한 구매 건수가 {1, 1, 2, 2}일 경우 고객(u_3)의 총 구매 건수는 6건일 때 고객(u_3)가 1차 사전 데이터를 기준으로 제공된 추천정보를 통해 고객(u_3)이 아이템카테고리 C_4에 대한 구매를 결정하여 구매 건수가 4회 증가하여 총 6회의 구매 건수를 갖는다고 하면, 사전 데이터는 2, 사후 데이터는 6이 되고, 전체 구매 건수는 10이 된다.

베이시안 네트워크는 고객의 구매 행위 결정으로 변화된 사후 데이터를 반영하여 아이템의 선호도 확률을 계산한다. 아이템카테고리(X = C_4)에 대해 아이템 구매 행위가 결정되었을 때 베이시안 네트워크에 의해 변경된 아이템카테고리 C_3에 대한 선호도의 사후 확률은 식 (6)과 같이 계산된다. 각 고객별로 특정 아이템에 대한 구매결정이 이루어졌을 때 구매 건수가 반영된 아이템카테고리에 대한 선호도의 사후 선호도로 변경된다면 보다 정확한 추천 아이템의 선호도 확률을 얻을 수 있다[Choi et al., 2002].

$$\begin{aligned} \mathrm{P}(C_{j,i}|X_i) &= \frac{P(C_{j,i}, X_i)}{P(X_i)} \\ &= \frac{P(C_{j,i})\, P(X_i|C_{j,i})}{P(X_i)}, \end{aligned} \quad (5)$$

$$\text{where } P(X_i) = \sum_{j=1}^{k} P(C_{j,i})\, P(X_i|C_{j,i})$$

식 (5)를 이용해서 고객(u_3)이 아이템카테고리 C_4에 대한 구매결정에 따른 구매 건수가 4회 증가하여 총 6회의 구매 건수를 갖는다고 하면, 사전 데이터는 2, 사후 데이터는 6이 되고, 전체 구매 건수는 10이 된다. 베이시안 네트워크는 고객의 구매 행위 결정에 의해 변화된 사후 데이터를 반영하여 아이템의 선호도 확률을 계산한다.

고객(u_3)이 1차 사전 데이터를 기준으로 제공된 추천 아이템을 통해 아이템카테고리($X = C_4$)에 대해 아이템 구매 행위가 결정되었을 때 베이시안 네트워크에 의해 변경된 아이템카테고리(C_3)에 대한 선호도의 사후확률은 식 (6)과 같이 계산된다.

$$P(C_3|X = C_4) = \frac{P(C_3)\, P(X = C_4|P(C_3))}{\sum_{j=1}^{4} P(C_j)\, P(X = C_4|C_j)} \quad (6)$$

$$\frac{0.33 \times \frac{2}{10}}{0.17 \times \frac{1}{10} + 0.17 \times \frac{1}{10} + 0.33 \times \frac{2}{10} + 0.33 \times \frac{6}{10}} \cong 0.23$$

식 (6)과 같은 방법으로 구매결정에 대한 전체 고객 선호도 확률을 구할 수 있다. 고객(u_3)에 대해 특정 아이템의 구매결정에 의한 사후 선호도 확률값과 사전 선호도 확률값을 비교하면 다음 <표 1>과 같다.

〈표 1〉 선호도 테이블 비교표

item / u_3	C_1	C_2	C_3	C_4
사전	$\frac{1}{6}$= 0.17	$\frac{1}{6}$= 0.17	$\frac{2}{6}$= 0.33	$\frac{2}{6}$= 0.33
사후	0.07	0.07	0.23	0.65

위의 <표 1>에서 고객(u_3)이 정한 아이템의 구매결정을 했을 때 특정 아이템의 선호도 확률값은 사전 확률보다 상향조정되었고 전체 데이터에 의해 현재의 선호도 패턴이 반영되었음을 알 수 있다.

또한 고객이 한 군집의 선호도 확률도 같은 방법으로 계산 가능하다. 위의 예제처럼 베이시안 네트워크 기법을 이용하면 아이템 선호도에 대한 사전확률 뿐만 아니라 사후확률 계산도 가능하다.

