Journal of Digital Contents Society Vol. 14 No. 4 Dec. 2013(pp. 419-428)
http://dx.doi.org/10.9728/dcs.2013.14.4.419

마이닝과 FRAT기반 가중치 선호도 군집을 이용한 추천 기법에 관한 연구

박화범*, 조영성**, 고형화***

요 약

유비쿼터스 컴퓨팅 환경의 전자상거래에서 실시간성과 추천의 정확도를 높이는 연구가 활발히 진행되고 있다. 대부분의 기존 추천기법들은 프로파일 방식의 문제로 고객의 관심도나 고객성향을 분석하기에는 많은 어려움과 비용의 문제가 있으며 고객은 여전히 만족하지 못하고 있다. 이는 구성되어있는 데이터베이스들의 문제가 아니라 기존 자료를 분석하기 위한 평가 자료인 신규로 프로파일을 생성하거나 다양한 프로파일을 생성하는데 문제가 있다. 또한 기존 추천기법에서는 다양한 특성을 가진 각 사용자 계층별로 차별화된 개인화 추천이 어렵다. 따라서 이 논문에서 기존의 평가 자료 방식과 다르게 구매로 인해 발생되어진 자료를 기반으로 사용자에게 번거로운 질의 응답 과정이 없이 묵시적인 방법을 이용하였다. 다양한 개인화 성향과 정확한 고객성향의 내용 분석이 가능한 FRAT 기법을 적용하였다.

키워드 : FRAT 기법, k-mean알고리즘, 협력필터링, 추천시스템

A Study on Recommendation Technique Using Mining and Clustering of Weighted Preference based on FRAT

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

Abstract

Real-time accessibility and agility are required in u-commerce under ubiquitous computing environment. Most of the existing recommendation techniques adopt the method of evaluation based on personal profile, which has been identified with difficulties in accurately analyzing the customers' level of interest and tendencies, as well as the problems of cost, consequently leaving customers unsatisfied. Researches have been conducted to improve the accuracy of information such as the level of interest and tendencies of the customers. However, the problem lies not in the preconstructed database, but in generating new and diverse profiles that are used for the evaluation of the existing data. Also it is difficult to use the unique recommendation method with hierarchy of each customer who has various characteristics in the existing recommendation techniques. Accordingly, this dissertation used the implicit method without onerous question and answer to the users based on the data from purchasing, unlike the other evaluation techniques. We applied FRAT technique which can analyze the tendency of the various personalization and the exact customer.

Keywords : FRAT, K-means Algorithm, Collaborative Filtering, Recommender System

※교신저자(CorrespondingAuthor): Wha-Beum Park
접수일:2013년 10월 26일, 수정일:2013년 11월 26일
완료일:2013년 12월 06일
* 광운대학교 전자통신공학과
e-mail:pwb430@nate.com
** 동양대학교 컴퓨터학과
*** 광운대학교 전자통신공학과

1. 서 론

최근 유비쿼터스 컴퓨팅 환경하에서 스마트폰과 스마트패드 같은 지능형 단말기 수요가 더욱 증대되고, 이로 인해 많은 빅데이터 속에서 정보를 찾아내는 기술이 부각되고 있다. 전자상거래에서 아이템 추천에 많이 사용되는 기존의 명시적인 방법의 협업 필터링(Collaborative Filtering)은 데이터의 널(Null) 값으로 인한 데이터 희박성(Sparsity) 문제, 아이템의 속성을 반영하지 못하는 문제, 사용자 수가 증가하거나 아이템 수가 많아져 대용량 데이터에서 유사한 사용자나 아이템 그룹을 찾기 위해서는 많은 시스템 리소스가 필요하며 추천에 소요되는 시간이 길어지는 확장성의 문제가 여전히 존재하며, 실행 속도의 저하, 실시간성(Real Time Accessibility) 및 민첩성(Agility)등에 문제가 있다. 본 연구에서는 추천을 위해서 시간의 변화를 반영한 구매이력 정보에 FRAT기반 세분화 기법을 적용하여 신속한 처리를 위한 고객점수와 아이템 점수에 대한 기반 정보를 구축 및 활용과 추천 아이템의 정확도와 신속한 처리를 위해서 마이닝과 FRAT기반 가중치 선호도 군집을 이용한 추천 기법에 대한 것이 연구의 목적이다.

추천 대상 고객인 로그인 사용자와 고객 성향과 고객점수가 같은 사용자들의 구매이력 데이터를 추출하여 구매이력 데이터에서 로그인 사용자의 개인별 선호도 분석을 통한 군집화 기반의 추천 기법에 대한 연구와 마이닝을 통한 연관 규칙 탐사에 대한 연구가 필요하다.

