Journal of The Korea Society of Computer and Information
Vol. 17 No. 6 , June 2012

2012-17-6-4-5

RFM기법과 k-means 기법을 이용한 개인화 추천시스템의 개발

조 영 성*, 구 미 숙**, 류 근 호***

Development of Personalized Recommendation System using RFM method and k-means Clustering

Young Sung Cho*, Mi Sug Gu**, Keun Ho Ryu***

요 약

기존 추천시스템의 명시적((Explicit) 협력 필터링 방법은 실용화 되었으나 정확한 아이템의 속성이 반영되지 않는 문제와 희박성과 확장성 문제가 여전히 남아 있다. 본 논문에서는 실시간성과 민첩성이 요구되는 유비쿼터스 상거래에서 고객에게 번거로운 질의 응답 과정이 없이 묵시적인(Implicit) 방법을 이용하여 RFM(Recency, Frequency, Monetary)기법과 k-means 기법을 이용한 개인화 추천시스템을 제안한다. 구매 가능성이 높은 아이템을 추출하기 위해서 고객데이터와 구매이력 데이터를 기반으로 아이템의 속성 반영이 가능한 RFM기법과 k-means 클러스터링을 이용한다. 제안 방법으로 추천의 효율성이 높은 아이템 추천이 가능하도록 고객정보의 속성 변수의 특징 벡터가 적용된 클러스터링 작업과 군집내의 아이템 카테고리 선호도 계산 작업의 전처리를 수행한다. 성능평가를 위해 현업에서 사용하는 인터넷 화장품 아이템 쇼핑몰의 데이터를 기반으로 데이터 셋을 구성하여 기존 시스템과 비교 실험을 통해 성능을 평가하여 효용성과 타당성을 입증하였다.

▶ Keywords : RFM기법, 협력 필터링, 클러스터링, 추천시스템

Abstract

Collaborative filtering which is used explicit method in a existing recommedation system, can not only reflect exact attributes of item but also still has the problem of sparsity and scalability, though it has been practically used to improve these defects. This paper proposes the personalized recommendation system using RFM method and k-means clustering in u-commerce which is required by real time accessablity and agility. In this paper, using a implicit method which is is

제1저자 : 조영성, 교신저자 : 구미숙, 책임저자 : 류근호
투고일 : 2012. 03. 01, 심사일 : 2012. 03. 30, 게재확정일 : 2012. 04. 10.
* 동양미래대학 전산정보학부(School of Computer Science & Information, DongYang Mirae University)
** 충북대학교 전기전자컴퓨터공학부(School of Electrical & Computer Enginnering, Chungbuk National University)
※ 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2012-0000478).

not used complicated query processing of the request and the response for rating, it is necessary for us to keep the analysis of RFM method and k-means clustering to be able to reflect attributes of the item in order to find the items with high purchasablity. The proposed makes the task of clustering to apply the variable of featured vector for the customer's information and calculating of the preference by each item category based on purchase history data, is able to recommend the items with efficiency. To estimate the performance, the proposed system is compared with existing system. As a result, it can be improved and evaluated according to the criteria of logicality through the experiment with dataset, collected in a cosmetic internet shopping mall.

▶ Keywords : RFM Method, Collaborative Filtering, Clustering, Recommend System

I. 서 론

유비쿼터스 컴퓨팅 환경하에서 유무선 인터넷이 생활의 일부가 되어가면서 정보의 양도 급속도로 늘어나고 있으며, 이로 인해 많은 데이터 속에서 정보를 찾아내는 기술이 부각되고 있다. 추천시스템은 고객를 대신하여 적합한 아이템을 빠른 시간 내에 추천하고, 추천된 내용이 또한 정확하다면 고객은 만족감을 얻을 수 있다. 기존의 명시적인 방법의 협력 필터링은 희박성과 확장성의 문제가 여전히 존재하며 아이템의 속성도 반영되지 않는 문제점이 남아 있다. 본 논문에서는 모바일 단말기를 사용하는 고객에게 번거로운 질의 응답 과정 없이 묵시적인 방법을 이용하여 아이템 속성 분석이 가능한 RFM기법과 k-means 클러스터링을 이용한 개인화 추천시스템을 제안한다. 지능형 추천시스템의 구성은 기업의 비즈니스 전략이 되어 가고 있다. 고객에 맞는 아이템 세분화 분석 기법인 RFM 이용한 개인화 추천 방법에 대한 연구[1,2,3,4,5]가 활발히 진행되고 있다. 본 논문은 기존의 협력 필터링의 확장성의 문제와 아이템의 속성이 반영되지 않는 문제, 추천시 대규모 거래 데이터 처리 시간의 지연 문제, 추천 처리의 실시간성과 민첩성 확보 문제에 대한 방법을 제공한다. 본 논문의 구성은 다음과 같다. 제 Ⅱ장은 관련 연구를 다루었으며 제 Ⅲ 장에서는 제안 추천시스템 설명하며 제 Ⅳ 장에서는 실험 및 성능 평가를 실행하며 마지막으로, 제 Ⅴ장에서는 본 논문의 결론과 향후 연구에 대하여 기술한다.

