韓國컴퓨터情報學會 論文誌
第17卷 第2號, 2012. 2.

2012-17-2-5-2

RFM기반 FP-tree 마이닝을 이용한 개인화 추천시스템

조 영 성*, 류 근 호**

Personalized Recommendation System using FP-tree Mining based on RFM

Young Sung Cho*, Ryu Keun Ho**

요 약

기존의 연관규칙을 이용한 추천시스템은 매번 계속적으로 대량의 데이터를 스캔해야 하므로 속도가 느릴 뿐 아니라 확장성 문제와 정확도 문제가 있다. 본 논문에서는 사용자의 평가 자료에 의존하지 않고 묵시적인(Implicit) 방법을 이용하여 RFM(Recency, Frequency, Monetary)기반 FP-tree 마이닝을 이용한 개인화 추천시스템을 제안한다. 구매 가능성이 높은 아이템을 찾기 위해서 고객정보와 구매이력정보를 기반으로 고객과 아이템의 속성 반영이 가능한 RFM기법과 FP-tree 마이닝을 이용한다. 제안 방법으로 RFM기반의 FP-tree 마이닝을 이용하여 후보집합의 발생없이 빈발항목을 구성하고 연관규칙을 생성한다. 생성된 연관규칙의 지지도, 신뢰도, 향상도를 사용하여 추천 효율성이 높은 아이템 추천이 가능하다. 성능평가를 위해 현업에서 사용하는 인터넷 화장품 아이템 쇼핑몰의 데이터를 기반으로 데이터 셋을 구성하여 기존의 시스템과 비교 실험을 통해 성능을 평가하여 효용성과 타당성을 입증하였다.

▸Keywords : RFM기법, 빈발패턴트리 마이닝, 추천시스템

Abstract

A exisiting recommedation system using association rules has the problem, such as delay of processing speed from a cause of frequent scanning a large data, scalability and accuracy as well.

• 제1저자 : 조영성, 교신저자 : 조영성, 책임저자 : 류근호
• 투고일 : 2011. 11. 3, 심사일 : 2011. 11.26, 게재확정일 : 2011. 12. 6.
* 동양미래대학 전산정보학부 (School of Computer Science & Information, DongYang Mirae University)
** 충북대학교 전기전자컴퓨터공학부 (School of Electrical & Computer Enginnering, Chungbuk National University)
※ 이 논문은 2011년도 정부(교육과학기술부)의 재원인 한국연구재단의 지원을 받아 수행된 연구(No. 한국연구 2011-0001044)와 교육과학기술부로부터 지원받아 수행된 연구(지역거점연구단육성사업 / 충북BIT연구중심대학사업단)와 국토해양부 첨단도시기술개발사업 - 지능형국토정보기술혁신사업과제의 연구비지원(과제번호:06국토정보B01)에 의해 수행되었습니다.

In this paper, using a Implicit method which is not used user's profile for rating, we propose the personalized recommendation system which is a new method using the FP-tree mining based on RFM. It is necessary for us to keep the analysis of RFM method and FP-tree mining to be able to reflect attributes of customers and items based on the whole customers' data and purchased data in order to find the items with high purchasability. The proposed makes frequent items and creates association rule by using the FP-tree mining based on RFM without occurrence of candidate set. We can recommend the items with efficiency, are used to generate the recommendable item according to the basic threshold for association rules with support, confidence and lift. 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, FP-tree Mining, Recommendation System

Ⅰ. 서론

유비쿼터스 네트워크 환경에서 스마트 폰과 같은 지능형 이동 단말기로 유무선 인터넷을 즐기는 것이 생활의 일부가 되어가면서 정보의 양도 급속도로 늘어나고 있으며, 이로 인해 많은 데이터 속에서 정보를 찾아내는 기술이 부각되고 있다. 추천시스템은 사용자를 대신하여 적합한 아이템을 빠른 시간 내에 추천하고, 추천된 내용이 또한 정확하다면 고객은 만족감을 얻을 수 있다.

인터넷 기업은 지능형 추천시스템을 구성하는 것이 비즈니스의 성공 전략이 되고 있다. 기존의 연관규칙을 이용한 추천시스템은 매번 계속적으로 대량의 데이터를 스캔해야 하므로 속도가 느릴 뿐 아니라 확장성 문제와 정확도 문제가 있다. 본 논문은 이런 문제점과 구매 가능성이 높은 추천 아이템을 위하여 고객의 니즈의 파악이 가능한 RFM기반의 FP-tree 마이닝을 이용한 개인화 추천시스템을 제안한다. 성능평가를 위해 현업에서 사용하는 인터넷 화장품 아이템 쇼핑몰의 데이터를 기반으로 데이터 셋을 구성하여 기존의 방법과 비교 실험을 통해 성능을 평가하여 효용성과 타당성을 입증한다. 본 논문의 구성은 다음과 같다. 제 Ⅱ장은 관련 연구를 다루었으며 제 Ⅲ 장에서는 제안 추천시스템 설명하며 제 Ⅳ장에서는 실험 및 성능 평가를 실행하며 마지막으로, 제 Ⅴ장에서는 본 논문의 결론과 향후 연구에 대하여 기술한다.