마이닝과 FRAT기반 가중치 선호도 군집을 이용한 추천 기법에 대한 연구를 위해서 추천 아이템의 정확도와 실시간 추천을 위한 신속한 처리 방안과 추천을 통한 최종 아이템에 대한 구매 결정시 구매 아이템에 대한 사전 및 사후 선호도 확률 계산을 통해 추천의 정확도를 향상시키는 방안을 적용한다. 최종 아이템 구매결정시 구매에 대한 사전 및 사후 선호도 계산을 위해 베이시안 네트워크를 이용한 선호도 기반의 학습에 대한 연구도 진행된다. 본 연구에서 제안한 추천 기법이 적용된 추천시스템에 대한 FRAT기반 적용분석을 통해서 타당성을 제시하며 시스템 평가를 통해 그 효용성을 입증한다.

2. 관련 연구

이 장에서는 이 논문과 관련된 선행연구를 검토하고 선행연구의 문제점을 기술한다. 고객 및 아이템에 대한 세분화를 통한 속성 반영을 위한 RFM 기법과 개선된 세분화 모델인 FRAT 기법에 대하여 기술한다. 다음으로는 마이닝을 통한 연관 규칙 생성과 아이템 선호도를 위한 군집화 기법에 대하여 기술한다.

2.1 선행 연구

추천 기법이 적용된 시스템에 대한 최근의 국내외 연구동향을 보면 협업 필터링 추천 시스템의 성능 향상을 위해 추천의 정확도를 향상시키기 위한 연구가 계속적으로 진행되고 있으며, 이용 패턴 정보 기반 멀티미디어 콘텐츠 추천 시스템 연구에서도 추천 정확도 향상을 위해 해결해야 할 핵심적인 문제로는 사용자 정보 취득·분석의 부정확성 문제와 사용자 정보의 희소성 문제를 해결할 수 있는 새로운 멀티미디어 콘텐츠 추천 시스템을 제안하였다[1]. 모든 고객의 거래데이터를 이용하여 추천 대상 고객과 유사한 고객, 즉 네이버를 연결하는 네이버 네트워크를 형성하는 기존의 CF와 대조되는 방법으로, 개별고객에게 유효한 네이버 탐색범위를 구조적으로 제한하여 추천정확성의 손실없이 시스템 속도를 획기적으로 높였다. 다음으로, 로컬 네이버 탐색 방법의 유망한 애플리케이션의 하나로 유비쿼터스 환경에서의 추천프로세스를 제안하였고 추천프로세스를 지원하는 고객네트워크인 Buying-net을 모델링하였다. 또한 멀티미디어 컨텐츠를 위한 이용빈도 기반 하이브리드 추천 시스템에 대한 연구를 들 수 있다[2].

군집화의 k-mean 클러스터링과 마이닝의 순차 패턴 기법을 이용한 VLDB(very large database) 기반의 상품 추천 시스템을 설계 및 구현에 대한 연구를 들 수 있다. 이 연구는 사용자의 정보를 일괄처리하고 다양한 카테고리를 계층적으로 정의하며, 탐색엔진에 순차 패턴 마이닝 기법을 이용하였다. 예측 모델을 만들기 위하여 사용자의 로그 데이터 중에서 카테고리에 대한 사용자의 선호도를 추출하여 이용하고 있다[3].

최근 추천 기법에 대한 주요 국내 연구로는 마이닝 기반 감성 분석을 이용한 디자인 추천 시스템[4]에서는 디자인 추천을 위해 연관 이미지 필터링과 협업 필터링을 결합한 연관 이미지 기반 협업 필터링에 관한 제안이 있다. 스마트 홈 환경에서 신뢰할 수 있는 사용자 인증 및 추천 시스템[5]에 관한 연구에서는 스마트 홈에서 소셜 네트워크 등을 통해 실시간으로 공유되는 컨텐츠의 양이 많기 때문에 개인에게 알맞은 추천 시스템이 필요하다. 협업 필터링 기법의 또 다른 문제점으로 알려져 있는 고객의 취향 변화를 반영해야 하는 문제를 개선하기 위하여, 검증된 웹 사용 패턴을 이용하는 동적 사용자 프로파일 생성 기법[6], 모바일 웹 마이닝 및 선호도 순위 기반의 개인화된 모바일 음악 추천 시스템 등에 대한 연구가 진행되고 있다[7].

2.2 RFM 기법

2.2.1 RFM 개요

RFM(Recency, Frequency, Monetary)분석은 구매가능성이 높은 우수고객에게 집중적으로 마케팅 전략을 실행하고 수익을 창출하기 위한 모형이다.

세개의 요소로 구성되어진 RFM은 3요소 최근성, 구매 빈도성, 구매액 각각에 점수를 부여하는 방식으로써 이 RFM을 이용하여 고객 데이터베이스는 세분화될 수 있다. 각 요소마다 5개의 세분화 세그먼트로 나뉘어 전체 고객데이터베이스는 결국 5 × 5 × 5 = 125 개의 세그먼트로 분할되어 진다. RFM은 가치 있는 고객을 추출해 내어 이를 기준으로 고객을 분류할 수 있는 매우 간단하면서도 유용하게 사용될 수 있는 방법으로 알려져 있다 [8].

