k-means 클러스터링과 순차 패턴 기법을 이용한 VLDB 기반의 상품 추천시스템

심 장 섭[†] · 우 선 미[††] · 이 동 하[†††] · 김 용 성[††††] · 정 순 기[†††††]

요 약

대용량 데이터베이스에서의 추천시스템은 많은 문제점들을 지니고 있으므로, 대규모 인터넷 쇼핑몰에 적합한 추천 시스템 구조와 데이터 마이닝 기법의 필요성이 요구되고 있다. 따라서 본 논문에서는 k-mean 클러스터링과 순차 패턴 기법을 이용한 VLDB(very large database) 기반의 상품 추천 시스템을 설계 및 구현한다. 본 논문에서는 사용자의 정보를 일괄처리하고 다양한 카테고리를 계층적으로 정의하며, 탐색엔진에 순차 패턴 마이닝 기법을 이용한다. 예측 모델을 만들기 위하여 사용자의 로그 데이터 중에서 카테고리에 대한 사용자의 선호도를 추출하여 이용한다. 본 논문에서는 실험과 성능 평가를 위하여 국내 인터넷 쇼핑몰에서 30일 동안 수집한 실제 데이터를 이용한다. 또한 성능평가를 위하여 추천 예측 정확율(PRP: Predictive Recommend Precision), 추천 예측 재현율(PRR: Predictive Recommend Recall), 정확도 인수(PF1 : Predictive Factor One-measure)를 제안하여 사용한다. 성능평가 결과 가장 빠른 추천시간 및 학습시간은 O(N)이었고, 다양한 실험에서의 측도들의 값이 상당히 우수하였다.

키워드 : 상품추천, K-means 클러스터링, 순차 패턴

Product Recommendation System on VLDB using k-means Clustering and Sequential Pattern Technique

Jang-sup Shim[†] · Seon-Mi Woo[††] · Dong-ha Lee[†††] · Yong-Sung Kim[††††] · Soon-key Chung[†††††]

ABSTRACT

There are many technical problems in the recommendation system based on very large database(VLDB). So, it is necessary to study the recommendation system' structure and the data-mining technique suitable for the large scale Internet shopping mall. Thus we design and implement the product recommendation system using k-means clustering algorithm and sequential pattern technique which can be used in large scale Internet shopping mall. This paper processes user information by batch processing, defines the various categories by hierarchical structure, and uses a sequential pattern mining technique for the search engine. For predictive modeling and experiment, we use the real data(user's interest and preference of given category) extracted from log file of the major Internet shopping mall in Korea during 30 days. And we define PRP(Predictive Recommend Precision), PRR(Predictive Recommend Recall), and PF1(Predictive Factor One-measure) for evaluation. In the result of experiments, the best recommendation time and the best learning time of our system are much as O(N) and the values of measures are very excellent.

Key Words : Product Recommendation System, K-means Clustering, Sequence Pattern

1. 서 론

최근 초고속 인프라의 확산과 인터넷의 폭발적인 대중화를 기반으로 다양한 형태의 전자 상거래가 급성장하고 있다. 또한 고객의 취향이나 관심에 초점을 맞추어 고객에게 상품이나 컨텐츠를 제공하는 고객맞춤 전략이 온라인 쇼핑몰이나 정보서비스 제공자에게 있어서 성공을 위한 필수적인 요소가 되고 있다[1]. 개인화된 상품 추천 시스템(personalized product recommendation system)은 해외의 경우 Amazon, J.C.Penny, Yahoo, CD Now, HP Shopping Village, 월마트, Half.com, 및 Musician' s Friend 등의 유수한 전자 상거래 사이트에서 이미 활용되고 있으며, 국내의 경우 삼성몰, 한솔CS클럽 및 SK몰 등 일부 유명 쇼핑몰 사이트 및 통신회사 등에서 활발히 도입하고 있다[1, 2, 4]. 현재 상품 추천 시스템에 주로 활용되고 있는 데이터마이닝 기법으로는 협

† 종신회원 : 정보통신연구진흥원 정보화추진팀장
†† 정 회 원 : 전북대학교 전자정보공학부 시간강사
††† 정 회 원 : (주) 넷스루 데이터 마이닝 연구소 연구소장
†††† 종신회원 : 전북대학교 전자정보공학부 교수
††††† 정 회 원 : 충북대학교 전기전자 컴퓨터공학부 교수

논문접수 : 2005년 9월 26일, 심사완료 : 2006년 9월 29일

업 필터링(collaborative filtering)이나 속성기반 필터링(contents based filtering) 등이 있으며, 이러한 기법들을 조합한 하이브리드(hybrid) 기법 또는 연관 규칙(association rule) 등이 활용되고 있다. 위와 같은 기법들을 이용하고 있는 기존 상품 추천 시스템의 문제점은 다음과 같다[5]. 첫째, 대형 온라인 쇼핑몰의 경우에 사용자와 상품 종류가 매우 많고, 계층구조를 갖는 상품 카테고리 수도 수만 개 이상이 되기 때문에 협업 필터링 기법이 가지고 있는 데이터 희소성(sparseness)과 시스템 확장성(scalability) 문제로 인하여 구현에 어려움이 있다[5, 6]. 둘째, 기존 상품 추천 시스템에서는 시간 변화에 따른 고객의 관심도의 변화를 고려하지 않고 있다.

본 논문에서는 이러한 기존 추천 시스템의 문제점들을 해결하기 위하여 대용량 계층구조를 가지는 데이터 처리 환경에서 고객에게 상품을 효과적으로 추천할 수 있는 상품 추천 시스템을 설계 및 구현한다. 본 논문에서는 대형 온라인 쇼핑몰에서 고객들의 과거 상품구매 이력으로부터 고객이 선호하는 상품 패턴을 추출하고, 추출된 상품 패턴들을 이용하여 상품들을 카테고리(category)화하며, 상품 카테고리 간의 연관성은 계층구조를 갖는 트리형태로 표현한다. 본 논문에서 제안하는 상품 추천 시스템은 시간 변화에 따른 고객의 상품구매 패턴을 실시간으로 학습하는 상품패턴 학습 모듈과 고객의 선호도를 분석하여 고객에게 최적의 상품을 추천해주는 추천 모듈로 구성된다. 본 논문에서 제안하는 개인화된 상품 추천 시스템의 성능은 초당 5만 명 이상의 고객접근을 전제로 하며, 상품추천에 소요되는 시간도 초당 5천명으로 가정하여 고객 1인당 처리시간이 2ms 이하가 되도록 한다. 고객들을 분할하는 고객 세그먼트의 구성에는 k-means 클러스터링 기법을 사용하고, 고객 세그먼트에 대한 상품패턴 학습에는 순차 패턴 기법을 사용한다. 또한 목표 고객에게 고객이 선호하는 상품들을 정렬하도록 하고, 지지도가 높은 상품을 추천하기 위하여 Match-Find 알고리즘을 제안하여 적용한다.

본 논문에서 제안한 추천 시스템의 성능평가를 위하여 50만명의 고객, 100만개 이상의 상품을 대상으로 하여 30일 동안 수집한 웹 로그 데이터를 이용한다. 그리고 5만개 이상의 상품 카테고리를 대상으로 하여 상품패턴의 학습 성능, 상품 추천 성능 및 상품 추천 타당성을 평가한다. 2장에서 기존 상품 추천 시스템들이 주로 사용하고 있는 데이터 마이닝 기법들을 알아보고, 상품 추천 시스템의 활용에 대한 국내외 연구 현황을 고찰한다. 3장에서는 본 논문에서 제안하는 VLDB(Very Large Database) 기반의 개인화된 추천 시스템을 설계하고, 4장에서 본 시스템의 실험 및 성능평가를 수행한다. 마지막으로 5장에서 본 논문의 결론 및 향후 연구 방향에 대하여 기술한다.