II. 관련연구

2.1 협력 필터링

협력 필터링은 고객들의 선호도 정보를 바탕으로 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 사용하는지에 따라 구분을 한다. 또한 추천 정보를 제시하는 방법으로 협력 필터링, 인구통계학적 필터링, 규칙기반 필터링, 내용기반 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[6,7]. 본 논문에서는 이러한 문제점 해결 방안으로 고객에게 번거로운 질의 응답 과정이 없이 묵시적인 방법으로 고객정보와 아이템정보, 구매이력정보를 이용한다.

2.2 클러스터링(clustering)

클러스터링은 임의의 데이터 집합으로부터 서로 유사한 속성을 가지는 데이터의 군집(cluster) 또는 세그먼트(segment)를 추출하는 기법을 의미한다. 고객의 구매 이력 정보로부터 구매 상품의 특징에 따라 고객들을 클러스터링하거나, 고객들의 신상 정보를 이용해 신상 정보의 유사성에 따라 고객들을 클러스터링 하는 것을 의미한다.

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 a 와 k 사이의 거리는 식1과 같이 계산하고, 식에서 a_i는 고객 a 의 속성(차원) i 에 대한 선호도 값을 의미한다.

$$d_{a,k}=\sqrt{\sum_i (a_i-k_i)^2} \qquad \text{식1}$$

본 논문에서는 고객점수 및 Social data로 구성된 인구통계학적 변수(나이, 성별, 직업, 고객 등)가 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[8].

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ & & \cdot & \\ & & \cdot & \\ & & \cdot & \\ x_{f1} & x_{f2} & \cdots & x_{fn} \end{pmatrix} \quad \text{식2}$$

다음은 k-means 기법의 처리 절차를 나타낸 것이다.

표 1. k-means 기법의 처리 절차 알고리즘

Table 1. Procedure algorithm for processing by k-means clusterring

Step 1: p개의 변수로 기술되는 n개의 자료에 대해 다음과 같은 자료 행렬을 만든다. 즉, X는 p차원에서의 n개의 벡터들을 나타낸다.

Step 2: 벡터의 형태로 k개의 군집의 위치를 임의로 정한다. 여기에서 군집의 개수 k는 분석자가 결정한다.

Step 3: 각각의 자료에 대해 k개의 평균 벡터와의 유클리드 거리를 계산하고 가장 가까운 군집에 그 자료를 할당한다.

Step 4: 각 군집에 대해 평균 벡터를 다음과 같이 다시 계산한다. 여기서 N_m은 m 번째 군집에 속한 자료의 개수이다.

$$Y_{mj}=\frac{1}{N_m}\sum_{i=1}^{N_m} X_{ij} \qquad \text{식3}$$

(j =1,2, ...,p: m=1,2, ...,k)

Step 5: 각각의 자료에 대해 그 자료가 속한 군집(m)을 포함한 모든 군집의 평균 벡터와의 거리를 계산하고 다음과 같은 조건이 만족되면 그 자료가 속한 군집을 바꾼다. (식3)은 m번째 군집에 속한 i번째 자료가 j번째 군집에 속해 있다고 가정할 경우 두 군집의 평균 벡터와의 차이를 의미한다.

$$\frac{N_m}{N_m+1}\sum_{j=1}^{p}(X_{il}-Y_{ml})^2-\frac{N_j}{N_j-1}(X_{il}-Y_{jl})>0$$

식4

Step 6: 자료의 재할당이 없을 때까지 Step4,5를 반복한다. k-means 기법에서 군집의 개수k는 분석자가 정할 수 있다.