2.2.2 RFM 점수화

데이터베이스 마케팅에서 고객의 가치를 평가하는데 가장 널리 사용되는 대표적인 것이 RFM 점수(Scoring)화이다. RFM 점수화는 RFM Matrix를 더욱 정교화 할 수 있는 방법 중 하나이다. 각각의 요소를 기준으로 고객들에 대해 점수를 부여하고 각각의 요소 기준의 가중치를 부여하여 RFM점수를 계산한다. 이 RFM점수를 고객 가치를 평가하는 지표로 삼는 방식이 RFM에 의한 고객 점수 부여 방법이라고 할 수 있다 [9].

RFM은 최근성(Recency), 빈도성(Frequency), 총구매액(Monetary)등 고객의 수익기여도를 나타내는 세 가지 지표들의 선형결합으로 구한 점수(Score)로 표현할 수 있는데 이를 구하는 방법으로 가장 일반적인 RFM 모형은 (식) 2.1과 같다.

$$RFM = A*\mathbf{R} + B*\mathbf{F} + C*\mathbf{M} \quad (2.1)$$

R(recency), F(frequency)
M(monetary:총구매액)

여기서 A, B, C는 각 요인에 대한 가중치가 되고, RFM은 각 요인들과 가중치의 선형결합에 의한 점수가 되는 것이다.

$$RFM점수 = w*(w1*R + w2*F + w3*M) \quad (2.2)$$

[여기서 w: 가중치, w1, w2, w3는 속성별 가중치, R, F, M: 각 항목의 점수]

위 RFM점수를 산출하는 식이 도출되면 이를 각 고객에게 적용하여 고객들에게 점수를 부여한다. 현재 구매일자를 비교하여, “고객이 얼마나 최근에 구입했는가?” 최근성에 대한 점수를 부여한다. 1년을 기준으로, “고객이 얼마나 자주 우리 상품을 구입했는가?” 빈도성에 대한 점수를 부여한다. 구매한 아이템에 총 금액으로, “고객의 총 구매액은 얼마인가?” 총 구매액에 대한 점수를 부여한다. R, F, M의 값은 각각 최고 5점, 최저 0점이다. 이 점수들의 합계는 최고 점수는 100점, 최하 점수는 0점이다.

2.3 FRAT 기법

고객 성향과 행동 데이타를 통한 전략 수립을 위해 고객세분화에서 보다 개선된 연구가 진행되고 있다[10]. 보다 개선된 FRAT(Frequency, Recency, Amount and Type of merchandise/Service) 세분화 모델에 대한 연구가 진행되고 있다. 세분화 분석 요소인 RFM 분석의 확장모델로 아이템의 종류(T)의 한 가지 요소가 더 추가된 FRAT는 F(frequency:빈도성), R(recency:최근성), A(amount:구매력), T(type of merchan-dise/service:구입상품수)의 개선된 분석요소를 가질 수 있다. 개선된 세분화 분석요소를 반영할 때 F, R, A, T의 값은 각각 최고 5점, 최저 0점

이다. 이 점수들의 합계는 통계적 분석 방법으로 각각의 가중치(계수)를 부여한 RFM등식에서 T의 요소가 추가된 것이다.

Robert Kestnbaum은 FRAT을 적용하여 기업들에서 컨설팅한 결과 실제로 RFM방식의 세분화 방법보다 더 좋은 결과를 얻었다[11]. 고객이 최근에 어떤 아이템을 샀다는 사실이 미래 구매 행동을 예측하고 다른 아이템을 교차판매 하는데 매우 중요한 요소로 작용하기 때문에 단순히 모델의 점수계산 목적뿐만 아니라, 아이템 거래 정보도 고객선별에 필수정보로서 계속 보관되어 분석데이터로 활용될 데이터인 것이다[12].

FRAT점수의 가중치(A, B, C, D)는 경영 상태이나 경영 전략에 따라 변경이 가능하다. FRAT점수의 합계는 최고점수는 100점, 최하점수는 0점이다. FRAT점수를 위해서 사용되는 F, R, A, T 요소는 예측력이 강한 변수이다. FRAT은 구매 가능성이 높은 고객을 선정하기 위한 데이터 분석 방법이다. 본 논문에서는 고객의 구매 가능성 세분화를 기반으로 FRAT 점수 기반의 고객과 취급되는 제품을 등급화하고 분석하여 추천에 활용한다.

2.4 협업 필터링

사용자간의 유사도를 기본으로 하는 협업 추천 방식은 GroupLens이 사용자간의 유사도를 계산하기 위해 통계학 분야의 피어슨(Pearson) 상관 계수를 사용할 것을 제안하였다[13]. 기본적으로 협업 필터링 알고리즘은 다음과 같이 4가지의 단계로 구현, 평가된다. <표 1> 선호도 평가를 예측하는 행렬은 협업 필터링 개인화 기법을 이용하는 아이템 추천 과정을 나타낸 것이다. 가중치는 5점 리커드(Likert)척도를 사용하였다. 이는 사용자와 아이템으로 구성되는 행렬에서 없는 값을 예측하는 방법과 동일하다. 다음 <표 1>는 새로운 사용자에 대한 아이템 D에 대한 선호평가를 예측하는 행렬에서 사용자 u_1, 사용자 u_2, 사용자 u_3의 기존의 사용자들이 평가한 선호도를 기반으로 새로운 사용자u_a의 미 평가 아이템 D에 대한 선호도를 예측하기 위한 행렬이다.