2. 관련 연구

본 장에서는 상품 추천 시스템에서 주로 사용하고 있는 기법들과 상품 추천 시스템의 활용에 대한 국내외 연구 현황을 살펴본다.

고객에게 개별적으로 상품을 추천하는 기술은 웹 사이트의 개인화 기술 중의 하나로서, 고객에게 추천 상품 리스트를 제시하여 고객이 원하는 상품을 용이하게 검색할 수 있도록 도와주는 정보 필터링 기술을 의미한다. 현재 가장 널리 이용되고 있는 기법은 협업 필터링과 속성기반 필터링 기법이고, 그 외에도 규칙기반(rule-based) 필터링 기법 및 에이전트(agent) 기술 등이 있다[5, 7, 8]. 온라인 쇼핑몰에서는 사용자의 의사결정에 필요한 정보를 처리하기 위하여 주로 협업 필터링(collaborative filtering) 기법을 사용하고 있다. 일반적으로 협업 필터링 기법을 이용하는 상품추천 과정은 크게 고객과 상품에 관련된 데이터 구성, 이웃고객 집단의 탐색, 그리고 추천 상품 결정의 3 단계로 이루어진다[9]. 이 방법은 목표 고객과 선호도가 유사한 고객들의 상품에 대한 관심도를 수치화하여 산출된 평가 점수를 활용하여 목표 고객에게 적합한 상품정보를 제공한다[6]. 협업 필터링 기법은 여러 분야에서 이용되고 있음에도 불구하고 예측대상이 되는 데이터의 희소성 문제, 시스템 확장성 및 처리속도 문제로 인하여 적용상에 많은 제한을 가지고 있다[10]. 따라서 고객 선호도의 시계열적 변화를 추적할 수 있는 기법을 도입하여 상품 추천뿐 아니라 상품 추천의 만족도까지도 계산할 수 있는 협업 필터링 기법에 대한 연구가 수행되고 있다[3]. 그리고 추천 시스템의 단점을 보완하기 위하여 데이터마이닝(data mining) 기법이 널리 이용되고 있다. 데이터마이닝이란 대용량 DB로부터 유효하고, 전에는 알려지지 않은 포괄적이고도 활동적인(actionable) 정보를 추출하는 방법으로서 중요한 비즈니스 의사결정에 이용되고 있다[11]. 클러스터링 기법은 임의의 데이터 집합으로부터 서로 유사한 속성을 가지는 데이터의 군집(cluster) 또는 세그먼트(segment)를 추출하는 기법이다. 순차 패턴이란 연관 규칙에 시간 개념을 도입하여 트랜잭션에서 사용되는 데이터들로부터 특정한 시간대에 가장 많이 사용되는 데이터 시퀀스(sequence)의 패턴을 추출하는 데이터마이닝 기법이다. 순차 패턴 기법은 트랜잭션의 발생 순서에 따라 각 트랜잭션에서 사용되는 상품 항목들의 연관성을 탐색하여 트랜잭션 간에 영향을 미치는 선행 항목을 검색한다. 순차 패턴 기법은 고객의 상품구매 패턴을 분석하여 향후 구매 가능한 상품의 예측은 물론, 의료 분야에서는 질병 발생순서의 패턴을 분석하여 진료 및 투약 정보 제공에 이용되고 있다. 최근에는 웹 로그 데이터를 이용하여 고객이 주로 방문하는 웹 사이트의 경로를 분석하여 사이트의 재구성 및 광고용 배너의 재배치 등에 사용되고 있다.

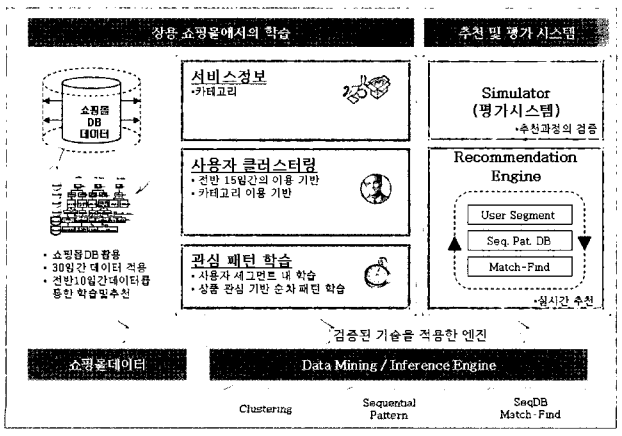
추천 시스템에 대한 국내 주요 연구로는 지수적(exponential) 가중치를 적용하는 협력적 상품 추천 시스템에 관한 연구가 있다[12]. [12]는 전자상점에서 주로 사용하는 추천 시스템으로서 최근의 고객 거래 데이터에 높은 가중치를 부여하고, 오래된 고객 거래 데이터에는 지수적으로 감소하는 낮은 가중치를 적용함으로써 유사 고객집합에게 최근의 상품을 추천할 수 있는 확률을 높여주는 방법을 사용하고 있다. 개인

화된 추천 시스템 분야에서는 상품의 계층 구조나 고객 선호도를 이용한 방법에 관한 연구가 많이 진행되고 있다. 이러한 연구에는 고객의 선호도 변화를 반영할 수 있는 상품 추천 방법, 웹 마이닝과 상품의 계층구조를 이용하는 협업 필터링 기법, 모바일 상거래에서 고객의 피드백을 고려하는 멀티미디어 컨텐츠 추천 방법 등이 있다. 개인화 추천시스템에 대한 해외 주요 연구로는 웹 기반의 데이터마이닝 기법을 사용하여 의사결정 과정을 트리구조로 표현하여 상품을 추천하는 연구와 시간 프레임(time frame) 추적(navigation) 클러스터링 기법 및 연관 규칙을 이용한 상품추천 방법 등이 있다[13]. 웹 기반의 데이터마이닝 기법은 웹 사이트에서 고객의 이동 경로(클릭 스트림) 정보를 이용하여 웹 로그로부터 고객들의 접속 관계, 패턴 및 규칙 등을 분석하고, 이를 모델화하여 유용한 마케팅 정보로 변환시킨다[14].

3. VLDB 기반의 상품 추천 시스템 설계

본 논문에서 제안하는 추천 시스템의 구조는 (그림 1)과 같이 학습 모듈, 추천 모듈 및 평가 모듈로 구성된다.

추천 시스템을 구성하는 학습 모듈과 추천 모듈의 연동을 통하여 상품을 추천하는 과정을 간단히 설명하면 다음과 같다. 첫째, 학습 모듈에서는 목표 고객에 대한 이웃고객(neighborhood)을 추출한다. 이웃고객이란 유사한 상품 선호도를 가지는 고객군집을 의미하며, 웹 로그로부터 고객들이 선호하는 상품 및 상품 카테고리 정보를 기준으로 k-means 클러스터링 알고리즘을 적용하여 이웃고객을 추출한다. 둘째, 생성된 이웃고객으로부터 접근 빈도수가 높은 상품들의 순차 패턴을 추출한다. 이웃고객들이 선호하는 상품들의 순차 패턴을 일일 단위로 추출하여 순차 DB를 생성한다. 셋째, 특정한 고객이 추천 시스템을 접근하면 고객의 과거 구매이력 데이터를 이용하여 상품의 접근패턴을 생성한다. 추천 모듈의 Match-Find 알고리즘을 이용하여 순차 DB으로부터 특정 고객의 상품접근 패턴을 검색한다. 즉 특정 고객의 상품접근 패턴이 순차 DB에 존재하면 순차 패턴내의 상품을 추천 상품으로 선택한다.