Ⅱ. 관련연구

2.1 RFM기법

RFM에서 R값은 최근 구매일자를 의미하며, F값은 일정기간 동G안의 총 구매횟수를 그리고 M값은 일정기간 동안의 총 구매금액을 의미한다. RFM분석을 이용하여 고객 데이터베이스는 세분화될 수 있다. 각 요소마다 5개의 세분화 세그먼트로 나누어지게 되어 전체 고객데이터베이스는 결국 5 × 5 × 5 = 125 개의 세그먼트로 분할되어 진다. 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[1]. 다음은 RFM점수를 산출식으로 나타낸 것이다.

RFM점수 = (A*R + B*F + C*M) * 20 식1

이렇게 산출된 RFM점수는 고객별, 아이템별 RFM점수를 구하여 데이터를 세분화하여 각종 분석에 활용할 수 있다. RFM점수에서의 가중치(A, B, C)는 그 회사의 경영 상태이나 경영 전략에 맞추어 변경이 가능하다. RFM이 적용된 시스템의 이점은 여러 가지 정보를 손쉽게 알 수 있다는 것이다. 예를 들면, 고객 정보의 고객 분류코드와 RFM점수를 비교하여 현재 아이템의 인기도나, 선호도, 관심 및 추천 아이템 등을 쉽게 파악할 수 있다.

2.2 연관규칙(Association Rule)

연관규칙 탐사는 대규모로 축적되어 있는 트랜잭션 데이터베이스를 바탕으로 지지도(support)와 신뢰도(confidence)를 이용하여 연관성이 강한 항목들을 찾아내는 것으로 정의할 수 있다[2]. 연관규칙을 찾기 위하여 일반적으로 지지도, 신뢰도, 향상도(lift)라는 척도를 사용한다. 지지도는 생성된 연관규칙이 전체 항목에서 차지하는 비율을 뜻한다. 전체 거래 중 X와 Y를 포함하는 거래의 정도를 나타내는 식이다.

$$SUPP(R) = \frac{P(X\ U\ Y)}{T} \quad \text{식2}$$

신뢰도는 X를 포함하는 거래 중에서 Y가 포함된 거래의 정도를 의미하며 연관규칙의 강도를 의미한다.

$$CONF(R) = P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} \quad \text{식3}$$

향상도는 규칙을 모를 때에 비하여 규칙을 알 때에 얼마나 향상되는가를 나타내고 있다.

$$LIFT(R) = \frac{P(Y|X)}{P(Y)} \quad \text{식4}$$

지지도는 연관규칙을 구성하는 항목집합을 포함하는 트랜잭션이 전체 트랜잭션의 몇 퍼센트나 되는지를 나타내는 것이고, 신뢰도는 규칙의 성립 정도를 나타낸다. 신뢰도가 규칙의 강도를 나타낸다면 지지도는 이 규칙이 전체 데이터베이스에서 가지는 통계적인 중요성을 나타낸다고 할 수 있다. 대표적으로 Apriori, FP-tree, SETM, DIC, ARHP 알고리즘등이 있다. 연관규칙을 찾아주는 알고리즘 중에서 가장 먼저 개발되었고 가장 많이 사용되고 있는 것은 Apriori 알고리즘이다. 이 알고리즘은 먼저 최소 지지도 설정 값에 따라 빈도수가 높은 항목의 집합들을 찾아내고, 이들 집합들 중에서 주어진 신뢰도를 만족하는 연관규칙을 찾아낸다. Apriori 알고리즘은 이진 연관규칙에 대한 빈발항목 집합을 찾아내는데 유용한 알고리즘이지만 이는 대량의 후보집합 생성이 필요하다.