2.3 RFM

RFM은 세 가지 요소로 구성되어진다. 첫째, 최근성은 최근에 구매한 고객이 앞으로 구매할 가능성이 높다는 판단 하에 최근 구매일이 가까울수록 높은 점수를 부여한다. 둘째, 빈도성은 일정 기간 동안의 거래 빈도에 따라 고객을 세분화하는 것으로, 빈도가 높은 고객일수록 앞으로 구매할 가능성이 높다는 판단 하에 거래 빈도가 높을수록 높은 점수를 부여한다. 셋째, 총구매액은 일정기간 동안의 아이템 구입에 사용한 총 구매금액에 따라 고객을 세분화하는 것으로, 총 구매금액이 높은 고객이 앞으로 구매할 가능성이 높다는 판단 하에 총 구매금액이 높을수록 높은 점수를 부여한다. 고객 데이터는 RFM을 이용하여 세분화가 가능하다. 각 요소마다 5개의 세분화 세그먼트로 나누져서 전체 고객은 결국 5 × 5 × 5 = 125 개의 세그먼트로 분할된다. RFM은 가치 있는 고객을 추출해 내어 이를 기준으로 고객을 분류할 수 있는 매우 간단하면서도 유용하게 사용될 수 있는 방법으로 알려져 있다. 따라서 RFM은 구매 가능성이 높을 고객을 선정하기 위한 데이터 분석 방법이다. 세 가지 요소를 기준으로 고객 각각에 대해 점수를 부여하고 세 가지 기준의 가중치를 주어 RFM점수를 계산하게 된다. 이 RFM점수를 고객 가치를 평가하는 지표로 삼는 방식이 RFM에 의한 고객 점수 부여 방법이다[5]. 다음은 RFM점수 산출식의 예를 나타낸 것이다.

RFM점수 = (A*R + B*F + C*M) * 20 식5

RFM점수의 가중치(A, B, C)는 경영 상태이나 경영 전략에 따라 변경이 가능하다. RFM점수의 합계는 최고점수는 100점, 최하점수는 0점이다. RFM점수를 위해서 사용되는 R, F, M 요소는 예측력이 강한 변수이다.

III. RFM기법과 k-means 클러스터링을 이용한 개인화 추천시스템

이 장에서는 별도의 고객 프로파일을 생성하지 않고 고객 데이터와 구매이력데이터를 기반으로 클러스터링을 위한 고객정보 변수의 정의와 아이템 카테고리 관련 정의와 k-means 기법을 이용한 이웃고객 생성 알고리즘을 기술하고, 제안시스템 구조 및 제안시스템 절차 알고리즘을 기술한다.

[정의 3.1] 고객정보 변수가 k개의 속성을 가지다면, 고객정보 변수 집합 P = { P_1, P_2,…, P_k}로 표현한다.

예를 들어 k = 3인 경우 P = { P_1, P_2, P_3}이다. 여기에서 속성 P_j가 S_j개의 다른 속성값을 가지면, P_j = { q_1, q_2…, q_{S_j}} 로 표현된다.

[정의 3.2] 고객정보 변수가 속성을 n개씩 클러스터링 할 경우 P_j를 n개씩 클러스터링 할 경우 P_{j1}={ $q_1, q_2, \ldots, q_n$}, P_{j2} ={ q_{n+1}, q_{n+2}, q_{2n}} …. P_{jm}={ q_{m+1}, q_{m+2}, q_{2m}}로 표현한다. 예를 들어 P = {age, gender, occupation}라면 P_2 = gender이고, P_2 = {남,여}로 구성된다. 이 경우 P_{2_1} = {남}, P_{2_2} = {여}이다. 본 논문에서는 이 장에서의 [정의 3.2]를 기반으로 추천시스템을 설계한다. 인터넷 쇼핑몰에서 구성된 고객의 신상정보인 고객정보의 고객 점수 및 고객 분류코드의 속성 변수를 이용하여 고객 데이터베이스와 구매이력 데이터베이스를 기반으로 고객 특성 벡터를 반영하여 k-means 기법을 이용하여 클러스터링한다. 이웃 고객 군집과 이웃 구매이력 군집으로부터 고객 특성 벡터를 적용하여 고객이 가장 선호하는 아이템들을 추출할 때 소요되는 시간의 복잡도를 Big O에 가깝게 감소시킬 수 있다. 따라서 아이템 추천 과정에 소요되는 전체 처리시간을 감소시킬 수 있다. 또한 고객집합 U는 U = { u_1, u_2, u_3…, u_n}, 아이템집합 I는 I = { i_1, i_2, i_3… , i_m}, 아이템 카테고리 집합 C는 C= { c_1, c_2, c_3 … , c_r}로 표현한다. 이때 쇼핑몰의 고객은 u_j, 쇼핑몰의 아이템을 i_j, 아이템 카테고리를 c_k로 나타낸다. 아이템들은 특징에 따라 특정한 카테고리에 속하게 되며 아이템 카테고리는 단일 계층구조(single hierarchy)의 트리구조를 갖는다. 아이템 카테고리의 구조는 [정의 3.3]과 같다.

[정의 3.3] 추천 시스템에서 사용하는 아이템 카테고리는 트리구조를 갖는다.

[정리 3.3] 아이템 i_j(i∈m)이 소속된 카테고리를 C_k(i_j)= C_j(j∈ r) 라고 하면, i_j의 상위 카테고리는 TC(i_j)= C_k(k∈r)이다. 이때 C_k를 C_j의 부모 카테고리라 한다.