(그림 1) 상품 추천 시스템의 구조

3.1 상품 카테고리와 상품 선호도

고객이 대형 인터넷 쇼핑몰을 이용하여 상품을 구매할 경우 추천 시스템은 고객의 모든 DB 접근정보를 웹 로그에 기록 및 유지 관리하며, 웹 로그에 저장된 데이터를 분석하여 고객에게 유용한 상품 정보를 제공한다. 본 논문에서 추천 시스템은 상품 수가 100만 개 이상, 상품 카테고리 수가 5만개 이상 그리고 대상 고객이 천만 명 이상이 되는 대형 쇼핑몰을 전제로 하여 설계한다.

본 논문에서는 고객의 신상정보 대신에 상품 카테고리에 대한 고객의 관심도를 이용하여 고객들을 클러스터링한다. 고객들의 클러스터링에는 k-means 클러스터링 기법을 사용하므로 고객군집으로부터 고객이 가장 선호하는 상품들의 순차 패턴을 추출할 때 소요되는 시간 복잡도를 O(N)에 가깝게 감소시킬 수 있다. 고객집합 U는 U = {u1, u2, u3, ... ,un }, 상품집합 P는 P = {p1, p2, p3, ... ,pm }, 상품 카테고리 집합 C는 C = {c1, c2, c3, ... ,cr}와 같이 표현한다. 이때 u_j는 쇼핑몰의 고객을, p_i는 쇼핑몰의 상품을, c_k는 상품 카테고리를 나타낸다.

본 논문에서 제안하는 추천 시스템은 상품과 컨텐츠 모두를 취급한다. 상품들은 특징에 따라 특정한 카테고리에 속하게 되며 상품 카테고리는 단일 계층구조(single hierarchy)의 트리구조를 갖는다. 상품 카테고리의 구조는 (정의 1)과 같다.

【정의 1】 추천 시스템에서 사용하는 상품 카테고리는 트리구조를 갖는다.

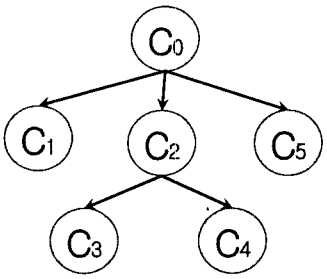
【정리 1】 상품 $p_i(i \in m)$이 소속된 카테고리를 $C(p_i)=C_j(j \in r)$라고하면, p_i의 상위 카테고리는 $TC(p_i)=C_k(k \in r)$ 이다. 이때 C_k를 C_j의 부모 카테고리라고 한다.

【증명】 (정의 1)에 의해서 상품 카테고리는 트리구조를 갖기 때문에 함수 TC(pi)의 값은 Ck가 된다. 따라서 Cj의 부모 노드는 Ck이다. (정의 1)과 (정리 1)로부터 다음과 같은 결과를 도출할 수 있다.

1) 최상위(루트) 카테고리 C0로부터 하위 카테고리로 이어지는 경로상의 모든 카테고리들은 C0의 자식 카테고리가 된다.
2) 자식 카테고리가 없는 카테고리를 리프(leaf) 카테고리라고 하며, 각각의 상품은 리프 카테고리에 포함된다.
3) 트리는 균형을 유지할 필요가 없다.

트리의 노드는 상품 카테고리 ID와 링크로 구성된다. 임의의 날짜에 고객이 접근한 상품 카테고리에 대한 로그 데이터의 일부를 예를 들면 다음과 같다.

{<u1, c1, 5>, <u1, c2, 10>, <u2, c2, 5>,
<u2, c3, 10>, <u2, c4, 5>, <u2, c5, 5>}



(그림 2) 카테고리 트리 구조

위의 로그 데이터에서 u_i는 고객, c_i는 접근한 카테고리 그리고 숫자는 카테고리에 대한 선호도를 나타낸다. 그리고 상품 카테고리의 트리구조는 (그림 2)와 같다고 가정한다.

하위 카테고리의 고객 선호도를 상위 카테고리의 선호도에 반영하지만, 상위 카테고리의 고객 선호도는 하위 카테고리로 반영하지 않는다. 그리고 고객의 목표 카테고리를 탐색할 때 반드시 트리의 경로를 고려할 필요는 없다. 이유는 특별 관리 대상이 되는 카테고리가 존재할 수 있으며, 고객의 접근 빈도수가 높은 인기 카테고리에 대한 직접 링크(hot link)가 존재할 수 있기 때문이다. 하위 카테고리의 고객 선호도를 상위 카테고리에 반영하기 위해서는 상향식 방법을 이용한다. 앞의 예로 든 로그 데이터에서 하위 카테고리 c3와 c4에서 고객 u2의 선호도(Pref_Count)를 합하여 상위 카테고리 c2의 선호도에 가산시키고, 그 결과를 다시 루트 카테고리 c0의 선호도에 합산시키는 순서를 나타내면 다음과 같다.

Pref_Count(u2, c3) = 10,
Pref_Count(u2, c4) = 5,
Pref_Count(u2, c2) = 10+5+5=20,
Pref_Count(u2, c0) = 20+5=25

위와 같은 고객의 선호도 반영과정에서 고객 자신이 선호하는 개별적인 카테고리에 대한 선호도를 보다 자세히 분석하기 위해서 다음과 같이 각 카테고리의 선호도 값을 루트 카테고리의 선호도 값으로 나눈다.

Pref_Count(u2, c2) = 20/25 = 0.8,
Pref_Count(u2, c3) = 10/25 = 0.4,
Pref_Count(u2, c4) = 5/25 = 0.2,
Pref_Count(u2, c5) = 5/25 = 0.2

위에서와 같이 하위 카테고리의 선호도를 상위 카테고리로 반영할 경우 상위 카테고리가 항상 높은 선호도 값을 가지게 된다. 따라서 서로 다른 레벨에 있는 카테고리를 추천할 경우 항상 상위 카테고리가 선택된다. 본 논문에서는 이러한 문제를 해결하기 위하여 카테고리의 레벨에 따라 레벨 바이어스(bias) 가중치를 부여하는 선호도 보정식을 (식 1)과 같이 제안한다.

Modified_Pref
= Pref_Count-(LevelBias*sum(Child_Pref_Count)) (식 1)

(0 <= 가중치 <= 1)인 조건에서 레벨 바이어스(LevelBias)가 0.5가 될 경우, Modified_Pref(u2, c2)는 0.8 - (0.5 * (0.4 + 0.2)) = 0.5가 된다. 레벨 바이어스가 1.0이 될 경우, 하위 카테고리의 선호도는 상위 카테고리로 전달되지 않는다.