2.3 FP-tree 마이닝

FP-tree 마이닝 알고리즘은 단계 1에서 빈발항목이 아닌 항목들을 제거하고, 단계 2에서 빈발항목들로만 구성된 거래들을 모두 FP-tree 구조에 저장한 다음, 단계 3에서 FP-tree로 부터 빈발항목 집합을 유도해 낸다. FP-tree 마이닝 알고리즘은 DB의 거래 내용을 필터링한 모든 거래 내용을 메모리에 구축한 후 그 곳으로부터 빈발항목 집합을 추출한다. FP-tree는 빈발패턴에 대한 지지도를 저장하는 Prefix 트리 구조이며 상위 노드들은 높은 지지도 값을 가지는 항목들이 위치하고 낮은 지지도의 노드일수록 하위 노드에 위치하는 방식으로 트리를 구성한다. FP-tree는 각 노드의 항목들의 카운트 값을 유지하기 위해서 헤더 테이블 데이터구조를 가지며, 트리에 삽입되는 모든 패턴들은 항목들의 카운트 값을 포함한다. 본 논문에서 <그림 1>은 화장품 아이템 구매이력 데이터로부터의 모든 빈발한 패턴 탐사를 위한 Prefix 기반의 알고리즘인 FP-tree 구성 알고리즘을 나타낸 것이고 <그림 2>는 패턴의 트리 삽입 프로시저를 나타낸 것이다.

```
Input: (1) Expression data set D ;
       (2) minimum support Minsup.
Output: FP-tree corresponding to and satisfying Minsup.
1) Scan D once and collect the set of frequent items F
and their supports.
2) Sort F in support descending order as L, the list of
frequent items.
3) If several items have the same support, and their
names are numbers, sort the items in ascending order
of their names.
4) Create the root R of a new FP-tree and label it as
"null".
5) Create frequent-item header table with |F| entries.
Set all head of node-link pointers to null.
6) for each transaction data d ∈ D do // Read D the
second time.
7) Select only frequent items of d into a record P ;
8) Sort P in the order of L ;
9) Call insert_tree (P, cd, R) ;
10) end for
```

그림 1. FP-tree 마이닝 구성 알고리즘
Fig. 1. The algorithm for construction of FP-tree mining

```
Procedure insert_tree (P, c, R );
1) Let P = [p |P-p], where p is the first element of P,
and P-p is the
remaining list.
2) if R has a child N such that N.item_name=p then
3) N.count = N.count + 1;
4) else {
5) create a new node N;
6) N.count = 1; N.item_name = p;
8) N.parent = R; N.node-link = H (p).head;
9) H(p).head = N;
10) }
11) H(p).count = H(p).count + 1;
12) if P-p ≠ φ then
13) Call insert_tree(P-p, c, N) recursively.
```

그림 2. 패턴의 트리 삽입 프로시저
Fig. 2. The procedure of insertion for compact pattern tree

탐사된 화장품 아이템 패턴은 대용량이며 많은 중복 패턴들을 포함한다. 따라서, 유용한 아이템 패턴만을 추출하기 위해서 압축 패턴트리를 이용하여 중복 패턴들을 제거할 수 있다.

III. RFM기반 FP-tree 마이닝을 이용한 개인화 추천시스템

이 장에서는 빈발패턴(FP)의 문제 정의와 함께, 빈발한 패턴 탐사방법을 화장품 아이템을 예를 들어 설명한다. 추천받는 대상이 되는 특정 로그인 사용자와 고객 분류코드가 동일하고 가장 많은 분포를 이루어진 아이템 RFM점수대의 군집화(Clustering)된 구매이력 데이터를 이용한다. 또한 탐사된 패턴들 사이의 유용성 측정을 위한 연관 관계 척도인 지지도, 신뢰도, 향상도를 측정하여 추천에 활용한다. 향상도의 특정 임계값(0.5) 이상의 데이터를 이용하여 FP-tree 마이닝을 이용한 연관규칙의 생성과 사용자 기반의 최소지지도를 만족하는 모든 빈발패턴 탐사방법에 적용한다.

[정의 1] 아이템 패턴(item pattern): 화장품을 하나의 항목(item)으로 가정하고 화장품의 패턴을 나타내는 항목집합(itemset)을 P={i_1, i_2, ...,i_n}라 할 경우, 1≤j≤n인 i_j은 하나의 화장품을 표현한다.

[정의 2] 부분 패턴(sub pattern), 상위 패턴(super pattern): 패턴, P= P={i_1, i_2, ...,i_n}가 조건, i'_{k1}, i_2=i'_{k2}, i_n=i'_{kn}을 만족하는 $k_1 < k_2 ... < k_n$인 정수들이 존재한다면 P는 다른 패턴 P′={i'_1,i'_2,...,i'_m}의 부분 패턴이라고 하며 반대로, P를′상위 패턴이라고 한다.

[정의 3] 빈발패턴(FPs: Frequent Patterns): 빈발패턴, FPs란 임계값인 최소지지도(Minsup)을 만족하는 트랜잭션의 부분 패턴이다. 화장품 빈발데이터로부터의 빈발화장품 아이템 패턴 탐사는 사용자가 미리 지정한 최소지지도를 만족하는 모든 빈발한 화장품들의 집합을 탐사하는 문제이다.