[증명] [정의 3.3]에 의해서 아이템 카테고리는 트리구조를 갖기 때문에 함수 TC(i_j)의 값은 C_k가 된다. 따라서 C_j의 부모 노드는 C_k이다. [정의 3.3]과 [정리 3.3]로부터 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 3.4] 상위 카테고리의 선호도는 고객의 하위 카테고리의 선호도를 평균내어 산출한다. 아이템 카테고리에서 하위 카테고리의 고객 선호도를 상위 카테고리의 선호도에 반영하지만, 상위 카테고리의 고객 선호도는 하위 카테고리로 반영하지 않는다.

[정의 3.5] 선호도를 나타내는 함수 Pre_icd(id, cd, date)는 고객(id), 아이템 카테고리에 속한 아이템 코드(cd)를 갖는 아이템, 구매일자에 대한 선호도를 나타낸다. 즉 임의의 구매일자(date)에 고객이 아이템 카테고리 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_1, 접근한 아이템은 i_i 이고 숫자는 아이템에 대한 선호도 구매건수를 나타낸다. 다음은 카테고리 구성도의 예를 나타낸 것이다.

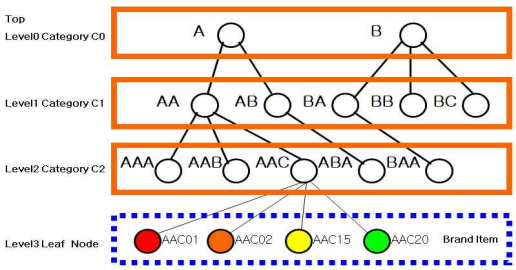


그림 1 카테고리 구성도의 예
Fig. 1. The example on configulatiom of category

3.1 k-means 기법을 이용한 이웃고객 생성 알고리즘

본 논문에서는 인터넷 쇼핑몰에서 설계된 고객의 신상 정보인 고객정보의 속성 변수(고객점수, 나이, 성별, 직업등)가 적용된 고객 데이터베이스와 구매이력 데이터베이스를 이용하여 아이템 카테고리 선호도 기반의 클러스터링하기 위하여 k-means 기법을 적용한다.

표 2. k-means 기법을 이용한 이웃고객 군집 알고리즘
Table 2. k-means clustering algorithm for the neighborhood of customer

입력 : 고객-아이템카테고리-선호도(UCP), 아이템카테고리 테이블(CT)

출력 : 특징 벡터(Feature Vector), 이웃 고객

begin

1. for(CT에서 모든 아이템카테고리)
소분류 아이템카테고리의 UCP내의 브랜드아이템의 선호도를 아이템카테고리에 합산한다.
2. 고객의 특징 벡터(Feature Vector) 생성
2.1 아이템카테고리별로 Pref_UC(u,c)를 합산하고, 이를 정규화 한다.
2.2 각 고객과 아이템카테고리간의 상대적 선호도 산출한다.
2.3 다음 공식을 이용하여 고객의 특징 벡터 V를 생성한다.
아이템카테고리가 M 개 일 때,

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

$$V_i = \sum_k (Pref\_UC(u_i, c_k)) / \sum_i (\sum_k (Pref\_UC(u_i, c_k)))$$

3. 선호도가 높은 아이템 카테고리 정보를 이용하여 이웃 고객을 생성한다.
// k-means 클러스터링 알고리즘을 이용하여 고객 군집을 추출한다.
end.

3.2 시스템 구조

본 시스템에서는 유무선 인터넷 쇼핑몰 환경하에서 제안시스템이 구동되기 위해서 분석 에이전트, 추천 에이전트, 학습 에이전트로 그 기능을 나누어져 서브 시스템이 구동된다. 선택사양으로 데이터마이닝 에이전트가 추가 가능하다. 유무선 웹환경에서 RFM기법과 k-means 클러스터링을 이용한 개인화 추천시스템을 개발하기 위해서 일반 웹 브라우저는 물론 휴대기기에서 풀 브라우저로 인터넷 접속이 가능하도록 하였다. 모바일 웹은 기존의 왑(WAP)의 피쳐폰(Feature phone) 및 스마트 폰 지원을 위한 웹표준을 준수한다. 다음은 추천시스템의 시스템 구성도를 나타낸 것이다.

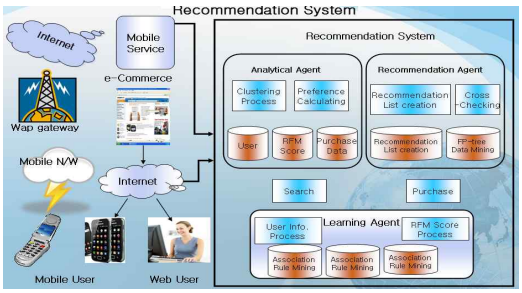


그림 2. 추천시스템의 전체 시스템 구성도
Fig. 2. system configuration for recommend system

3.3 제안시스템 절차 알고리즘

다음 <표 3>은 RFM기법과 k-means 클러스터링을 이용한 제안시스템의 절차 알고리즘을 나타낸 것이다.