웹 로그 DB의 스키마는 고객 ID, 상품 카테고리 ID, 상품 ID, 접근시간으로 구성된다. 이런 웹 로그 DB를 이용하여 날짜별 고객의 선호도를 나타내는 DB의 스키마는 고객 ID, 상품 카테고리 ID, 날짜, 계수(count)로 구성한다. 이때 계수는 상품 또는 카테고리에 대한 고객 선호도를 나타내며, 이는 고객이 접근한 웹 페이지의 힛팅 수를 의미한다. 고객의 개인별 상품 선호도는 특정 기간 내에 접근한 상품의 전체 항목 중에서 페이지 힛팅 비율이 높은 상품을 의미하며, 모든 상품 선호도의 합계를 정규화(normalization)하여 [0,1]이 되도록 한다. 각 상품 카테고리에 대한 평균 선호도는 각 카테고리의 힛팅 수를 루트 카테고리의 힛팅 수로 나눈 값이 된다. 따라서 루트 카테고리의 선호도 값은 1이 된다. 각 상품에 대한 선호도는 해당 상품이 속한 카테고리 선호도에 영향을 미치며, 하위 카테고리의 선호도는 상위 카테고리의 선호도에 영향을 미친다. 그리고 특정 고객의 u의 상품접근 총수를 P_Total(u)이라고 하고, 특정 고객 u의 상품 카테고리 접근 총수를 C_Total(u)로 표현한다.

임의의 날짜에 고객이 접근한 상품에 대한 웹 로그 데이터가 {<u1, p1, 2>, <u1, p2, 6>, <u2, p2, 1>, <u2, p3, 1>}와 같다면, 고객 u1의 선호도는 {(p1, 0.25), (p2, 0.75)}이 되며, 고객 u2의 선호도는 {(p2, 0.5), (p3, 0.5)}이 된다. 두 고객의 상품접근 수를 {<u1, 8>, <u2, 8>}로 표현했을 경우 P_Total(u1) = 8이 된다. 임의의 날짜에 고객이 접근한 상품 카테고리에 대한 웹 로그 데이터가 {<u1, c1, 5>, <u1, c2, 10>, <u2, c2, 5>, <u2, c3, 15>}과 같을 경우, 고객 u1의 상품 카테고리 선호도는 {(c1, 0.33), (c2, 0.67)}이 되고, 고객 u2의 상품 카테고리 선호도는 {(c2, 0.25), (c3, 0.75)}이 된다. 그리고 두 고객의 상품 카테고리 접근 수가 {<u1, 15>, <u2, 20>}와 같을 경우, C_Total(u1)=15가 된다. 특정 상품에 대한 선호도는 그 상품이 소속된 카테고리나 다른 상품 카테고리의 선호도에 영향을 미칠 수 있다. 상품 카테고리 {C(p1)=c1, C(p2)=c3, C(p3)=c2 }가 주어지고, 카테고리 c2는 c1의 하위 카테고리이고, c1은 c0의 하위 카테고리일 경우 카테고리 선호도와 해당 카테고리 내의 특정 상품에 대한 선호도의 비율은 CP_ratio=(카테고리 내의 상품에 대한 선호도 / 카테고리에 대한 선호도)로 표현한다. CP_ratio 값을 1로 고정하고, 상품이 소속된 카테고리에만 선호도를 반영한다면 카테고리 c1에 대한 선호도는 Pre_Count(u1, c1)=5+2=7로 계산할 수 있다. 만약 CP_ratio 값이 1이 아닌 경우 카테고리 c1에 대한 선호도는 Pref_Count(u1, c1)=(c1의 힛팅 수+(CP_ratio * c1))로 계산할 수 있다.

3.2 학습 모듈

학습 모듈에서 일괄처리 방법으로 처리되는 주요 작업과 수행과정은 다음과 같다.

(1) 웹 로그로부터 고객의 정보 추출
(2) 상품과 상품 카테고리 정보(카테고리 관심도)의 추출
(3) 고객의 특성 벡터(feature vector) 생성
(4) 카테고리 관심도를 기준으로 고객 군집화(k-mean clustering 이용), 즉 이웃 고객의 추출
(5) 고객 군집의 상품접근 순차 패턴 추출 및 학습

3.2.1 k-means 클러스터링을 이용한 이웃 고객의 추출

임의의 날짜 d 에서 고객 u의 상품 카테고리 ck에 대한 선호도를 나타내는 함수 Pref_CD는 (정의 2)와 같다.

【정의 2】 함수 Pref_CD(u, c_k, d)는 고객(u), 날짜(d)의 상품 카테고리 (c_k)에 대한 선호도를 나타낸다.

웹 로그로부터 이웃고객을 추출하는 알고리즘은 (그림 3)과 같다.

생성된 고객군집들의 예를 들면 다음과 같다.

고객군집 1 : 1 주일간의 카테고리 선호도가 10 이상이고, A 카테고리에 대한 선호도가 5이상인 고객
고객군집 2 : 1 주일간의 카테고리 선호도가 10 이상이고, A 카테고리에 대한 선호도가 5미만인 고객
고객군집 3 : 1 주일간의 카테고리 선호도가 10 미만인 고객

```
입력 : 고객-카테고리-선호도(UCI), 카테고리 트리(PCT), 기간
   (D) /*본 논문에서는 기간 D를 15일로 하였다. */
출력 : 특징 벡터(Feature Vector), 이웃고객
```

```
begin
1. for( PCT에서 레벨이 2 이상인 모든 카테고리 )
   하위 카테고리의 UCI들을 연쇄적으로 합산하여 레벨 1의
   카테고리에 합산한다.
2. 고객의 특징 벡터(Feature Vector) 생성
   2.1 D 별로 Pref_CD(u,c,d)를 합산하고, 이를 정규화 한다.
   2.2 각 고객과 레벨 1 카테고리간의 상대적 선호도 산출한다.
   2.3 다음 공식을 이용하여 고객의 특징 벡터 V를 생성한
       다. 레벨 1 카테고리가 M 개 일 때,
       V = (V1, V2,, ..., VM)
       Vi = Sumk(Pref_CD(u,ci,dk)) /
            Sumi(Sumk(Pref_CD(u,ci,dk)))
3. 선호도가 높은 카테고리 정보를 이용하여 이웃고객을 생성
   한다. /* k-means 클러스터링 알고리즘을 이용하여 고객군
   집을 추출한다.*/
end.
```

(그림 3) k-means 클러스터링 기법을 이용한 이웃고객 추출 알고리즘

3.2.2 상품접근 순차 패턴의 추출

순차 패턴 마이닝 기법을 이용하여 고객 군집으로부터 접근 빈도수가 높은 상품의 순차 패턴을 추출한다. 고객 군집 내의 각 고객들의 상품접근 정보를 일정기간(예: 1 주일)동안 수집하여 누적시킨다. 그리고 고객군집으로부터 상품접근 순차 패턴을 추출하여 순차 DB에 저장한다. 상품접근 순차 패턴 S와 특정 패턴 검색에 사용되는 함수를 (정의 3)과 같이 정의한다.

【정의 3】 순차 패턴 $S = (P_1, P_2, \ldots, P_n)$ 로 표현한다. 순차 패턴 S는 연결 리스트 $S = S_1 \rightarrow S_2 \rightarrow \ldots \rightarrow S_n$ 로 구성된다. 단, '→' 는 패턴의 링크(link)를 의미한다.

순차 패턴 S에서 특정 날짜에 해당하는 상품접근 패턴의 추출에는 다음과 같은 함수를 사용한다.