3.1 FP-tree 마이닝을 이용한 연관규칙 탐사

이번 절에서는 FP-tree 마이닝을 이용한 연관규칙의 생성과 사용자 기반의 최소지지도를 만족하는 모든 빈발패턴 탐사방법을 기술한다. 본 논문에서는 사용한 후보집합의 발생없이 빈발항목을 생성하는 빈발패턴트리(FP-tree : Frequent Pattern tree)를 사용한다. 다음<표1>은 빈발패턴트리를 만들기 위한 아이템 트랜잭션을 나타낸 것이다.

표1. 아이템 트랜잭션과 내림차순 정렬 데이터
Table 1. Items in transaction and Ordered frequent items by descending

Items in transaction	
TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Ordered frequent items	
TID	List of item_IDs
T100	I2, I1, I5
T200	I2, I4
T300	I2, I3
T400	I2, I1, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I2, I1, I3, I5
T900	I2, I1, I3

첫 번째로 데이터베이스를 스캔하여 빈발1-항목집합과 발생빈도를 찾아내는 것은 Apriori방법과 동일하다. 최소지지도(Minimum Support)를 2라고 가정하고 빈발항목집합은 지지도(Support)를 기준으로 하여 내림차순으로 정렬한다. 그 결과 집합 또는 리스트를 L로 표시하며 L=[I2:7, I1:6, I3:6, I4:2, I5:2]와 같다. FP-tree는 우선, “null”로 표시된 트리의 루트(root)를 생성하고 데이터베이스 D를 스캔한다. 각각의 아이템은 지지도에 따라서 내림차순 정렬의 트랜잭션마다 노드의 가지(branch)가 생성된다. 이때 트랜잭션 아이템은 먼저 생성된 노드가 부모 노드가 되고 그 다음 생성된 노드가 자식 노드가 되면서 트리를 구성하게 된다. 하나의 트랜잭션에 가지가 추가되는 경우에는 공통 노드를 하나씩 증가시키고 다음 순서의 트랜잭션에 노드를 생성한 후 링크를 연결해 준다. <그림 3>은 모든 트랜잭션을 스캔한 결과로 얻어진 트리이다. 생성된 FP-tree는 상향식으로 연관규칙을 생성한다. 따라서 데이터베이스에서 빈발패턴을 마이닝하는 문제는 빈발패턴트리를 마이닝하는 문제로 변환된다.

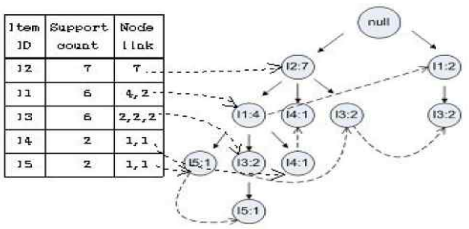

그림 3. 빈발패턴을 저장하는 빈발패턴트리
Fig. 3. FP-tree saving Frequent Pattern

FP-tree 마이닝은 초기의 접미 패턴(suffix pattern)에서 시작하여 조건부 패턴베이스(conditional pattern base)를 생성하고 조건부 빈발패턴트리를 생성한다. 그리고

이 트리에 대해서 재귀적으로 마이닝을 수행한다. 패턴의 증가는 접미부 패턴과 조건부 빈발패턴트리로부터 생성된 빈발패턴을 결합함으로써 얻어진다. 다음 <표2 >는 FP-tree 마이닝을 요약하여 나타낸 것이다.

표2. 조건부 패턴베이스를 생성하는 FP-tree 마이닝
Table 2. FP-tree mining generated conditional pattern base

아이템	조건부 패턴집합	조건부 빈발패턴트리	빈발패턴 아이템
I5	{(I2 I1:1), (I2 I1 I3:1)}	{I2:2, I1:2}	{I2 I5:2}, {I1 I5:2}, {I2 I1 I5:2}
I4	{(I2 I1:1), (I2:1)}	{I2:2}	{I2 I4:2}
I3	{(I2 I1:2), (I2:2), (I1:2)}	{I2:4, I1:2}, {I1:2}	{I2 I3:4},{I1 I3:4}, {I2 I1 I3:2}
I1	{(I2:4)}	{I2:4}	{I2 I1:4}

다음과 같은 방법으로 조건부 패턴집합을 생성하는 빈발패턴트리를 마이닝한 것이다. I5는 두 개의 노드에서 발견된다. I5를 접미부로 하면 두 개의 접두부 경로는 (I2 I1 I5:1), (I2 I1 I3 I5:1)이며, 접두부 경로들은 조건부 패턴 베이스를 구성한다. I5의 조건부 빈발패턴트리는 단일 경로 (I2:2, I1:2)가 되는데 I3이 포함되지 못하는 이유는 I3은 지지도가 1로서 최소 지지도보다 작기 때문이다. 이 단일 경로는 빈발패턴의 모든 조합 {I2 I5:2}, {I1 I5:2}, {I2 I1 I5:2}를 생성하게 된다. 이느 우수한 선택율을 제공하는 최소 빈발항목을 접미부로 사용함으로써 기존의 문제점을 해결한다. FP-tree는 빈발패턴에 대한 지지도를 저장하는 Prefix 트리 구조이며 상위 노드들은 높은 지지도 값을 가지는 항목들이 위치하고 낮은 지지도의 노드일수록 하위 노드에 위치하는 방식으로 트리를 구성한다.