<표 1> 선호도 평가를 예측하는 행렬

	Item A	Item B	Item C	Item D
User u_1	3	1	3	5
User u_2	1	3	1	4
User u_3		3	1	2
New User u_a	3	1		?
Step 1	To define and calculate the weighted of similarity with New users and Neighbors.			
Step 2	To determine how many neighbors with high similarity and by which criterion you will select in order to predict a preference of New user for particular item.			
Step 3	New users' preference predicts a value of preference for items which have not been input on the basis of preference for item of neighbors with similar preference.			
Step 4	To evaluate the result of cooperation filtering with preference of item which has not input a preference of new users and predicted preference by proper evaluation criterion.			

<Table 1> Matrix to predict preference rating

일반적으로 협업 필터링은 4개의 Step으로 수행된다. 협업 필터링은 사용자 기반의 협업 필터링과 아이템 기반의 협업 필터링으로 나뉜다[14].

아이템간의 유사도를 계산하기 위한 기존 연구로는 코사인 기반의 유사도, 피어슨 상관계수 기반의 유사도, 개선된 코사인 기반의 유사도가 있다[15]. 다음 (식) 2.3은 피어슨 관계 계수를 이용해 사용자 a와 사용자 k간의 유사도를 계산하는 식이다. (식) 2.3에서 j는 사용자 a와 사용자 k가 선호 평가를 매긴 아이템을 의미하고, r_{aj}와 r_{kj}는 각각 사용자 a와 사용자 k의 아이템 j에 대한 평가값이며 $\overline{r_a}$, $\overline{r_k}$ 는 각각 사용자 a와 k의 전체 정보에 대한 평가한 아이템 평균값이다. 분모는 각각 사용자 a와 사용자 k가 평가한 값의 편차이며, 사용자 a의 아이템 j 에 대한 선호도를 예측하게 된다.[16] W_{ak}는 사용자 a와 사용자 k 간의 유사도 가중치(Similarity Weight)이다.

$$W_{ak} = \frac{\sum_{i=1}^{n}(r_{aj}-\overline{r_a})\cdot(r_{kj}-\overline{r_k})}{\sigma_a \cdot \sigma_k} \qquad (2.3)$$

피어슨 관계 계수는 각각의 공통된 아이템에 대하여 현 사용자의 평가와 다른 사용자의 평가

를 비교하는 방식으로 구성되어 있다. (식) 2.4에서 n은 사용자수, P_{aj} 는 사용자 a의 아이템 j에 대한 예측값이다.

$$Paj = \overline{r_a} + \frac{\sum_{i=1}^{n} W_{ai}(r_{aj} - \overline{r_k})}{\sum_{i=1}^{n} W_{ai}} \quad (2.4)$$

(식)2.4는 (식)2.3의 사용자들의 유사도를 이용하여 사용자 a의 아이템 j에 대한 선호도를 예측(prediction)하는 식으로 아마존 같은 대형 전자상거래 시스템에서 사용되는 방식이다. GroupLens는 인터넷 뉴스를 추천하는 시스템으로 소개 된 이후로 아마존, 리바이스, CDNOW 등과 같은 사이트들에서 여러 형태로 널리 사용되고 있다.

2.5 데이터 마이닝의 연관규칙

데이터 마이닝 패턴 분석은 관련이 없는 규칙을 제거하고 흥미로운 규칙이나 패턴 발견 단계의 산출물로부터 패턴을 추출하는 것이다. 패턴 발견의 결과는 사용하기에 적절한 형식이 아니기 때문에 사용자가 쉽게 알 수 있는 형식으로 변경되어야 한다[17]. 기존 데이터 마이닝 기법 중에서 연관 규칙과 순차 패턴이 가장 많이 이용되고 있다. 대규모로 축적되어 있는 트랜잭션 데이터베이스를 바탕으로 지지도(Support)와 신뢰도(Confidence)를 이용하여 연관성이 강한 항목들을 찾아내는 것으로 정의할 수 있다[18].

연관규칙을 찾기 위하여 일반적으로 지지도, 신뢰도, 향상도(Lift)라는 척도를 사용한다. 지지도는 생성된 연관규칙이 전체 항목에서 차지하는 비율을 뜻한다. (식)2.5는 전체 거래 중 X와 Y를 포함하는 거래의 정도를 나타내는 식이다.

$$\text{SUPP(R)} = \frac{P(X \cup Y)}{T} \quad (2.5)$$

(식) 2.6의 신뢰도는 X를 포함하는 거래 중에서 Y가 포함된 거래의 정도와 연관규칙의 강도를 의미한다.