Head(Si) = S1, ..., Sn-1, Tail(Si) = Sn.

순차 패턴의 포함(inclusive) 관계는 (정의 4)와 같다.

【정의 4】 순차 패턴 $\alpha=a_1 \rightarrow a_2 \rightarrow \ldots \rightarrow a_n$과 $\beta=b_1 \rightarrow b_2 \rightarrow \ldots \rightarrow b_m$에서, 만일 $1 \leq i_1 < i_2 < \ldots < i_n \leq m$이면서, $a_1 \subseteq b_{i1}$, $a_2 \subseteq b_{i2}$, ..., $a_n \subseteq b_{in}$이 되는 정수 i_1, i_2, ... i_n이 존재한다면 α는 β에 포함되었다고 한다. 따라서 순차 패턴의 포함 관계는 $\alpha \subseteq \beta$이다.

그리고 고객군집 k에서 추출된 상품접근 순차 패턴을 저장하는 순차 DB를 SeqDBk라고 한다. 그러므로 k개의 고객 군집이 존재한다면 SeqDB도 k개가 존재하며, 각 SeqDB에는 n개의 순차 패턴이 존재한다. 순차 패턴 마이닝 기법을 이용하여 고객군집으로부터 상품접근 순차 패턴을 추출하는 알고리즘은 (그림 4)와 같다.

```
입력 : 고객군집의 집합, 상품/카테고리 선호도
출력 : 고객군집 수 k, k 개의 SeqDB
```

```
begin
1. 고객- 상품-카테고리 선호도를 날짜별로 처리한다.
2. 고객군집에 포함된 모든 고객의 상품/카테고리 정보에
   대하여 순차 패턴 마이닝을 수행, 고객군집의 순차 패
   턴을 추출하여 SeqDB에 저장한다.
end.
```

(그림 4) 고객군집으로부터 상품접근 순차 패턴 추출 알고리즘

3.3 추천 모듈

특정 고객이 추천 시스템에 접근하면 웹 로그에 기록된 고객의 과거 구매이력 데이터를 이용하여 고객이 주로 접근한 상품의 패턴을 생성한다. 그리고 Match-Find 알고리즘을 이용하여 특정 고객의 상품접근 패턴과 고객이 소속된 고객

군집으로부터 이미 생성된 SeqDB의 순차 패턴을 비교하여 고객의 관심도가 높은 순차 패턴 리스트를 추천한다. Match-Find 알고리즘은 SeqDB로부터 고객이 선호하는 상품 리스트를 탐색하기 위하여 순차 패턴의 포함 관계를 이용한다. Match-Find 알고리즘에서는 SeqDB에 포함된 특정 고객의 순차 패턴 su를 검색하기 위하여 함수 Include(Head(s), su)를 만족하는 순차 패턴의 Tail(s)를 검색한다. Tail(s)의 결과로 여러 개의 상품들이 생성될 경우 지지도(support)를 기준으로 상품들을 정렬하여 지지도가 높은 상품을 고객에게 추천한다. 추천 모듈에서 상품을 추천하는 과정을 예로 들어 설명하면 다음과 같다. 상품추천 과정의 이해를 돕기 위하여 SeqDB로부터 순차 패턴을 추출할 때 Tail(s)의 크기가 1 인 패턴만을 추출하는 것으로 가정한다. 추천하는 단계를 예를 들어 설명하면 다음과 같다.

1) 특정 고객이 소속된 고객군집의 순차 패턴 S는 5 개로 SeqDB에 다음과 같이 저장되어 있다고 가정한다.
   S1 : (1, 6) → (4) → (4),
   S2 : (1, 6) → (2) → (4),
   S3 : (1, 6) → (4) → (5),
   S4 : (6) → (2, 4) → (4),
   S5 : (1, 6) → (2, 4)

2) 특정 고객의 과거 구매이력을 분석한 결과 다음과 같은 순차 패턴 su가 추출되었다고 가정한다.
   Su : (1, 6) → (2)

3) SeqDB로부터 함수 Include(Head(S), su)를 만족하는, 즉 Head(S)가 su를 포함하는 순차 패턴을 검색하면 다음과 같다.
   S1 : (1, 6) → (4) → (4),
   S2 : (1, 6) → (2) → (4),
   S3 : (1, 6) → (4) → (5),
   S5 : (1, 6) → (2, 4)

4) 함수 Tail(S)의 수행 결과로 지지도가 높은 상품 (4), (5)를 고객에게 추천한다.

순차 패턴을 수행할때 목표로 하는 패턴수와 실행 시간과의 관계를 제어하기 위하여 L1 제한 값을(정의 5)와 같이 정의한다.

【정의 5】 L1('Large itemset size 1' or 'Size 1 large itemset')은 순차 패턴 중 길이가 1인 순차 패턴의 개수로 정의하고, L1 제한값은 길이가 1인 순차 패턴 길이의 최대값을 제한하는 값으로 정의한다.

추천 모듈에서 상품 추천에 사용하는 Match-Find 알고리즘은 (그림 5)와 같다.

```
입력 : 고객 u
출력 : p 개의 상품

begin
1. 고객 u의 상품접근 패턴 Su를 웹 로그로부터 추출한다.
2. 고객 u가 속하는 고객군집 i를 결정한다.
3. for(SeqDB_i 의 각 순차 패턴에 대하여)
   if (Include(Head(St), Su) then
      Tail(St)로부터 상품 리스트를 추출한다.
4. 지지도를 기준으로 상품 리스트를 정렬하고, 상위 p 개의
   상품을 선택한다.
end.
```

(그림 5) Match-Find 알고리즘

4. 실험 및 성능 평가

본 논문에서 제안하는 상품 추천 시스템의 실험 환경과 성능 평가를 기술한다.

4.1 실험 환경

본 논문에서 제안한 추천 시스템의 실험을 위한 환경은 다음과 같다.

1) 컴퓨터 하드웨어 : Intel PentiumTM 4, 2.8 GHz, 512MB
2) 운영체제 : MS Windows 2003TM
3) 추천 시스템의 주요 모듈의 구현에 사용된 도구는 <표 1>과 같다.

〈표 1〉 주요 모듈의 구현 환경

시스템	모듈	구현 도구	처리 내용
추천 시스템	전처리 모듈	WiseLog PremiumTM, SQL Server 2000TM	웹 로그 데이터의 전처리 수행
	학습모듈	Visual C++TM	고객 군집 생성, 순차 패턴 생성
	추천모듈	Visual C++TM	Match-Find 알고리즘 수행
평가 시스템	평가모듈	SQL Server 2000TM	상품추천 성능평가 수행

4) 실험 데이터 : 현재 운영중인 대형 인터넷 쇼핑몰에서 수집한 웹 로그 데이터로서, 고객 식별자(ID)를 기준으로 5% 범위 내에서 임의 샘플링하여 고객 수를 제한하였다. 로그 수집기간은 30일, 대상 고객은 50만 명, 카테고리 수는 5만개, 상품 개수는 100만개이다. 원본 로그 데이터를 전처리하여 실험에서 사용할 목적으로 변형된 30일 분량의 접근 상품 수는 900만개이고 접근한 상품 카테고리의 수는 300만개이다.