3.2 FP-tree 마이닝을 이용한 연관규칙 생성

기존의 연관규칙을 이용한 추천시스템은 대규모 데이터베이스의 거래 데이터 처리에서 빈발항목 집합을 찾아내는데 매번 계속적으로 대량의 데이터를 스캔해야 하므로 속도가 느릴 뿐 아니라 확장성 문제와 정확도 문제가 있다. 구매가능성이 높은 아이템을 추출하기 위해서 아이템 속성 반영이 가능한 RFM기반의 FP-tree 마이닝을 이용한다. 연관규칙 생성 시 후보집합을 생성하지 않으면서 연관규칙을 생성하기 때문에 반복적인 데이터베이스 접근을 필요로 하지 않는다. 또한 연관규칙을 생성하여 룰기반으로 제공하므로 실시간 추천에서 요구되는 즉시성을 마련할 수 있다. <표3>은 제안 시스템에서 RFM기반 FP-tree 마이닝을 이용한 연관규칙 생성시 지지도 10%이상이고 신뢰도가 80%, 향상도 5%이상의 경우의 화장품 아이템 연관규칙을 적용한 예이다. <표4>는 향상도 구간별 연관규칙수를 나타낸 것이고 <그림 4>는 향상도별 연관규칙수 현황을 나타낸 것이다.

표3. FP-tree 마이닝의 연관규칙
Table 3. FP-tree mining generated association rule

아이템1 → 아이템2	지지수	지지도	신뢰도	향상도
AAA21 → ABA06	50	0.3	100	2
ABA06 → AAA21	50	0.3	100	2
ABD17 → CAC05	30	0.18	100	3.33
ACB12 → ACC12	156	0.92	98.73	0.61
ACC12 → ACB12	156	0.92	96.3	0.61
BAA20 → BAB06	20	0.12	100	5
BAB05 → BAC02	99	0.59	100	1.01
BAB06 → BAA20	20	0.12	100	5
BAC02 → BAB05	99	0.59	100	1.01
CAC01 → CAD20	24	0.14	100	4.17
CAC02 → CAC13	30	0.18	100	0.95
CAC05 → ABD17	30	0.18	100	3.33
CAD20 → CAC01	24	0.14	100	4.17
CAE07 → CAE34	35	0.21	87.5	1.23
CAE26 → CAE34	30	0.18	85.71	1.21

표4. FP-tree 마이닝의 향상도별 연관규칙수
Table 4. FP-tree mining generated number of association rule by each lift rate

향상도	rule수
0.5 ~ 0.9	28
1.0 ~ 1.99	27
1.5 ~ 1.99	5
2.0 ~ 2.99	14
3.0 ~ 3.99	3
4.0 ~ 4.99	2
5.0 ~ 5.99	6
6.0 ~ 6.99	23

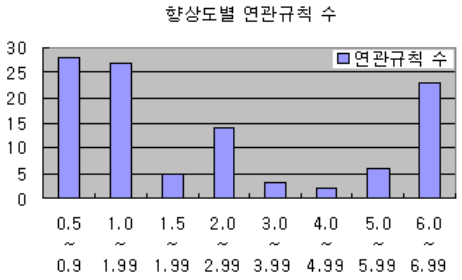

그림 4. 향상도별 연관규칙수 현황
Fig. 4. The result for number of association rule by each lift rate