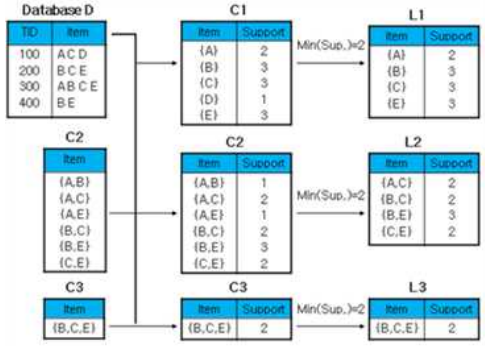
$$\text{CONF(R)} = P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} \quad (2.6)$$

만약, 규칙 'X⇒Y'가 의미가 있다면 전체 트랜잭션의 수에서 아이템 Y를 포함하고 있는 트랜잭션의 비율보다는 아이템 X를 포함하는 트랜잭션에서 아이템 Y를 포함하고 있는 트랜잭션의 비율이 더 클 것이다. 따라서 아이템 X와 아이템 Y를 포함하고 있는 트랜잭션이 서로 상호관련이 없다면(두 아이템이 독립이라면), $P_r(Y|X)$는 $P_r(Y)$와 같게 된다. 이를 상대적으로 표현하기 위한 규칙 'X→Y'의 향상도(Lift)는 식 2.7과 같이 나타낸다. 향상도는 규칙을 모를 때에 비하여 규칙을 알 때에 얼마나 향상되는가를 나타내고 있다.

$$\text{LIFT(R)} = \frac{P(Y|X)}{P(Y)} \quad (2.7)$$

상호 대칭적으로 향상도 (X→Y) = 향상도(Y→X)이다. 따라서 $P_r(Y|X)$의 값은 $P_r(Y)$의 값보다 향상도의 배수만큼 크다. 연관규칙을 찾아내는 대표적인 알고리즘으로는 Apriori, FP-tree, SETM, DIC, ARHP 알고리즘 등이 있다. 연관규칙을 찾아주는 알고리즘 중에서 가장 먼저 개발되었고 가장 많이 사용되고 있는 것은 Apriori 알고리즘이다. 연관 규칙과 순차 패턴이 다른 점은 연관성 규칙은 X→Y, Y→X가 모두 의미있지만, 순차패턴은 X→Y만 의미가 있다. 즉, 연관규칙은 X, Y 중 어느 것이 먼저 일어나도 관계없지만 순차패턴은 반드시 X가 먼저 발생하는 것을 말한다. 다음 (그림 1)은 연관규칙을 추출하는 과정을 나타낸 것이다.

(그림 1) 연관규칙을 추출하는 과정



(Figure 1) Creating routine for association rule

빈발 아이템집합 생성의 첫 번째 과정으로 모든 빈발 아이템집합 C1에서의 아이템집합(Itemset)은 {A}, {B}, {C}, {D}, {E}이고, 최소

지지도(Minimum Support) 2미만을 제외한 L1의 아이템집합은 {A}, {B}, {C}, {E}로 구성된다. 두 번째 과정은 L1에서 아이템집합의 조인과정을 거쳐 C2의 Apriori알고리즘[20]은 아이템집합을 구성하고, 다시 데이터베이스의 스캔과정을 거쳐 새로운 빈발 아이템집합이 발견되지 않을 때까지 계속된다.

2.6 군집화

군집화(Clustering)는 임의의 데이터 집합으로부터 서로 유사한 속성을 가지는 데이터의 군집(Cluster) 또는 세그먼트(Segment)를 추출하는 기법을 의미한다. 거리 기반 군집화 방법으로 고객의 선호도를 다차원 공간상의 점으로 표시하고, 거리를 계산함으로써 전체 고객들의 집합을 k개의 군집으로 나눈다. 고객 a 와 k 사이의 거리는 (식) 2.8과 같이 계산하고, 식에서 a_i는 고객 a 의 속성(차원) i에 대한 선호도 값을 의미한다.

$$d_{a,k} = \sqrt{\sum_{i}(a_i - k_i)^2} \qquad (2.8)$$

본 논문에서는 고객점수 및 Social Data로 구성된 인구통계학적 변수가 적용된 군집화(Clustering)을 위해 k-means 기법을 적용한다. 적용된 k-means 기법은 계산 속도가 빠르고 대량의 자료에서 군집을 발견하는데 상당히 효과적인 것으로 알려져 있다[19].

$$X = \begin{bmatrix} x_{1,1} x_{1,2} \cdots x_{i,n} \\ x_{2,1} x_{2,2} \cdots x_{2,n} \\ \cdot \\ \cdot \\ x_{p,1} x_{p,2} \cdots x_{p,n} \end{bmatrix} \qquad (2.9)$$

3. 마이닝과 FRAT기반 가중치 선호도 군집을 이용한 추천 기법

3.1 군집화를 위한 사용자정보 변수정의

이 절에서는 군집화를 위한 사용자정보 변수에 대한 정의를 하고 아이템에 관련된 분류체계와 선호도에 관한 정의를 하고, 마이닝과 FRAT 기반 가중치 선호도 군집을 위한 관련된 가정, 요구조건 그리고 정의와 그 방법을 기술한다.