본 논문에서는 상품검색 성능의 평가척도로서 IRS(Information Retrieval System)에서 사용하는 정확률(precision ratio)과 재현

율(Recall ratio)을 변형한 추천 예측 정확율(PRP: Predictive Recommend Precision ratio)와 추천 예측 재현율(PRR: Predictive Recommend Recall ratio)을 정의하여 사용한다. PRP는 추천리스트 중에서 몇 개의 상품을 고객이 실제로 좋아했는지를 의미하고, PRR은 고객이 좋아하는 상품 중에서 몇 개를 찾아냈는가를 의미한다.

【정의 6】 추천 예측 정확율(PRP: Predictive Recommend Precision ratio)과 추천 예측 재현율(PRR: Predictive Recommend Recall ratio)은 다음과 같이 정의한다.

$$PRP = 100 \times \frac{\text{추천대상 상품 수}}{\text{고객이 검색에 성공한 총 상품 수}}$$

$$PRR = 100 \times \frac{\text{추천 대상 상품 수}}{\text{고객이 검색한 총상품 수(=성공수+실패수)}}$$

일반적으로 웹 사이트를 검색할 때 검색대상이 되는 웹 페이지는 수만 페이지 이상이 되지만 고객이 실제로 검색하는 중요 페이지 수는 매우 적기 때문에 PRR은 낮게 된다. 그러나 검색된 웹 페이지들 중에서 고객이 주로 선호하는 페이지 수는 증가하기 때문에 PRP는 PRR 보다 높게 된다. 또한 PRP와 PRR은 서로 상충적인 특성을 가지고 있으므로, 추천시스템의 성능이 어느 정도 양호한 가를 평가하기 위한 종합적인 추천 성능평가 척도로서 정확도 인수(PF1 : Predictive Factor One-measure)를 정의하여 사용한다. 그리고 상품 추천에 성공한 고객의 비율을 나타내기 위하여, 인터넷 쇼핑몰의 관점에서 의미가 있는 추천 성공 고객 비율(SCR: Success Customer Ratio) 척도를 정의하여 사용한다.

【정의 7】 정확도 인수(PF1 : Predictive Factor One-measure)와 추천 성공 고객 비율(SCR : Success Customer Ratio)은 다음과 같이 정의한다.

$$PFI = \frac{2 \times PRP \times PRR}{PRP + PRR}$$

$$SCR = 100 \times \frac{\text{추천 대상상품 중에서 상품을 선택한 고객의 수}}{\text{전체 고객 수}}$$

4.2 실험

본 논문에서 제안한 상품 추천 시스템의 주요 모듈별로 실험 방법을 설명한다.

4.2.1 전처리 모듈

전처리 모듈에서는 웹 로그 데이터를 분석하는 WiseLog PremiumTM과 MS SQL ServerTM를 이용하여 30일 분량의 고객들에 대한 인터넷 쇼핑몰 이용 정보를 가공한다. 아래와 같은 단계로 각 날짜별로 고객의 관심 상품 및 최 상위 관심 상품 카테고리 접근 정보를 수집, 가공하여 학습모듈과 추천모듈에서 활용할 수 있는 형태로 정보를 가공한다.

1) 웹 로그 분석기를 이용하여 웹 로그 데이터와 쿠키(cookie) 정보를 조합하여 고객 별로 접근한 상품 및 상품 카테고리에 대한 정보를 텍스트 파일 형태로 출력한다. 이 데이터는 날짜별로 수집한다.
2) 출력된 텍스트 파일 형태의 고객 ID, 상품 ID 및 상품 카테고리 ID 등을 숫자형으로 변환시켜서 DB 테이블에 저장한다.
3) DB 테이블의 내용, 즉 일정 기간의 고객 접근(관심) 상품 정보 및 고객 접근(관심) 카테고리 정보를 학습 모듈 및 추천 모듈에서 활용할 수 있는 형태로 변환시킨다.

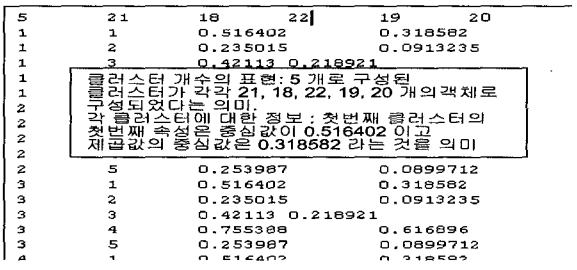
4.2.2 학습 모듈

학습 모듈에서 생성된 순차 패턴들을 결과 순차 DB(result sequential pattern database)라고 부른다. 학습 모듈에서 수행하는 이웃고객 결정과 순차 패턴 학습은 매일 일괄처리한다. 학습 모듈에서는 원본 로그 데이터 중에서 전반부 15일에 해당하는 데이터를 이용하여 이웃고객 결정 작업과 상품접근 순차 패턴 추출 작업을 수행한다. 50만 명의 고객 중 전반부 15일 동안 이용 로그가 남아있는 사용자는 약 30만 명이고, 상품접근 로그수는 약 400만개이다. 상품접근 순차 패턴 추출 작업은 고객군집 수를 5개로 정하고, k-means 클러스터링 알고리즘을 수행하여 각각의 고객군집 내에서 순차 패턴을 추출하였다. 이 과정에서 최소 지지도를 변경하면서 수행 시간을 측정하였다. <표 2>는 학습 모듈의 실험 대상을 나타낸다.

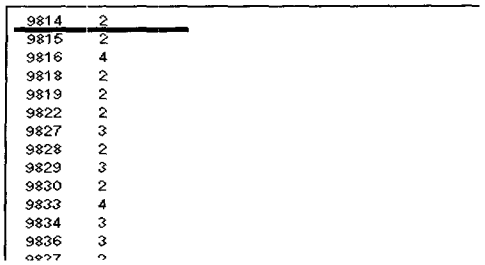
(그림 6)은 고객군집 결정을 위한 클러스터링의 실행 상태를 보여 준다. 30만 명의 고객에 대한 클러스터링을 수행하는데 6.812초가 소요되었다. (그림 6)에서 첫 번째 고객군집의 속성은 중심값이 0.516402이고 제곱값의 중심값이 0.318582임을 나타내고 있다.

〈표 2〉 학습 모듈의 실험 대상

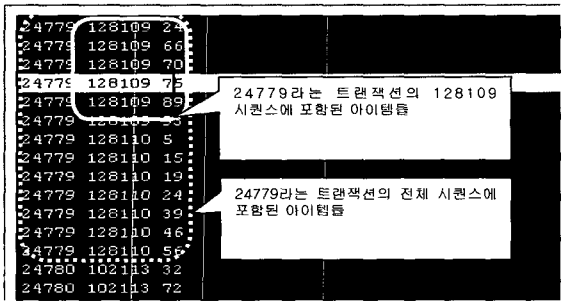
실험 대상	실험 데이터	측정 값
이웃 고객 결정	대상 고객 수, 초기 군집 수	수행 시간 및 군집 생성 수
순차 패턴의 학습 과정	최소 지지도	수행 시간 및 순차 패턴의 수



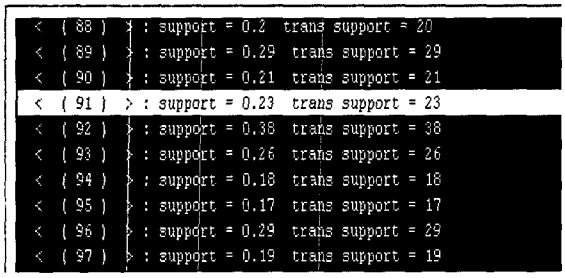
(그림 6) 고객군집 결과



(그림 7) 이웃고객 결정 결과



(그림 8) 순차 패턴 처리결과



(그림 9) 텍스트 형태로 나타난 순차 패턴 처리결과

실제 군집 결과는 (그림 7)과 같이 9814 고객이 2번째 군집으로 결정되어 있다고 나타낼 수 있다.