표 3. 제안시스템 절차 알고리즘
Table 3. Procedure algorithm for Proposing system

Step 1 : 회원가입시 사용자의 social data를 통해 고객 분류코드 및 고객점수를 부여하여 고객정보를 생성 및 관리한다.
Step 2 : 고객정보에서 로그인 사용자의 고객 분류코드와 같은 군집(Cluster)을 탐색하여 선택한다.
Step 3 : 가장 많은 분포가 이루어진 RFM의 아이템점수대를 적용하여 해당군집을 발췌한다.
Step 4 : 해당군집의 소분류 아이템 카테고리의 선호도를 탐색한다.
Step 5 : 해당 군집에 속한 구매데이터를 기반으로 선호도가 높은 아이템 카테고리를 선택하여 그 카테고리에 속한 단말 브랜드아이템의 선호가 높은 순으로 추천아이템들을 생성한다.
Step 6 : 추천 아이템의 선호도를 스캔하여 선호도가 높은 아이템들을 Top-N의 추천 아이템 목록을 생성한다.
Step 7 : 추천시 추천된 아이템을 로그인 사용자 구매 이력정보와 체크하여 중복 추천되지 않도록 한다.

Ⅳ. 실험 및 성능 평가

4.1 실험 환경

구현 및 실험 환경은 윈도우 운영체제하에서 해당 평가하기 위해서 다음과 같은 웹서버 환경을 사용하였다[5].

- OS: Window XP
- Web Server: Apache HTTP Server Version 2.2.8 / WAP 2.0
- XML/WML2.0/HTML/XHTML/CSS2/JAVASCRIPT
- Server-Side Application : JSP/ PHP 5 Version

5.2.5
- Java 2 SDK, SE 1.4.2_08 - http://java.sun.com/
- Tomcat 5.0.28 - http://jakarta.apache.org/
- http://www.mysql.com/products/mysql/
- MySQL Connector/J 3.0
- http://www.mysql.com/products/connector/j/
- http://wb.mysql.com/ - workbench

RFM기법과 k-means 클러스터링을 이용한 개인화 추천 시스템의 성능을 평가하기 위해서 제안시스템과 기존 시스템의 성능평가를 실험을 통해서 알아본다.

4.2 실험 데이터 구성

RFM기법과 k-means 클러스터링을 이용한 개인화 추천 시스템은 윈도우 XP 환경에서 인터넷 화장품 쇼핑몰을 위한 데이터베이스가 구축되었다. 기존의 명시적인 자료 기반의 추천 방법에 문제점을 해결하고 다음과 같은 제안 방법을 적용하여 제안시스템에 대한 평가를 위해서 실험 데이터를 구성한다. 첫째, 기존의 추천 방법은 아이템의 속성 대한 선호도가 반영되지 않는 문제점이 있어 아이템의 속성 반영을 위해 RFM기법을 적용한다. 둘째, 기존의 추천 방법이 아이템에 대한 다른 고객들의 선호도를 기반으로 하기 때문에 제안 방법의 추천의 효율성을 입증하기 위해 아이템 카테고리 선호도 기반의 클러스터링하기 위하여 k-means 기법을 적용한다. 셋째, 기존의 명시적인 자료 기반의 추천 방법에 여전히 희박성이 존재하고 있어 제안 방법으로 모바일 환경하에 고객에게 번거로운 질의 응답 과정이 없이 묵시적인 방법을 적용한다. 쇼핑몰을 이용해 본 경험이 있는 고객 319명의 고객정보와 현재 화장품을 전문적으로 판매하는 인터넷 화장품 쇼핑몰인 P사의 아이템 분류에서 사용하는 화장품 아이템 580개를 대상으로 구매한 1600건의 구매 데이터를 이용하였다. 2009년 2월 부터 2010 2월까지의 12개월간의 과거의 구매 데이터를 학습 셋으로 사용하였고, 2010년 3월 부터 2010년 5월까지 3개월간의 미래 구매 데이터를 테스트 셋으로 사용하였다 [5].