3.3 시스템 구성

본 시스템에서는 고객정보와 구매이력정보 바탕의 묵시적인 방법으로 고객의 Social data로 구성된 인구통계학적

변수(나이, 성별, 직업, 고객성향 등)가 적용된 고객 분류코드에 맞게 개인화된 추천시스템을 구성한다. 유무선 인터넷 쇼핑몰 환경하에서 제안 시스템이 구동되기 위해서 분석 에이전트, 추천 에이전트, 학습 에이전트, 마이닝 에이전트로 그 기능을 나누어 서브 시스템이 구동된다. 다음은 각 에이전트 모듈의 주요 기능이다. 분석 에이전트는 회원 가입을 통하여 고객집단을 분류 가능하도록 고객 분류코드로 부여하여 고객정보를 생성 및 관리한다. 고객정보에서 로그인 사용자의 고객분류코드를 인지한다. 또한 기존 시스템[3,4,5]과 다르게 고객 분류코드를 이용하여 군집화된 군집(Cluster)을 탐색하고 인지한다. 추천 에이전트는 선호도가 높은 아이템 카테고리내에서 구매 가능성이 높은 아이템을 Top-N의 추천 아이템 목록을 생성한다. 이때 로그인 사용자의 구매 이력정보와 체크하여 중복 추천되지 않도록 한다. 학습에이전트는 고객 분류코드를 갖는 고객정보 처리 및 관리하며 아이템 선호도 계산 작업을 수행한다. 구매가 이루어진 고객의 구매데이터에 대한 아이템 RFM점수가 DB처리하여 실시간으로 적용될 수 있도록 한다. 마이닝 에이전트는 연관규칙 기반 추천을 위한 모듈이다. 아이템간의 연관규칙을 찾아, 그 결과로 추천 아이템을 생성하는 모듈이다. 생성된 연관규칙의 지지도, 신뢰도, 향상도를 산출하여 추천에 사용한다. 유무선 웹환경에서 RFM기반 아이템 카테고리 선호도를 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 아이폰의 사파리 및 안드로이드 기반의 구글 크롬 웹브라우저에서도 이용 가능하도록 웹표준을 준수한다. 다음 <그림 5>는 추천시스템의 시스템 구성도를 나타낸 것이다.

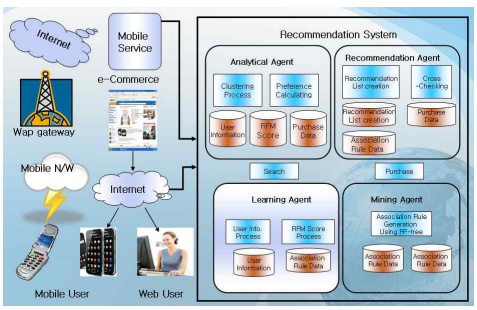

그림 5. 추천시스템의 전체 시스템 구성도
Fig. 5. system configuration for recommend system

Ⅳ. 실험 및 성능 평가

4.1 실험 환경

구현 및 실험 환경은 윈도우 운영체체하에서 해당 평가하기 위해서 다음과 같은 웹서버 환경을 사용하였다.

- 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/
- MySQL 4.0.24
- http://www.mysql.com/products/mysql/
- MySQL Connector/J 3.0
- http://www.mysql.com/products/connector/j/

RFM기반 아이템 카테고리 선호도를 이용한 개인화 추천시스템의 성능을 평가하기 위해서 제안 시스템과 기존 시스템의 성능평가를 실험을 통해서 알아본다.

4.2 실험 데이터 구성

RFM기반 FP-tree 마이닝을 이용한 개인화 추천시스템은 윈도우 XP 환경에서 인터넷 화장품 쇼핑몰을 위한 데이터베이스가 구축되었다. 시스템에 대한 평가를 위한 실험데이터의 구성은 쇼핑몰을 이용해 본 경험이 있는 고객 319명의 고객 정보와 , 그리고 현재 화장품을 전문적으로 판매하는 인터넷 화장품 쇼핑몰인 P사의 아이템 분류에서 사용하는 화장품 아이템 580개를 대상으로 그들의 추천 1600건의 구매 데이터를 이용하여 2009년 2월 부터 2010 2월까지의 12개월간의 과거의 구매 데이터를 학습 셋으로 사용하였고, 2010년 3월 부터 2010년 5월까지 3개월간의 미래 구매 데이터를 테스트 셋으로 사용하였다.

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 있

는 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 기존시스템을 실험하였다. 우선 첫 번째, 실험으로 MAE에 의해 예측의 성능을 평가하였다. 추천시스템의 예측 값의 정확성을 평가하기 위해 MAE(Mean Absolute Error)를 사용하였고 식5와 같이 산출하였다[6].

$$MAE = \frac{\sum_{i=1}^{N} |\epsilon_i|}{N} \qquad 식5$$

N은 총 예측 회수를 나타내고, ε_i는 예측 값과 실제 값의 오차를 나타내며 i는 각 예측 단계를 나타낸다. <표5>는 식2를 이용하여 예측값의 정확성 평가를 수행한 결과이다.