군집화 방법은 선호도가 유사한 사용자들을 미리 같은 군집으로 만든다. 군집 내에 속한 사용자들을 이웃으로 선정하여 예측을 수행하므로 군집화의 결과는 예측의 정확도에 직접적인 영향을 준다. 본 논문에서는 추천시스템을 설계하기 위해서 다음과 같은 가정을 전제로 한다.

【가정 3.1】 인터넷 쇼핑몰의 운영 및 관리 측면에서 고객의 신상정보는 지불이 이루어지는 전자상거래관계이므로 정확한 정보로 활용될 수 있음을 가정한다. 따라서 이 논문에서는 고객의 회원정보 이외에 평가 자료를 위한 불필요한 사용자 프로파일은 생성하지 않는다.

【가정 3.2】 인터넷 쇼핑몰의 운영 및 관리 측면에서 생성되는 웹 로그 파일은 축적된 대용량 웹로그 데이터를 복잡한 전처리 과정을 거쳐 분석하게 되면 실시간 처리의 도입에 어려움과 전처리 과정의 부정확성 우려로 활용하지 않는다고 가정한다.

일반적으로 실제적인 데이터 마이닝의 개념은 특징, 클래스 및 패턴 등을 발견하는 일련의 과정을 말한다. 데이터 마이닝은 KDD(Knowledge Discovery in Database)와 동격의 과정으로 볼 수 있다. (그림 2)은 데이터 마이닝의 처리과정을 나타낸 것이다.

(그림 2) 데이터 마이닝의 처리과정

Selec-tion	Pre-processing	Transfor-mation	Analysis for Result	Data Visualiza-tion

(Figure 2) The routine of processing of data mining

3.2 아이템 카테고리 기반의 군집화

<표 2>은 고객의 구매 아이템 목록을 예로 나타낸 것으로 인터넷 쇼핑몰의 장바구니를 거쳐 실제로 구매과정을 거친 고객의 구매 아이템 목록을 기준으로 군집화를 수행하면 <표 2>와 같이 G1, G2, G3.... G14의 14개의 군집이 형성된다. <표 3>는 아이템 코드 번호 기반 고객 군

집 테이블을 예로 나타낸 것이다.

<표 2> 고객 구매 아이템 목록

Customer	Brand Item Code No.
C1	AAA10
:	:
C12	BAC01, CAD34, CAE22
:	:
C20	AAB25

<Table 2> Customer Purchase List

<표 3> 아이템 코드 번호 기반 고객 군집 테이블

Item Code No	Category	Customer	Cluster
AAA01~AAA50	AAA	{C1,C4,C8,C14, C17,C18}	G1
:	:	:	:
BAC01~BAC14	BAC	{C2. C13}	G11
:	:	:	:
CAE01~CAE36	CAE	{C7, C12, C15}	G14

<Table 3> Table of cluster for customer based on brand item code no.

<표 4> 아이템 카테고리 테이블과 소분류별 아이템 종류

Large Group	Medium Group	Small Group	Sorts of Item
A Skin Care	A Skin	A All Skin type	50
		B Dryness	32
		:	:
		E Mixed	24
	B Lotion	A All Skin type	41
		:	:
		E Mixed	24
	C Cream	A Water/Moisture	42
		B Pores of Skin	37
		C Trouble	51
B Man Care	A Skin Care	A Skin	46
		B 2in1	6
		C Lotion	41
Total			580

<Table 4> Item category table and sort of brand item in small grouping

<표 4>는 제안 시스템의 대/중/소분류 아이템 카테고리 구성표와 소분류에 속한 브랜드아이템의 종류(갯수)를 나타낸 것이다. 일반적으로 추천시스템에서 전체 데이터가 대단위일 경우는 유사도 및 선호도 계산 작업량이 많아지게 된다. 먼저 고객별 아이템에 대한 선호도를 계산하고, 고객별 아이템 카테고리 선호도는 아이템 카테고리내의 고객의 아이템에 대한 선호도를 평균을 계산해서 생성한다.

4. 실험 및 성능 평가

이 장에서는 제안 기법의 적용 및 평가에 대한 실험을 위해서 이전 및 기존 기법이 적용된 시스템의 적용 및 실험 환경 소개와 실험 데이터 셋 구성을 위한 실험적용 영역을 설정하고 이런 환경에서 제안 기법을 화장품을 전문적으로 취급하는 인터넷 쇼핑몰을 대상으로 실험하여 FRAT의 적용 분석을 바탕으로 결과를 도출하여 검증한다. 검증된 결과를 바탕으로 시스템을 분석 및 평가를 하여 제안 기법의 효용성 입증한다.

4.1 데이터베이스 구조

본 절에서는 실험 및 평가를 위한 인터넷 쇼핑몰의 화장품 아이템 기본 데이터베이스를 구축하고 고객의 취향에 맞는 화장품 아이템을 추천한다. 화장품 아이템 정보 데이터베이스는 회원정보(Member), 아이템(Items), 구매이력 정보(Sale), 아이템분류(gp), 아이템점수(s_point)로 구성한다.