4.3 분석 및 성능 평가

추천시스템의 전체적인 성능 평가는 두 방향으로 나누어 진행하였다. 예측 값과 실제 값의 차이를 표시하여 정확성 측면에서 성능을 평가하기 위한 MAE방식과 정확도와 재현율을 함께 사용해서 시스템의 전체적인 성능을 평가 할수 있는 F-measure 방식을 사용하였다. MAE는 예측의 정확성을 판단하는데 가장 많이 쓰이는 방법이고, F-measure는 값이 클수록 추천이 우수함을 의미한다. 본 논문에서는 MAE 및 정확도와 재현율, 그리고 F-measure 방식에 대한 실험을 제안시스템과 기존 시스템을 실험하였다. 우선 첫 번째, 실험으로 MAE에 의해 예측의 성능을 평가하였다. 추천시스템의 예측 값의 정확성을 평가하기 위해 MAE(Mean Absolute Error)를 사용하였고 식3과 같이 산출하였다[9].

$$MAE = \frac{\sum_{i=1}^{N} |\varepsilon_i|}{N} \qquad \text{식6}$$

N은 총 예측 회수를 나타내고, ε_i는 예측 값과 실제 값의 오차를 나타내며 i는 각 예측 단계를 나타낸다. <표 4>는 식6를 이용하여 예측값의 정확성 평가를 수행한 결과이다.

표 4. 제안 및 기존 시스템의 MAE에 의한 성능평가
Table 4. The result for table of MAE by comparing proposal system with existing system

	P_count	Proposal	Existing
MAE	50	0.47	0.65
	100	0.23	0.32
	300	0.07	0.08
	500	0.05	0.06

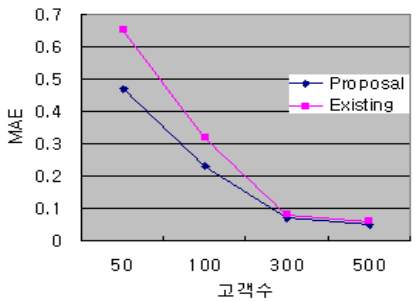


그림 3. 제안 및 기존 시스템의 MAE에 의한 성능평가
Fig. 3. The result for the graph of MAE by comparing proposal system with existing system

다음은 두 번째, 실험으로 정확도와 재현율, F-measure에 대한 실험이다. 성능은 social data에 기반한 화장품 아이템 추천에서의 추천의 유효성과 추천시스템의 전체적인 성능 평가 방향으로 진행하였다. 우선 초기 화장품 아이템 추천의 유효성을 실험에 참가한 고객들의 구매데이터와 제시되는 화장품 아이템의 비교를 통해 이루어졌으며, 추천의 정확성을 평가하기 위하여 정보검색 분야에서 보편적으로 사용되는 평

가 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 아이템 갯수 중에서 실제로 고객이 구매한 아이템의 비율이다. 재현율은 추천시스템의 추천 제품 중에서 실제로 고객이 구매한 제품의 비율이다. F-measure는 정확률과 재현율을 보완하기 해서 결합한 평가방법으로 시스템의 전체적인 성능을 평가 할수 있는 척도로 사용하였다.

$$\text{정확률} = \frac{\text{고객이 구매한 아이템 갯수}}{\text{추천아이템 갯수}} \quad \text{식7}$$

$$\text{재현율} = \frac{\text{추천시스템의 추천아이템중 고객이 구매한 아이템 갯수}}{\text{고객이 구매한 아이템 갯수}} \quad \text{식8}$$

$$\text{F-Measure} = \frac{2(\text{정확률} * \text{재현율})}{\text{추천아이템 갯수}} \quad \text{식9}$$

추천받는 대상이 되는 특정 로그인 고객과 고객 분류코드가 동일한 상황에서 RFM의 아이템점수대의 특징 벡터가 적용되지 않은 기존 시스템의 군집 데이터와 제안시스템의 경우 가장 많은 분포를 이루고 있는 RFM의 아이템점수대의 고객 특징 벡터가 적용된 군집 데이터를 비교한다. <표 5>은 군집별 추천의 정확도와 재현율을 분석한 결과를 나타낸 것이다.

표 5. 군집별 추천의 정확도와 재현율 결과
Table 5. Result of Precision and Recall for Recommendation by clusters

군집	제안시스템			기존 시스템		
	정확률	재현율	F-measure	정확률	재현율	F-measure
G1	56.98	91.44	65.90	52.10	50.89	47.18
G2	100.00	27.27	42.86	38.03	15.18	20.71
G3	100.00	28.88	44.81	42.08	16.07	22.34
G4	48.79	55.70	48.41	48.02	31.32	35.31
G5	49.36	52.53	48.09	47.82	29.54	34.39
G6	55.50	23.93	32.19	43.29	21.81	27.32
G7	52.49	38.37	41.95	51.67	34.98	39.18
G8	50.41	47.40	45.24	48.60	43.21	42.02
G9	50.93	37.23	40.03	48.56	36.60	38.26
G10	47.41	27.27	32.60	47.41	26.81	32.26
G11	43.60	37.23	38.17	42.63	36.60	37.08
G12	46.68	28.45	32.62	44.78	25.19	29.71
G13	67.18	20.69	31.17	42.77	18.32	23.58
G14	67.23	62.50	60.94	61.93	55.34	53.87

다음은 <표 5>의 기존 시스템과 제안시스템을 비교하기 위한 SQL을 이용한 실험 데이터 결과의 일부분를 나타낸 것이다.