(그림 8)은 전체 고객에서 순차 패턴을 추출하는 실행 결과 화면이다.

(그림 8)에서 순차 패턴 추출에 사용된 최소 지지도가 0.05%이고 수행시간이 25069초(=약 7시간)라는 것을 보여주고 있다. 그리고 24779 고객이 66, 70, 75 등 13개의 상품에 대하여 관심기록을 남겼음을 알 수 있다. (그림 9)는 추출된 순차 패턴을 텍스트 형태로 나타낸 결과를 보여 준다.

(그림 9)에서 순차 패턴과 순차 패턴에 대한 지지도를 개수 기준과 비율 기준으로 나타낸다. 예를 들어, < (2) (71) > : support = 0.07 trans support = 7 의 의미는 상품 (2)를 구매한 사람이 일정 시간이 지나 상품 (71)을 구매하는 패턴이 있고 이런 사람이 전체의 7%(= 지지도)라는 것이다.

4.2.3 추천 모듈

추천 모듈에서는 상품추천 적합성에 대한 평가가 필요하다. 학습이 끝난 후, 고객의 수를 각각 100명, 1,000명, 10,000명, 100,000명으로 변경시키면서 추천에 소요된 평균 처리시간을 분석한다. 추천 결과의 타당성을 조사하기 위하여 고객 10,000명에 대하여, 추천한 상품수를 각각 1개, 5개, 10개로 변경시키면서 특정 고객에 대한 추천 결과와 해당 고객이 후반부 15일에 관심을 보인 상품 목록과의 비교를 통하여 추천의 정확성을 평가한다.

〈표 3〉 추천 모듈의 실험 대상

실험종류	변경 파라메터	평가척도
추천 속도	추천 결과 수	처리시간
추천 타당성	추천 결과 수	PRP, PRR, PF1, SCR

4.3 성능 평가

본 절에서는 실험 앞 절에서 설명한 학습 모듈 및 추천 모듈의 실험 결과를 이용하여 본 논문에서 제안한 추천 시스템의 성능을 평가한다.

4.3.1 학습 성능 평가

(그림 10)은 학습 모듈에서 실험대상 고객들로부터 고객군집을 추출할 때 소요되는 처리 시간을 나타낸다.

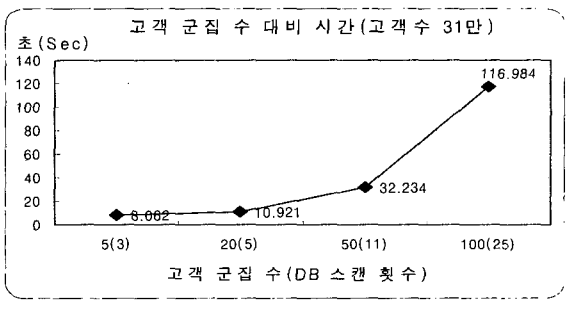
실험을 통하여 20개 이하의 고객군집 수 결정은 10초 이내에 처리되었고, 30여만 명 정도의 고객에 대한 군집 수 결정은 2분 이내에 처리되었다. 실험 평가를 위한 고객군집 ID에 따른 고객수와 상품관심횟수는 <표 4>와 같다.

〈표 4〉 고객군집 당 고객 수와 상품 관심 횟수

고객군집 ID	고객 수	상품 관심 횟수
1	8,522	58,588
2	224,743	2,586,351
3	39,887	337,242
4	37,013	784,984
5	2,953	32,267

(그림 11)은 고객군집 수가 <표 4>와 같을 때, 고객의 수 변화에 따른 고객군집의 추출 성능을 나타낸다.

(그림 11)에서 고객 수의 변화에 따른 고객군집의 추출시간은 거의 선형에 가깝다. 10만 명까지의 고객에 대한 고객군집 추출시간은 3초 이내에 가능하며, 30만 명의 처리에 10초 이내, 그리고 120만 명에 대한 처리도 40초 이내에 수행됨을 알 수 있다. 그리고 실험 결과, 고객 당 평균 실행시간은 0.03 msec 로서 추천 결과 수나 고객 수에 크게 영향을 받지 않고 있음을 알 수 있다.



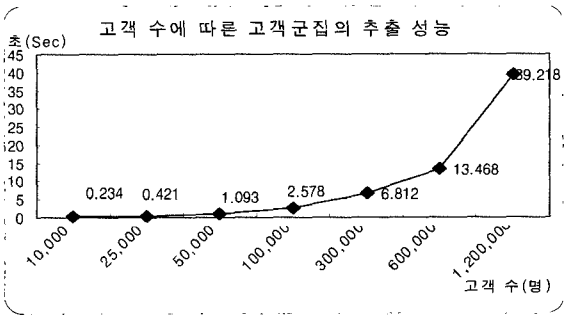
(그림 10) 고객군집 수 대비 추출 수행시간

순차 패턴의 실행시간은 지지도에 따라 급격히 변하는 특성을 가지고 있다. (그림 12)는 고객군집 ID 4의 환경에서 지지도 변화에 따른 순차 패턴 실행 성능을 나타내고 있다.

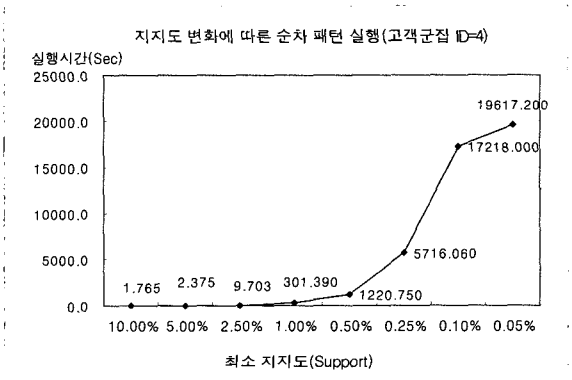
순차 패턴의 경우, 최소 지지도에 따라 수행 속도가 매우 심하게 변하는 특성을 가지고 있어서 실제 응용환경에 적용하기가 어렵다. 본 논문에서는 (그림 13)과 같이 순차 패턴의 패턴 수를 기준으로 최소 지지도를 자동 조절하여, 수행 속도를 선형 이하로 수행되도록 최적화하였다.

고객군집 수에 따라 지지도를 낮게 설정하면 비교적 안정된 수의 순차 패턴을 추출할 수 있다. (그림 13)과 같이, 결과 패턴수를 기준으로 최소 지지도를 자동 변경함으로써, 수행 시간이 급격히 안정화되고 있는 것을 알 수 있다.

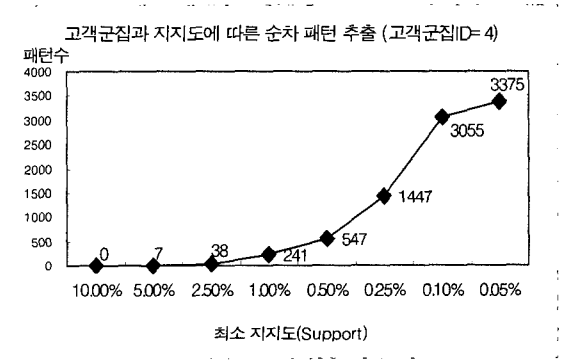
<표 5>는 고객군집수와 지지도에 따라 학습시킨 순차 패턴의 학습결과 데이터이다.