3. K-means 기법과 베이시안 네트워크 기반 가중치 선호도 군집 방법을 이용한 추천시스템

이 장에서는 묵시적인 방법으로 고객정보 변수를 이용하여 발췌된 구매 이력 데이터를 바탕으로 고객점수 가중치를 적용한 가중치 선호도를 이용한 새로운 군집방법을 제안한다. 이를 위해 학습 에이전트에서 전처리 방식으로 가중치 선호도에 대한 사전확률이 계산되며 고객에게 제시된 1차 추천을 바탕으로 구매를 결정한 시점의 사후 확률 계산의 2차 추천에 따른 보다 정확한 추천을 통한 구매가 유도될 것이다. 이를 위해 고객 프로파일을 생성하지 않고 고객정보와 구매이력 데이터를 이용하고 고객 선호도 계산을 위해 고객정보와 구매 아이템에 대한 정의와 함께, 선호도 함수를 정의한다.

또한 고객점수 가중치를 이용한 가중치 선호도 적용 절차 알고리즘과 가중치 선호도를 이용한 k-means 기법을 위한 절차 알고리즘을 기술한다.

[정의 1] 고객정보 변수가 k개의 속성을 가지다면, 고객정보 변수 집합 P = {P_1, P_2,…, P_k}로 표현한다.

[정의 2] 고객정보 변수가 속성을 n개씩 클러스터링 할 경우 P_j를 n개씩 클러스터링 할 경우 P_{j1}={$q_1, q_2, \cdots, q_n$} P_{j2}={q_{n+1}, q_{n+2}, q_{2n}}$\cdots P_{jm}$={q_{m+1}, q_{m+2}, q_{2m}}로 표현한다.

[정의 3] 추천시스템에서 사용하는 아이템카테고리는 트리를 갖는다.

[정리 3] 아이템 i_i(i∈m)가 소속된 카테고리를 $C_k(i_i)$=C_j(j∈r)라고 하면, i_i의 상위 카테고리는 TC(i_i) = C_k(k∈r)이다. 이 때 C_k를 C_j의 부모 카테고리라 한다.

[증명] [정의 3]에 의해서 아이템카테고리는 트리구조를 갖기 때문에 함수 TC(i_i)의 값은 C_k가 된다. 따라서 C_j의 부모 노드는 C_k이다. [정의 3]과 [정리 3]으로부터 다음과 같은 결과를 도출할 수 있다.

첫째, 최상위(루트) 카테고리 C_0로부터 하위 카테고리로 이어지는 경로상의 모든 카테고리들은 C_0의 자식 카테고리가 된다.

둘째, 경로상의 자식 카테고리 C_2는 자식 카테고리가 없는 카테고리를 리프(leaf) 카테고리라고 하며 각각의 아이템은 리프 노드로 브랜드아이템이 된다.

셋째, 트리는 균형을 유지할 필요가 없으며 트리의 계층은 대분류, 중분류, 소분류, 리프 노드로

구성된다.

[정의 4] 상위 카테고리의 선호도는 고객의 하위 카테고리의 선호도의 평균값으로 산출한다. 아이템카테고리에서 하위 카테고리의 고객 선호도를 상위 카테고리의 선호도에 반영하지만, 상위 카테고리의 고객 선호도는 하위 카테고리로 반영하지 않는다.

[정의 5] 선호도를 나타내는 함수 Pre_icd(id, code, date)는 고객(id), 아이템카테고리에 속한 아이템 코드(code)를 갖는 아이템, 구매 시에 대한 선호도를 나타낸다. 즉 임의의 구매일자(date)는 구매일자와 time stamp를 포함하며 고객이 아이템카테고리 속한 아이템(브랜드아이템)에 대한 선호도를 나타낸다. 예를 들면 {<u_1, i_1, 5>, <u_1, i_2, 10>, <u_2, i_2, 5>, <u_2, i_3, 10>, <u_2, i_4, 5>, <u_2, i_5, 5>}로 고객은 u_i, 접근한 아이템은 i_i이고 숫자는 아이템에 대한 구매 건수를 나타낸다.