4.2 FRAT 적용 분석

본 연구에서는 F값, R값, A값, T값을 각각 5개의 세그먼트로 나누어 분석한다. 우선 F값에 의해 고객을 분류하고 다음으로 R값, A값, T값에 의해 고객을 분류함으로써 최종적으로 625개의 FRAT 세그먼트를 생성한다. 그리고 관리의 효율화를 높이기 위하여 가중치를 부여하여 점수화를 실시하고 이를 다시 5개의 고객 세분화 그룹으로 나누어 분석한 후 각 그룹별로 유의성 검정을 통하여 분류 모형을 측정한다.

4.3 제안 추천기법의 시스템 적용

4.3.1 아이템 추천 과정

로그인 사용자가 사이트에서 구매한 사실 유무에 따라 추천방법을 다르게 적용한다. 로그인 사용자의 사용자정보 변수를 적용한 고객분류 코드가 같은 사용자 군집을 발췌하여 고객 성향이 유사한 고객들이 구매한 구매 데이터를 기준으로 사용자별 아이템 카테고리별 선호도를 산출하여 제품점수가 높은 아이템 TOP-4를 추천한다. 반면에 로그인 사용자가 사이트에서 구매한 사실이 없을 경우에는 로그인 사용자의 사용자정보 변수를 적용하여 분류코드가 같은 사용자 집단을 군집화하여 사용자 FRAT점수가 80점 이상의 사용자들이 구매한 아이템 TOP-4를 추천한다.

4.3.2 아이템 선호도에 의한 추천 실험

아이템 선호도에 의한 실험은 로그인 사용자의 해당 군집화에 속한 사용자 정보와 구매데이터를 이용한다. 로그인 사용자의 해당 군집의 구매데이터를 화장품 대분류, 스킨케어 중분류, 아이템 카테고리 소분류를 아이템 군별로 데이터를 재구성하여 사용자별-아이템 카테고리별 선호도를 구한다. 가장 선호도가 높은 아이템 카테고리에 아이템 FRAT점수를 반영하여 추천한다. 다음 <표 5>은 사용자의 화장품 범주 선호도 테이블의 실험 데이터를 나타낸 것이다. 구매가 많이 이루어졌을 경우 각각의 사용자가 화장품 범주에 대해 갖는 선호도가 2.5%(임계치) 이상의 값을 가질 경우 선호도가 높은 아이템으로 구매될 가능성이 높다.

<표 5> 각 사용자의 화장품 범주 선호도 테이블

ID	Skin	Lotion	Cream	...	Pack
phb	0.048	0.052	0.062	...	0.058
msc	0.072	0.093	0.082	...	0.033
hgd	0.262	0.119	0.039	...	0.064
:	:	:	:	:	:
stj	0.048	0.052	0.061	...	4
yjs	0.048	0.052	0.061	...	4

<Table 5> Table for preference rate of item category by each user

한 사용자가 총 구매한 화장품의 빈도 중 화장품이 속한 아이템 범주가 차지하는 고유한 선호도 즉 한 사용자의 총 구매빈도 중 화장품이 차지하는 비율을 계산하고, 각각의 아이템 범주의 총 구매빈도 중에서 그 사용자가 차지하는 비율을 계산한다. 사용자가 화장품 범주에 대해 갖는 선호도가 임계값을 넘는 화장품 범주에 한하여 다음의 사용자간의 아이템별 유사도와 선호도를 구하는데 사용된다. 선호도 임계값의 적용은 구매가 많이 이루어지는 구매데이터의 추세를 분석하여 적용할 수 있으며 구매 데이터 건수가 많을 경우 추천의 정확도를 높이기 위해서 임계값을 적용할 수 있다. 산출된 선호도 확률을 기준으로 임계값 산정할 수 있다. 산출된 선호도 확률 0.25를 임계값으로 산정한다면 아래 <표 4.1>에서 0.25이상의 값을 가지는 화장품 범주 즉 화장품이 각 사용자의 선호가 있는 카테고리가 된다.

5. 결 론

본 논문의 연구 내용은 사용자정보 변수와 고객 FRAT점수 기반의 군집을 이용한 구매이력 데이터를 군집화하여 추천을 위한 데이터 처리 기반을 구축, 고객별 선호도를 파악하기 위해서 아이템 분류체계 기반의 고객별 아이템 카테고리를 통한 선호도 분석의 기반을 구축, 베이시안 네트워크를 이용하여 시간의 변화를 반영한 구매이력 정보 기반 가중치 선호도의 계산 및 학습관리를 위한 FRAT기반 고객점수와 아이템 점수의 정보를 구축, 그리고 마이닝을 이용한 연관규칙 생성으로 고객의 선택 사양으로 연관규칙의 척도에 따른 교차판매 및 상위판매가 가능한 규칙기반으로 마이닝 결과를 제공하도록 제시하였다. 향후 연구로는 대규모 거래가 이루어지는 전자상거래에 제안 기법을 적용하여 추천의 적용분석을 통한 안정성을 확보하는 것이 필요하다.

References

[1] S.G. Park, “Study on Multimedia Content Recom-mendation System based on Usage Pattern Information”, Kwangwoon University, Doctoral Dissertation, 2012.