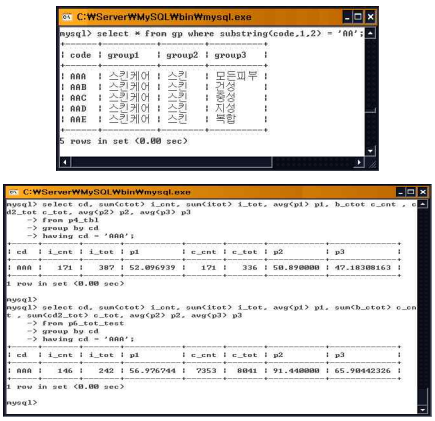


그림 4. SQL을 이용한 실험데이터 결과
Fig. 4. The result of experimental data using SQL

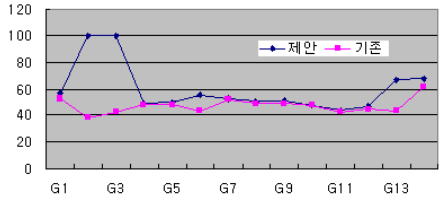


그림 5. 군집별 정확률에 따른 추천평가 결과
Fig. 5. The result of recommending ratio for recommendation each cluster by precision

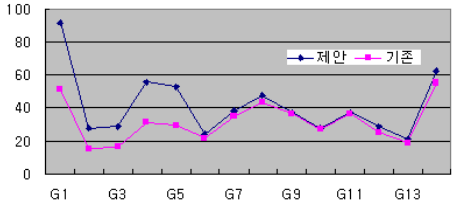


그림 6. 군집별 재현율에 따른 추천평가 결과
Fig. 6. The result of recommending ratio for recommendation each cluster by recall

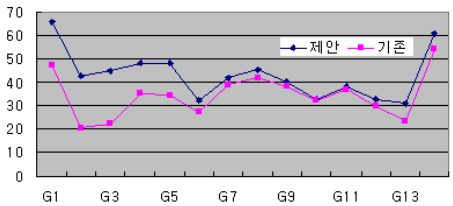


그림 7. 군집별 F-measure에 따른 추천평가결과
Fig 7. The result of recommending ratio for recommendation each cluster by F-measure

<그림 5>. <그림 6>. <그림 7>은 <표 5>의 결과를 바탕으

로 군집별 정확도와 재현율 그리고 F-measure의 성능 평가이다. 제안시스템은 기존 시스템보다 12.63% 높은 정확도와 9.79% 높은 재현율, 그리고 8.68% 향상된 F-measure의 결과를 나타내었다. 다음은 웹상의 RFM기법과 k-means 클러스터링을 이용한 개인화 추천시스템에서 추천된 화장품 사이트을 보여주고 있으며 스마트폰으로도 이용가능하다.

그림 8. 화장품 아이템 추천 결과
Fig 8. The result of recommending items of cosmetics

기존의 시스템은 대규모 데이터베이스의 거래 데이터처리로 인한 추천 아이템을 위한 선호도 계산 작업 등에 많은 시간이 소요되었다. 제안시스템은 고객정보의 RFM기법과 k-means 클러스터링을 이용한 개인화된 추천시스템으로 추천시스템의 성능도 향상되었으며 군집내의 아이템 카테고리 선호도 작업을 처리함으로써 처리능력 향상과 실시간에서의 즉시성 확보가 가능하였다.

V. 결론 및 향후 과제

요즘 실시간성과 민첩성이 요구되는 유비쿼터스 컴퓨팅 환경하에 응용 분야로써의 유비쿼터스 상거래는 각광을 받고 있다. 고객들은 모바일 단말기 화면의 제약으로 원하는 정보를 검색하거나 평가자료 작성을 위한 설문에 대한 번거로운 질의과정에 정확하게 답하는 것은 불편하고 어려운 일 것이다. 본 논문에서는 묵시적인 방법을 이용하여 구매 가능성이 높은 아이템을 찾기 위해서 아이템 속성 분석이 가능한 RFM 기법과 k-means 클러스터링을 이용한 개인화 추천시스템을 제안하였다. k-means 기법을 적용하여 이웃 고객 군집과 이웃 구매 이력 군집 형성을 위해 클러스터링 작업을 전처리하여 고객이 가장 선호하는 아이템들을 추출할 때 소요되는 시간감소가 가능하여 아이템 추천 과정에 소요되는 전체 처리시간을 감소시킬 수 있었다.. 클러스터링 작업과 군집내의 선호도 계산 작업을 전처리 함으로써 유비쿼터스 상거래에서 요구되는 추천 처리의 즉시성과 민첩성 확보가 가능하다. 성능평가를 위해 현업에서 사용하는 인터넷 화장품 아이템 쇼핑몰의 데이터를 기반으로 데이터 셋을 구성하여 기존의 방법과 비교실험을 통해 성능을 평가하여 효용성과 타당성을 입증하였지만 제한점으로 실용화 단계에서 대용량 데이터베이스에 대한 실험을 통한 시스템 튜닝 작업이 남아 있다. 아울러 향후 연구로는 유비쿼터스 컴퓨팅 환경하에 구매 가능성이 높은 아이템을 추천하기 위해 묵시적인 방법으로 RFM 가중치를 적용한 클러스터링기법 이용한 개인화 추천시스템에 대한 고찰과 연구가 필요할 것으로 생각된다.