〈표 2〉 고객점수 가중치를 이용한 가중치 선호도 적용 절차 알고리즘

Step 1 : 로그인 고객과 성향이 같은 구매이력 데이터를 추출한다.
Step 2 : 학습에이전트에서 전처리과정으로 고객에 대한 RFM 점수와 아이템점수를 계산 및 관리한다.
Step 3 : 시스템은 로그인 고객과 성향이 같은 구매이력 데이터를 기준으로 고객점수 구간별 구매 건수를 집계한다.
Step 4 : 시스템은 집계된 구매 건수를 바탕으로 구매건수 점유율을 산정하여 고객 수량 가중치 비율을 구한다.
Step 5 : 시스템은 발췌된 구매데이터를 기준으로 아이템 선호도를 계산할 때 사전 확률계산으로 구매 건수에 가중치 비율을 적용한다.
Step 6 : 가중치가 적용된 사전 선호도 확률이 높은 아이템 카테고리 중에서 구매를 위한 선택이 결정되면 최종 구매 결정을 위한 사후 선호도 확률을 계산한다.
Step 7 : 구매하려는 아이템의 카테고리에 속한 단말 브랜드아이템의 선호도가 높은 순으로 추천 아이템들을 생성한다.
Step 8 : 추천 아이템의 선호도를 스캔하여 선호도가 높은 아이템들을 Top-N의 추천 아이템 목록으로 생성한다.
Step 9 : 추천 시 추천된 아이템을 로그인 고객 구매 이력정보와 체크하여 중복 추천되지 않도록 한다.

[정의 6] 가중치 선호도 : 선호도 P에 대한 가중치 선호도(Weighted Preference)는 다음과 같이 정의된다. 가중치 Weight(P)는 고객점수(Score) 등급별 가중치를 나타낸다.

로그인 고객 id가 속한 고객 등급 가중치는 고객의 점수 등급의 구매건수를 전체 구매건수로 나누어 계산한다. 즉 가중치 선호도는 구매건수에 고객 등급 가중치를 곱하여 계산한다.

3.1 고객점수기반 가중치 선호도 적용

본 논문에서는 로그인 고객(Login User)은 고객정보를 읽어 고객정보 변수(나이, 성별, 직업, 피부타입 등)와 고객점수를 이용하여 전 처리된 구매이력 데이터에서 로그인 고객과 성향이 같은 이웃 고객군집(Neighborhood Cluster of User)을 생성한다.

로그인 고객과 성향이 유사한 이웃 고객군집과 조인하여 구매이력 데이터를 추출한다. 추출된 구매이력 데이터를 기준으로 아이템 선호도를 계산할 때 사전 확률계산으로 구매 건수에 가중치 비율을 적용한다. 선호도 추천시스템의 학습에이전트에서 구매 패턴 예측을 위한 추출된 구매 데이터 바탕으로 시간에 따라 변화하는 고객의 선호도는 시간이 지나감에 따라 바뀔 수 있으므로, 시간에 따른 고객 선호도 변화가 유연하게 선호도 계산이 가능한 베이시안 네트워크 방법을 적용한다. 다음은 추천의 정확도를 높이기 위한 고객점수 가중치를 이용한 가중치 선호도 적용 절차 알고리즘이다.

3.2 k-means 기법을 이용한 이웃고객 생성 알고리즘

효과적인 추천을 위해서 인터넷 쇼핑몰에서 설

계된 고객정보인 고객정보 변수(나이, 성별, 직업, 피부타입 등) 및 고객점수가 적용된 고객 데이터베이스와 구매이력 데이터베이스를 이용한다.

로그인 고객과 성향이 같은 고객의 구매이력 데이터를 바탕으로 가중치 아이템 카테고리 선호도를 이용하여 군집화 한다. 다음은 가중치 선호도를 이용한 k-means 기법을 위한 절차 알고리즘을 나타낸 것이다.

〈표 3〉 가중치 선호도를 이용한 k-means 기법을 위한 절차 알고리즘

입력 : 고객-아이템카테고리-선호도(UCP), 아이템카테고리 테이블(CCT)
출력 : 특징 벡터(Feature Vector), 이웃 고객 구매데이터

begin
1. 로그인 고객의 특징 벡터(Feature Vector)고객과 성향이 같은 구매이력 데이터를 추출하기 위해서 고객의 특징 벡터(Feature Vector)를 이용한다. // 분류코드(나이, 성별, 직업, 피부타입) 및 고객점수를 이용한다.
2. for (CCT에서 모든 브랜드 아이템카테고리)
 2.1 집계합수를 이용하여 아이템카테고리 내의 평균 브랜드 아이템 선호도(Pref_UC(u,c)를 계산한다.
 //고객-아이템카테고리-선호도(UCP)내의 평균 브랜드 아이템의 선호도를 계산한다.
 endfor;
 2.2 for (CCT내의 모든 아이템카테고리)
 집계합수를 이용하여 평균 아이템카테고리 선호도를 계산한다.
 // 아이템카테고리내의 평균 아이템 카테고리 선호도를 계산한다.
 endfor;
 2.3 다음 공식을 이용하여 고객의 특징 벡터 V를 생성한다.
 아이템카테고리가 M개 일 때,

$$V = (V_1, V_2, V_3, \cdots, V_m)$$

$$V_i = \sum_k (\mathrm{Pref\_UC}(u_i, c_k \times Weight_i) / \sum_i (\sum_k (\mathrm{Pref\_UC}(u_i, c_k \times Weight_i)))$$