표5. 제안 및 기존 시스템의 MAE에 의한 성능평가
Table5. The result for table of MAE by comparing proposal system with existing system

	P_count	Proposal	Existing
MAE	50	0.52	0.65
	100	0.24	0.32
	300	0.07	0.08
	500	0.04	0.06

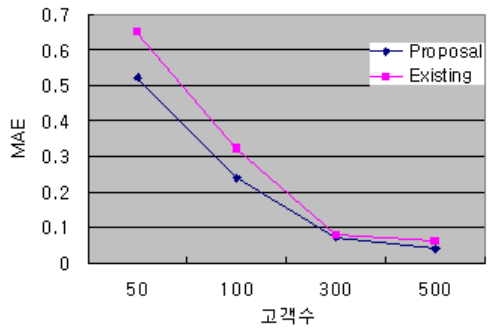

그림 6. 제안 및 기존 시스템의 MAE에 의한 성능평가
Fig. 6. The result for the graph of MAE by comparing proposal system with existing system

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

$$\text{정확률} = \frac{\text{고객이 구매한 아이템 갯수}}{\text{추천아이템 갯수}} \qquad 식6$$

$$\text{재현율} = \frac{\text{추천시스템의 추천아이템 중 고객이 구매한 아이템 갯수}}{\text{고객이 구매한 아이템 갯수}} \qquad 식7$$

$$\text{F-Measure} = \frac{2(\text{정확률} * \text{재현율})}{\text{추천 아이템 갯수}} \qquad 식8$$

추천받는 대상이 되는 특정 로그인 고객과 고객 분류코드가 동일하고 가장 많은 분포를 이루어진 아이템점수대의 군집 데이터를 이용한다. <표6>은 군집별 추천의 정확도와 재현율을 분석한 결과를 나타낸 것이다.

표6. 군집별 추천의 정확도와 재현율 결과
Table6. Result of Precision and Recall for Recommendation by clusters

군집	제안 시스템			기존 시스템		
	정확률	재현율	F-measure	정확률	재현율	F-measure
C1	62.88	91.44	70.01	56.98	50.89	50.21
C2	46.20	55.70	45.90	48.79	31.32	35.64
C3	43.20	52.53	45.76	49.36	29.54	35.06
C4	40.99	23.93	30.05	44.26	21.81	27.65
C5	56.06	38.37	43.77	52.49	34.98	39.75
C6	53.94	27.27	34.90	47.41	26.81	32.26
C7	45.07	37.23	38.29	43.60	36.60	37.82
C8	64.08	28.45	37.19	46.68	25.19	30.28
C9	60.00	20.69	30.77	46.53	18.32	25.10
C10	73.85	62.50	64.42	67.23	55.34	57.10

다음은 <표6>의 기존 시스템과 제안 시스템을 비교하기 위한 SQL을 이용한 실험 데이터 결과의 일부분를 나타낸 것이다.

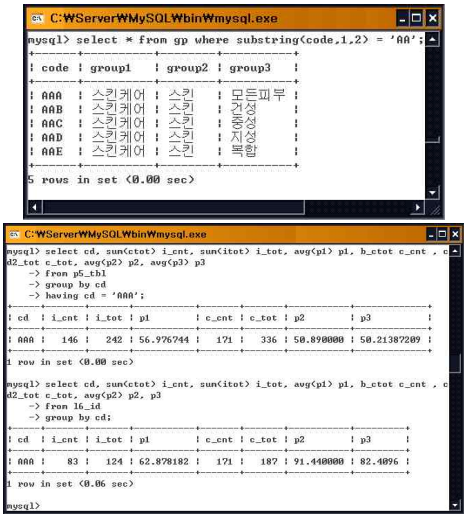

그림 7. SQL을 이용한 실험데이터 결과
Fig. 7. The result of experimental data using SQL

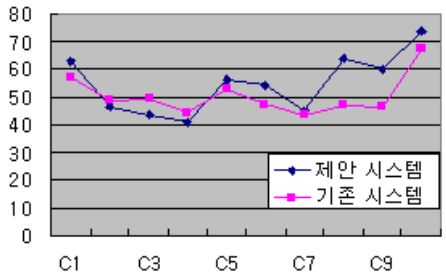

그림 8. 정확률에 따른 추천평가 결과
Fig. 8. The result of recommending ratio for recommendation each cluster by precision

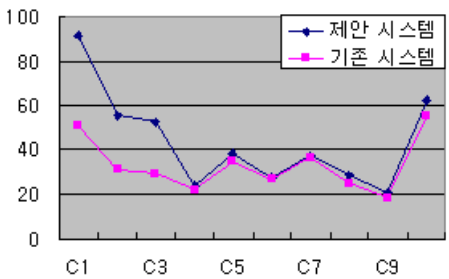

그림 9. 재현율에 따른 추천평가 결과
Fig. 9. The result of recommend ing ratio for recommendation each cluster by recall

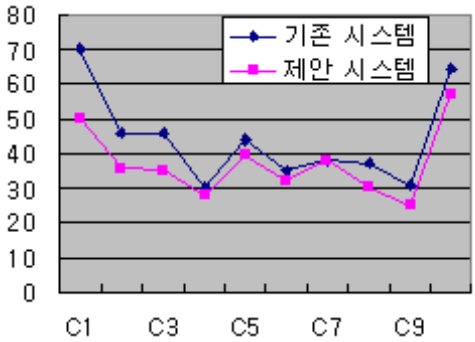

그림 10. F-measure에 따른 추천평가결과
Fig 10. The result of recommending ratio for recommendation each cluster by F-measure