참 고 문 헌

[1] Young Sung Cho, Moon Haeng Heo, Keun Ho Ryu, “Implementation of Personalized Recommendation System using RFM method in Mobile Internet Environment”, KSCI, 13th-2 Vol, pp 1-5, Mar, 2008

[2] Young Sung Cho, Keun Ho Ryu, "Implementation of Personalized Recommendation System using Demographic data and RFM method in e-Commerce", 2008 IEEE International Conference on Management of Innovation & Technology Publication, 2008.

[3] Jin Byeong Woon, Young Sung Cho, Keun Ho Ryu, “Personalized e-Commerce Recommendation System using RFM method and Association Rules”, KSCI, 15th-12 Vol, pp 227-235, Dec, 2010

[4] Young Sung Cho, Seon-phil Jeong, Keun Ho Ryu,

"Implementation of Personalized u-Commerce Recommendation System using Preference of Item Category based on RFM", the 6th International Conference on Ubiquitous Information Technologies & Applications, pp109-114, Dec, 2011

[5] Young Sung Cho, Keun Ho Ryu, "Personalized Recommendation System using FP-tree Mining based on RFM ,KSCI, 17th-2 Vol, Feb., 2012

[6] P. Resnick, et. al., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," Proc. of ACM CSCW'94 Conference on Computer Supported Cooperative Work, pp.175-186, 1994.

[7] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of Dimensionality Reduction in Recommender System - A Case Study," In ACM WebKDD 2000 Web Mining for E-Commerce Workshop, 2000.

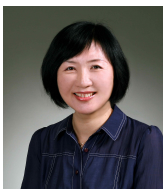
[8] Yang-Koo Lee, Won-Tae Kim, Young-Jin Jung, Kwang-Deuk Kim, Keun-Ho Ryu, "Cluster Analysis of Climate Data for Applying Weather Marketing", Journal of the Research Institute for Computer and Information Communication, 12th-2 Vol, Nov, 2004

[9] Jonathan L. Herlocker, Joseph A. Kosran, Al Borchers, and John Riedl, "An Algorithm Framework for Performing Collaborative Filtering", Proceedings of the 1999 Conference on Rechearch and Development in Information Retrival, 1999

저 자 소 개

조 영 성

1989년 연세대학교 전산학과(공학석사)
2008년 충북대학교 전산학과(공학박사)
1982~2009년 국세청 전산실, 미국 CDC Cyber System/SE Manager, 미국 Stratus FT System/SE Manager, 네오아이엔씨(CEO), 가나소프트(대표)
2010년~현재 가나소프트(고문), 한국경영기술컨설팅협회(전문위원), 기술지도사(중기청), 동양미래대학 전산정보학부 산업체겸임/교수
관심분야 : 시공간 데이터베이스, 유비쿼터스 컴퓨팅 및 GIS, 데이터 마이닝, 기계학습, 웹서비스, ebXML
Email : gumisug@dblab.chungbuk.ac.kr



구 미 숙

1988년 충남대학교 영문학과(문학사)
2005년 충북대학교 전산학과(이학석사)
2005년~현재 충북대학교 전산학과 박사과정수료
관심분야 : XML, 시공간 데이터베이스, 유비쿼터스 컴퓨팅, 바이오 인포매틱스, 데이터 마이닝
Email : gumisug@dblab.chungbuk.ac.kr

류 근 호

1976년 숭실대학교 전산학과(이학사)

1980년 연세대학교 공학대학원 전산전공 (공학석사)

1988년 연세대학교 대학원 전산전공 (공학박사)

1976~1986년 육군군수 지원사 전산실(ROTC 장교), 한국전자통신연구원(연구원), 한국방송통신대 전산학과 (조교수) 근무

1989년~1991년 Univ. of Arizona Research Staff (TempIS 연구원, Temporal DB)

1986년~현재 충북대학교 전기전자 컴퓨터공학부 교수

관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 지식기반 정보검색 시스템, 유비쿼터스 컴퓨팅 및 스트림데이터처리, 데이터 마이닝, 데이터베이스 보안, 바이오 인포메틱스

Email : khryu@dblab.chungbuk.ac.kr