 // 가중치 선호도는 고객 등급 가중치를 건수에 곱하여 계산한다.
3. 집계합수를 이용하여 평균 가중치 아이템 선호도를 계산한다.
4. 선호도가 높은 아이템 카테고리 정보를 이용하여 이웃 고객을 생성한다.
 // k-means 클러스터링 알고리즘을 이용하여 고객 군집을 추출한다.
end.

3.3 추천시스템 절차 알고리즘

다음 <표 4>는 RFM기반 아이템 선호도 베이시안 네트워크를 이용한 추천시스템의 절차 알고리즘을 나타낸 것이다.

〈표 4〉 추천시스템 절차 알고리즘

Step 1 : 회원가입 시 고객정보 변수(나이, 성별, 직업, 피부타입 등)와 RFM의 고객점수를 통해 고객 분류코드 및 고객점수를 부여하여 고객정보를 생성 및 관리한다.
Step 2 : 고객정보에서 로그인 고객과 성향이 같은 군집(Cluster)을 탐색하여 선택한다.
Step 3 : Step 2에서 선택된 고객정보와 전처리 관리된 구매 이력데이터를 조인하여 해당되는 구매이력 데이터 군집을 발췌한다.
Step 4 : 해당 군집에 속한 구매데이터를 기반으로 아이템 카테고리의 선호도를 계산한다.
Step 5 : 사전 선호도 확률이 높은 아이템 카테고리 중에서 구매를 위한 선택 결정되면 최종 구매 결정을 위한 사후 선호도 확률을 계산한다.
Step 6 : 구매하려는 아이템의 카테고리에 속한 단말 브랜드아이템의 선호도가 높은 순으로 추천 아이템들을 생성한다.
Step 7 : 추천 아이템의 선호도를 스캔하여 선호도가 높은 아이템들을 Top-N의 추천 아이템 목록으로 생성한다.
Step 8 : 추천시 추천된 아이템을 로그인 고객 구매 이력정보와 체크하여 중복 추천되지 않도록 한다.

4. 실험 및 성능 평가

4.1 실험 환경

시스템 구현 및 실험 환경은 윈도우 운영체제 하에서 AMP(Apache/MySQL/PHP) 웹서버 환경 하에서 사용하였고 버전은 다음과 같다.

- OS : Windows XP SP3(64),
- Web Server : Apache 2.2.14/WAP 2.0
- Server-Side Script : JSP/PHP 5.2.12
- Client-SideScript : XHTML4.0/WML2.0/ HTML5.0/CSS3/JAVASCRIPT
- Database : MySQL 5.1.39
- J2SDK(1.7.0_11)
- MySQL JDBC
- jQuery Mobile
- jakarta-tomcat(5.0.28)

4.2 실험 데이터 구성

고객 구매 패턴 예측을 위한 RFM 점수기반 가중치 빈발 패턴 마이닝을 이용한 추천시스템은 윈도우 XP 환경에서 인터넷 화장품 쇼핑몰을 위한 데이터베이스가 구축되었다.

시스템에 대한 평가를 위한 실험데이터의 구성은 쇼핑몰을 이용해 본 경험이 있는 고객 319명의 고객정보와 그리고 현재 화장품을 전문적으로 판매하는 인터넷 화장품 쇼핑몰인 P사의 화장품 분류에서 사용하는 화장품 580개를 대상으로 그들의 추천 1,600건의 구매 데이터를 이용하여 화장품 아이템 분류체계에서 대.중.소 분류로 실제 쇼핑몰에서 사용되는 실 데이터를 중심으로 실험데이터 셋을 구성하였다[Cho et al., 2012].