[2] Yong Kim, “A Study on Hybrid Recommendation

System based on usage frequency for multimedia Contents",Yonsei University, Doctoral Dissertation , 2006.

[3] J.S.Sim, "Product Recommendation System on VLD B using k-means Clustering and Sequential Pattern Technique", Chung-Buk University, Doctoral Disse rtation, 2005.

[4] H.I. Jung, "Design recommendation system using mi ning based sensibility analysis," Sang Ji University, Doctoral Dissertation, 2013.

[5] S.C.Kim, "A Study on Recommendation Technique or a Credible User Authentication in Environment of Smart-home," Chong-Ang University, Doctoral Dissertation, 2012.

[6] K.S.An, S.J.Ko, J.Jung, P.K.Lee, "Generator of Dyna mic User Profiles Based on Web Usage Mining," KMIS-Education, Vol. 9-B, No.4, pp.389 -398, 2002.

[7] S.K.Lee, Y.H.Jo, Y.B.Jo, S.H.Kim, "A Mobile Music Recommending System be personalized Based on M obile Web Mining or ranking preference," KMIS-Ed ucation, pp.582-587, 2004.

[8] Huges, Arthur M, "Strategic Database Marketing ," 2nd., Mcgraw-Hill, pp. 105-137, 2000.

[9] B.K.Kim, "Management of Customer for DataBase Marketing," Namdu Publishing Co, 2000.

[10] J.Y.Woo, "Segmented CRM strategy via custom -er information visualization and its extension to bus iness convergence", KAIST Doctoral Disserta- tion, 2005.

[11] http://www.anzmac.org/conference/ 1998/Cd_rom /Gunaratne120.pdf, p866 by K Asoka Gunaratne.

[12] S.H.Ha, S.M.Bae, S.C.Park, "Customer's time va- riant purchase behavior and corresponding mar- keting strategies: an online retailer's case," Computer s & Industrial Engineering Vol. 43, pp. 801-820, 2002.

[13] B.Sarwar, Karypis.G, Konstan,J and Riedl.J, "Applic ation of Dimensionality Reduction in Reco- mmender System a Case Study," In Proceedings of ACM WebKDD-2000 Workshop, pp.285-295, 2000.

[14] Y.S.Cho, , S.C.Moon, S.C.Noh, K.H.Ryu, "Imple- mentation of Personalized recommendation System Using k-means Clustering of Item Category based on RFM", IEEE, June.2012.

[15] K.Y.Jung and J.H.Lee, "User Preference Mining thr ough Hybrid Collaborative Filtering and Content-ba sed Filtering in Recommendation System," IEICE T ransaction on Information and Systems, Vol. E87-D, No.12, pp.2781-2790, 2004.

[16] Konstan, J.A, B.N.Miller, D.Maltz, J.L.Herlocker, L. R.Gordon, and J.Riedl, "GroupLens: Applying Collab orative Filtering to Usenet News," Communication of the ACM, Vol.40, No.3, pp.77-87, 1997.

[17] K.Y.Jong,"Personalization Recommendation using Context Awareness based Information Filtering," N RF-Education, Report of Result, pp.31-38, 2008.

[18] J.S.Kim,"A Dynamic Recommending System Using User's Ordering Pattern and Similarity of Document within Cluster", InHa University, a master´s thesis, 2001..

[19] Y.K.Lee, W.T.Kim, Y.J.Jung, K.D.Kim, K.H.Ryu, "C luster Analysis of Climate Data for Applying Weath er Marketing", Journal of geographic information sys tem association of Korea, Vol.7, No.3, pp.33-44, 2005.

[20] Y.J. Lee, S.B. Seo, K.H. Ryu, H.K. Kim, "Discoverin g Temporal Relation Rules from Temporal Interval Data", KIISE, vol.28 no.3, pp.301-314, 2001

박 화 범

1994년 : 광운대학교 대학원(공학석사)
2004년 : 광운대학교 대학원(공학박사수료-전자통신공학)

1991년~1993년 : 제성전산원
1994년~1996년 : 태산엘시디
2010년~현 재 : 보이저아이엔씨, 연구원
관심분야 : 디지털영상압축, 워터마킹, 데이터마이닝

조 영 성

1989년 : 연세대학교 대학원(공학석사)
2008년 : 충북대학교 대학원(공학박사-전산학)

1989년~2000년: CDC/Stratus SE Manager
2000년~2013년: 네오아이엔씨(CEO)
2006년~현 재: 동양미래대학 겸임교수, 정보처리 기술지도사(중기청), (주) 컴트리 연구소장
관심분야 : 시공간 데이터베이스, 유비쿼터스 컴퓨팅, 데이터 마이닝, 기계학습, u-커머스, ebXML

고 형 화

1982년 : 서울대학교 대학원(공학석사)
1989년 : 서울대학교 대학원(공학박사- 전자공학)

1979년~1980년: (주) 금성사 중앙연구소
1985년~ 현재 : 광운대학교 전자통신공학과 교수
관심분야 : 영상통신, 문자인식, 생체인식, JBIG2 Wavelet,