<그림8>. <그림9>. <그림10>은 <표6>의 결과를 바탕으로 군집별 정확도와 재현율 그리고 F-measure의 성능 평가이다. 제안시스템은 기존시스템보다 4.3% 높은 정확도와 10.73% 높은 재현율, 그리고 7.02% 높은 F-measure의 결과를 나타내었다. 실험 결과적으로 기존 시스템 보다 높은 성능이 산출되었음이 입증되었다. <그림10>은 웹상의 RFM 기반 FP-tree 마이닝을 이용한 개인화 추천시스템에서 추천된 화장품 사이트를 보여주고 있으며 스마트폰으로도 이용가능하다.

그림 11. 화장품 아이템 추천 결과
Fig 11. The result of recommending items of cosmetics

제안 시스템은 구매 가능성이 높은 아이템을 추출하기 위해서 RFM기반의 FP-tree 마이닝을 이용하였다. 연관규칙 생성 시 후보집합을 생성하지 않으면서 연관규칙을 생성하기 때문에 반복적인 데이터베이스 접근을 필요로 하지 않는다. 또한 연관규칙을 생성하여 룰기반으로 제공하므로 실시간 추천에서 요구되는 즉시성이 확보되었다.

V. 결론 및 향후 과제

유비쿼터스 네트워크 환경에서 스마트 폰과 같은 지능형 이동 단말기로 유무 인터넷을 즐기기 위한 수요가 더욱 증대되고 있다. 이러한 시점에서 지능형 추천시스템의 구성은 기업의 비즈니스 전략이 되어 가고 있다. 본 본문에서는 RFM기반의 FP-tree 마이닝을 이용한 개인화 추천시스템을 제안하였다. FP-tree는 빈발항목 집합으로부터 최소지지도와 최소신뢰도를 만족하면서 강한 아이템간의 연관규칙을 생성한다. 이러한 연관규칙기반에 향상도와 고객 분류코드를 적용하여 정확도를 향상시키고 추천시스템의 확장성 문제를 해결하였다. FP-tree는 후보집합을 생성하지 않아 Apriori 알고리즘보다 우수하고 최소 빈발항목을 접미부로 사용함으로써 속도와 탐색비용을 감소시킨다. 또한 정확도 측면에서도 더욱 향상된 추천시스템을 구현할 수 있다. 성능평가를 위해 현업에서 사용하는 인터넷 화장품 아이템 쇼핑몰의 데이터를 기반으로 데이터 셋을 구성하여 기존의 방법과 비교실험을 통해 성능을 평가하여 효용성과 타당성을 입증하였다. m-Commerce가 증가되는 시점에 시간과 장소에 제약을 받지 않는 유무선 웹을 이용한 추천시스템에 개선된 개인화된 추천시스템을 제안함으로써 인터넷 쇼핑몰 추천시스템의 프레임워크를 제시함에 큰 의미가 있다. 아울러 향후 연구는 유비쿼터스 컴퓨팅 환경하에 구매 가능성이 높은 아이템을 추천하기 위해 묵시적인 방법으로 RFM기반의 u-커머스 추천시스템에 대한 고찰과 연구가 필요할 것으로 생각된다.

참고문헌

[1] Chan Wook Park. “DataBase Marketing-Strengthen for Competited Enterprise using Customer's Information ”, YeonAm Press, 1996.

[2] Agrawal, R. and Srikant, R, “Fast Algorithms for Mining Association Rules in Large Databases,” In Proceedings of the VLDB, Santiago, Chile, pp.487-499, September, 1994.

[3] 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

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

[5] 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

[6] Jonathan L. Herlocker, Joseph A. Kosran, Al Borc hers, and John Riedl, “An Algorithm Framework for Performing Collaborative Filtering", Proceedings of the 1999 Conference on Rechearch and Developm ent in Information Retrival, 1999

저 자 소 개

조 영 성

1989 : 연세대학교 전산학과(공학석사)

2008 : 충북대학교 전산학과(공학박사)

1982~2009 : 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 산업체 겸임/교수,

관심분야 : 시공간 데이터베이스, 유비쿼터스 컴퓨팅 및 GIS, 데이터 마이닝, 기계학습, 웹서비스, ebXML

Email : youngscho@empas.com

류 근 호

1976 : 숭실대학교 전산학과(이학사)

1980 : 연세대학교 공학대학원 전산전공(공학석사)

1988 : 연세대학교 대학원 전산전공 (공학박사)

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

1989~1991 : 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, 데이터마이닝, 데이터베이스, 보안, 바이오 인포메틱스

Email : khryu@dblab.chungbuk.ac.kr