4.3 분석 및 성능 평가

추천시스템의 전체적인 성능 평가는 예측 값과 실제 값의 차이를 표시하여 정확성 측면에서 성능을 평가하기 위한 MAE(Mean Absolute Error) 방식을 사용하였다. MAE는 예측의 정확성을 판단하는데 가장 많이 쓰이는 방법이다.

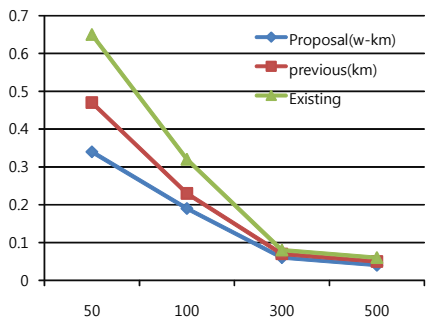
본 논문에서는 MAE에 대한 실험을 제안방법 적용 시스템과 이전 방법 적용시스템을 실험하였다. 우선 첫 번째, 실험으로 MAE에 의해 예측의 성능을 평가하였다. 추천시스템의 예측 값의 정확성을 평가하기 위해 MAE를 사용하였고 식 (7)과 같이 산출하였다[Herlocker et al., 1999].

$$MAE = \frac{\sum_{i=1}^{N} |\epsilon_i|}{N} \quad (7)$$

N은 총 예측 회수를 나타내고, ε_i는 예측 값과 실제 값의 오차를 나타내며 i는 각 예측 단계를 나타낸다. <표 5>는 식 (7)을 이용하여 예측 값의 정확성 평가를 수행한 결과이다.

〈표 5〉 제안 및 기존시스템의 MAE에 의한 성능평가

	P_count	Proposal (w-km)	previous (km)	Existing
MAE	50	0.34	0.47	0.65
	100	0.19	0.23	0.32
	300	0.06	0.07	0.08
	500	0.04	0.05	0.06



〈그림 1〉 제안 및 기존시스템의 MAE에 의한 성능평가

다음은 두 번째, 실험으로 정확도와 재현율, F-measure에 대한 실험이다. 성능은 social data에 기반한 화장품 추천에서의 추천의 유효성과 추천시스템의 전체적인 성능 평가 방향으로 진행하였다. 우선 초기 화장품 추천의 유효성을 실험에 참가한 고객들의 구매데이터와 제시되는 화장품

의 비교를 통해 이루어졌으며, 추천의 정확성을 평가하기 위하여 정보검색 분야에서 보편적으로 사용되는 평가 척도인 정확률(Precision)과 재현률(Recall)을 응용하여 사용하였다.

제안 추천시스템을 통해 추천된 추천 선호도가 높은 Top-N개를 추천하였고, 이 N의 추천 목록에 대하여 정확률, 재현율, F-measure를 평가하였다. 정확률은 추천의 정확률을 평가하기 위한 방법으로 추천 목록의 정확성이 어느 정도 정확한가를 평가하기 위한 방법으로서, 추천시스템이 고객에게 추천한 제품 갯수 중에서 실제로 고객이 구매한 제품의 비율이다.

재현율은 추천시스템의 추천 제품 중에서 실제로 고객이 구매한 제품의 비율이다. F-measure는 정확률과 재현율을 보완하기 위해서 결합한 평가 방법으로 시스템의 전체적인 성능을 평가할 수 있는 척도로 사용하였다.

$$\text{정확률} = \frac{\text{고객이 구매한 아이템 갯수}}{\text{추천아이템 갯수}} \quad (8)$$

$$\text{재현율} = \frac{\text{추천시스템의 추천아이템 중 고객이 구매한 아이템 갯수}}{\text{고객이 구매한 아이템 갯수}} \quad (9)$$

$$\text{F-Measure} = \frac{2(\text{정확률} \times \text{재현율})}{\text{추천 아이템 갯수}} \quad (10)$$

추천받는 대상이 되는 특정 로그인 고객과 고객 분류코드가 동일하고 가장 많은 분포로 이루어진 제품점수대의 군집 데이터를 이용한다. <표 6>는 군집별 추천의 정확도와 재현율을 분석한 결과를 나타낸 것이다.

<그림 2>, <그림 3>, <그림 4>는 <표 6>의 결과를 바탕으로 군집별 정확도와 재현율 그리고 F-measure의 성능 평가이다. 제안시스템은 비록 군집별 평균 정확도에서 30.31%의 결과를 나타났지만 기존 시스템보다 군집별 평균 재현율에서 13.41%의 결과와 군집별 평균 F-measure에서 17.76% 높은 결과를 나타내었다. 이로써 실험의 결과가 기존 시스템보다 향상된 결과가 도출되었음이 입증되었다.

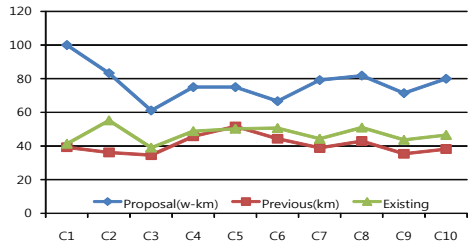
다음 <그림 5>는 웹상의 추천시스템을 위한 k-means 기법과 베이시안 네트워크를 이용한 가중치 선호도 군집 방법의 사이트를 보여 주고 있다.

제안 방법은 시간에 따라 변화하는 고객의 선호도는 시간이 지남에 따라 바뀔 수 있으므로, 시간에 따른 고객 선호도 변화가 유연하게 선호도 계산이 가능한 베이시안 네트워크 방법과 가중치 선호도 방법을 이용한 k-means 군집화를 통하여 최종 추천 단계에서 보다 정확한 추천 정보를 제공

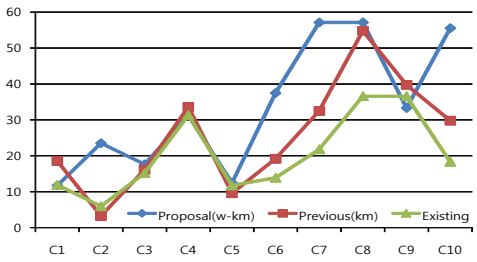
〈표 6〉 군집별 추천의 정확도와 재현율 결과

군집	Proposal(w-km)			Previous(km)			Existing		
	Precision1	Recall1	F-measure1	Precision2	Recall2	F-measure2	Precision3	Recall3	F-measure3
1	100	11.76	21.05	39.21	18.55	23.63	41.29	11.90	17.51
2	83.33	23.53	36.06	36.11	3.23	5.89	55.19	5.95	10.47
3	61.11	17.65	26.39	34.52	16.13	21.14	38.97	15.18	20.88
4	75.00	31.25	43.04	45.68	33.60	36.42	48.79	31.32	35.64
5	75.00	12.50	21.11	51.67	9.60	15.40	50.22	11.74	18.28
6	66.67	37.50	46.02	44.27	19.20	25.84	50.60	13.88	21.03
7	79.17	57.14	64.05	38.83	32.49	33.64	44.26	21.81	27.65
8	81.82	57.14	65.67	42.93	54.79	44.89	50.93	36.60	39.64
9	71.43	33.33	44.29	35.33	39.73	36.60	43.60	36.60	37.82
10	80.00	55.56	63.91	38.17	29.79	31.99	46.53	18.32	25.10

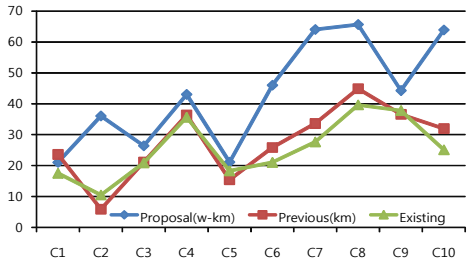
하였으며 학습 에이전트에서 전처리과정으로 가중치 선호도 방법을 이용한 군집화 및 가중치 선호도 계산을 수행함으로써 실시간 추천에서 요구되는 즉시성을 확보할 수 있었다.



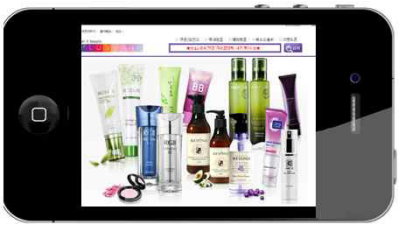
〈그림 2〉 정확률에 따른 추천평가 결과



〈그림 3〉 재현율에 따른 추천평가 결과



〈그림 4〉 F-measure에 따른 추천평가 결과



〈그림 5〉 화장품 추천 사이트

4. 결론 및 향후 과제

대규모 거래가 이루어지는 전자상거래에서 거래되는 아이템들에 대한 가중치는 효과적인 속성값이 반영되어 동적 환경인 실제 유비쿼터스 상거래 추천에서 정확성을 제공해야 한다.

본 논문에서는 고객점수 가중치를 적용한 가중치 선호도를 이용한 새로운 군집방법으로 추천시스템을 위한 k-means 기법과 베이시안 네트워크를 이용한 가중치 선호도 군집 방법을 제안하였다.

성능평가를 위해 현업에서 사용하는 인터넷 화장품 쇼핑몰의 데이터를 기반으로 데이터 셋을 구성하여 기존의 방법과 비교 실험을 통해 성능을 평가하여 효용성과 타당성을 입증하였다. 유비쿼터스 상거래에서 실시간성이 요구되는 모바일 웹 앱을 이용한 추천에서 이전 방법보다 개선된 결과로 추천시스템을 구현할 수 있었다. 향후 연구로는 의사결정 트리(Decision tree)를 이용한 추천방법에 관한 연구가 진행될 예정이다.

참 고 문 헌

[1] Cho, Y. S., Gu, M. S., and Ryu, K. H., “Development of Personalized Recommendation System using RFM method and k-means Clustering”, In : *KSCI*, Vol. 17, No. 6, 2012, pp. 163-171.

[2] Cho, Y. S., Jeong, S. P., and Ryu, K. H., “Implementation of Personalized u-commerce Recommendation System using Preference of Item Category based on RFM”, the 6th International Conference on Ubiquitous Information Technologies and Applications. In : the 6th International Conference on Ubiquitous Information Technologies and

Applications, Dec, 2011, pp 109-114.

[3] Cho, Y. S., Moon, S. C., Noh, S. C., and Ryu, K. H., "Implementation of Personalized recommendation System using k-means Clustering of Item Category based on RFM", In: 2012 IEEE International Conference on Management of Innovation and Technology Publication, June. 2012.

[4] Cho, Y. S., Moon, S. C., Jeong, S. P., Oh, I. B., and Ryu, K. H., "Clustering Method using Item Preference based on RFM for Recommendation System in u-Commerce", In : Cho, Y. S., Moon, S. C., Jeong, S. P., Oh, I. B., Ryu, K. H. (eds.), Ubiquitous Information Technologies and Applications, *LNEE, Springer, Heidelberg*, Vol. 214, 2012, pp. 353-362.

[5] Kim, J. H., Kim, Y. J., Jeong, K. Y., Rim, K. W., and Lee, J. H., "User and Item based Collaborative Filtering Using Classification Property Naive Bayesian", In : *KCA*, Vol. 7 No. 11, 2007, pp. 23-33.

[6] Lee, Y. K., Kim, W. T., Jung, Y. J., Kim, K. D., Ryu, K. H., "Cluster Analysis of Climate Data for Applying Weather Marketing", In: *Journal of geographic information system association of Korea*, Vol. 7, No. 3, 2005, pp. 33-44.

[7] Choi, J. H., Kim, D. S., and Rim, K. W., "Dynamic Recommendation System for a Web Library by Using Cluster Analysis and Bayesian Learning", In : *KCI*, Vol. 12, No. 5, 2002, pp. 385-392.

[8] Herlocker, J. L., Kosran, J. A., Borchers, A., and Riedl, J., "An Algorithm Framework for Performing Collaborative Filtering", In : Proceedings of the 1999 Conference on Research and Development in Information Retrival, 1999.

저자소개

박 화 범

광운대학교 전자통신공학 석사, 광운대학교 전자통신공학 박사 수료 후, 2000년부터 남서울대학교 겸임교수로 재직 중이다. 주여 관심분야는 디지털 영상 압축, 워터마킹, 데이터마이닝 등이다.

조 영 성

연세대학교 전산학과 공학석사, 충북대학교 전산학과 공학박사, 미국 CDC/Stratus System S.E Manager, (주)네오아이엔씨 CEO역임, 현재 정보처리기술지도사(중기청), (주)컴트리 연구소장, 동양미래대학 겸임교수, 주요 관심분야는 시공간 데이터베이스, 유비쿼터스 컴퓨팅, 데이터마이닝, 기계학습, u-커머스, ebXML 등이다.

고 형 화

서울대학교 전자공학과 학사, 서울대학교 전자공학과 석사, 서울대학교 전자공학과 박사 후, 1985년부터 광운대학교 전자통신 공학과 교수로 재직 중이다. 주요 관심분야는 영상통신, 2진문서 압축, 임베디드 시스템, Wavelet 부호화 등이다.