(그림 11) 고객 수에 따른 고객군집(ID 5)의 추출 성능



(그림 12) 지지도 변화에 따른 순차 패턴 실행 성능



(그림 13) 고객군집과 지지도에 따른 순차 패턴 추출 성능

〈표 5〉 고객군집별 순차 패턴 학습결과 데이터

고객군집ID=1				
최소 지지도 (%)	실행 시간 (sec)	패턴 수	최대 패턴 수	조정된 지지도
10.00%	0.234	2	1	10.000%
5.00%	0.218	6	1	5.006%
2.50%	0.390	19	2	2.509%
1.00%	2.047	74	2	1.006%
0.50%	9.953	213	3	0.509%
0.25%	54.016	657	4	0.260%
0.10%	390.078	3179	5	0.107%
0.05%	1688.590	10650	5	0.059%
0.03%	1720.000	10650	5	0.059%
고객군집ID=2				
최소 지지도(%)	실행 시간 (sec)	패턴 수	최대 패턴 수	조정된 지지도
10.00%	6.750	0	0	10.000%
5.00%	6.484	0	0	5.000%
2.50%	8.094	1	1	2.500%
1.00%	17.187	23	1	1.000%
0.50%	188.750	102	1	0.500%
0.25%	1877.170	332	2	0.250%
0.10%	37138.400	1516	3	0.100%
0.05%	62714.900	2157	3	0.080%
고객군집ID=3				
최소 지지도 (%)	실행 시간 (sec)	패턴 수	최대 패턴 수	조정된 지지도
10.00%	0.671	0	0	10.001%
5.00%	1.000	7	1	5.000%
2.50%	3.500	32	1	2.501%
1.00%	73.953	172	2	1.000%
0.50%	185.875	344	2	0.501%
0.25%	489.921	853	3	0.251%
0.10%	2530.480	4288	4	0.100%
0.05%	8902.250	14946	5	0.050%
고객군집ID= 4				
최소 지지도 (%)	실행 시간 (sec)	패턴 수	Max Pattern	조정된 지지도
10.00%	1.765	0	0	10.002%
5.00%	2.375	7	1	5.002%
2.50%	9.703	38	1	2.502%
1.00%	301.390	241	2	1.003%
0.50%	1220.750	547	2	0.503%
0.25%	5716.060	1447	3	0.251%
0.10%	17218.000	3055	3	0.157%
0.05%	19617.200	3375	3	0.149%
고객군집ID= 5				
최소 지지도(%)	실행 시간 (sec)	패턴 수	최대 패턴 수	조정된 지지도
10.00%	0.109	3	1	10.027%
5.00%	0.109	7	1	5.014%
2.50%	0.250	30	2	2.507%
1.00%	2.078	147	3	1.016%
0.50%	12.265	488	4	0.508%
0.25%	52.000	1599	5	0.271%
0.10%	727.735	17471	5	0.102%
0.05%	726.500	17471	5	0.102%

추천시스템에서의 실행속도는 필요한 데이터 구조를 로딩(loading) 하는 시간과 실제 추천에 필요한 시간으로 구분될 수 있다. 제안된 시스템에서 최소 지지도를 0.2%로 했을 때, 전반기 고객들의 이용 패턴을 주기억장치로 읽어 들이는 준비시간(Warm-up time)은 약 3.406초가 걸렸다. 최소 지지도를 자동으로 변경하여 순차 패턴 수행 시간이 너무 오래 수행되어 시스템의 안정성을 위협하지 않도록 하였는데, 자동 변경의 기준은 순차 패턴 중 길이가 1인 순차 패턴 개수의 최대값으로 하였다.

4.3.2 추천 성능 평가

<표 6>은 고객 수와 추천한 상품 수의 변경에 따른 추천 실행 속도를 나타낸다.

실험 결과, 10,000명의 고객에게 10개의 상품을 동시에 추천할 경우, 5초 미만의 시간에 추천이 가능하였다. 그리고 추천하는 상품의 개수에 따라 처리시간이 급격히 증가하지만 학습된 순차 패턴 수가 한정되어 있기 때문에 일정한 추천 개수에 도달하면 처리시간은 안정됨이 검증되었다. 추천 상품 수와 고객 수에 대하여 추천 실행 시간이 크게 영향을 받지 않고 있으며, 1인당 평균 추천 실행 시간이 0.5 msec 이하로 측정되었다. 또한 본 시스템에서는 초당 2,000개 이상의 동시 처리가 가능하였다.

<표 7>은 추천 시스템의 성능평가 과정을 나타낸다. 단계 1에서 순차 패턴 학습이 완료된 후, 단계 2-1에서는 추천 모듈을 반복적으로 수행하여 그 결과를 파일로 출력한다. 출력된 파일을 DB에 적재한 후에 단계 2-2와 단계 3을 수행한다.

〈표 6〉 고객 수와 추천 결과 상품 수의 변경에 따른 추천 실행 속도

실행 시간(sec) / 고객 수	추천 상품 1개	추천 상품 5개	추천 상품 10개
100	0.032	0.031	0.046
1000	0.407	0.437	0.438
10000	4.61	4.593	5.109
100000	41.782	43.172	44.657

〈표 7〉 추천 시스템의 성능평가 과정

단계 1	수집된 30일분의 로그 데이터 중에서 전반부 15일 데이터를 이용하여 이웃고객을 결정하고, 순차 패턴을 학습한다.
단계 2	후반부 15일 동안에 접근한 고객(100,000명)들에 대하여 다음 단계 2-1과 2-2를 반복 수행한다.
단계 2-1	추천 모듈을 통해 특정 고객에게 추천대상이 되는 5개의 상품을 선택한다.
단계 2-2	특정 고객이 후반부 15일 동안에 추천된 상품 5개에 대해 접근한 기록이 있는지를 확인한다.
단계 3	고객 100,000명에 대하여 다음과 같은 추천 성능 평가척도를 연산 한다 : PRP, PRR, PF1, SCR

<표 8>은 고객 100,000명에 대하여 추천 결과 상품수를 1개, 5개, 10개로 하였을 때의 추천 성능 평가 결과를 나타낸다.

〈표 8〉 추천 성능 평가 결과

추천대상상품수	PRP	PRR	PF1	SCR
1	3.7%	0.3%	0.5%	3.7%
5	1.9%	0.7%	1.0%	8.1%
10	1.5%	1.1%	1.3%	12.0%

추천대상 상품수가 증가함에 따라 PRP는 감소되었다. 이것은 추천하는 상품의 개수가 증가할수록 추천하는 상품에 포함된 상품을 고객이 볼 가능성이 적어지기 때문이다. PRP와는 달리 추천하는 상품 수가 증가함에 따라 PRR은 증가하였다. 이는 추천하는 상품의 개수가 늘어나는 경우 추천하는 상품 중 고객이 하반기 15일동안 접근한 상품이 포함되어 있을 가능성이 증가하기 때문이다. 추천 상품 수가 10개일 경우 PF1이 가장 우수한 것으로 검증되었다. 추천 상품 수가 증가하면 SCR도 점진적으로 증가하였다. 일반적으로 SCR이 증가하면 쇼핑몰 운영자 입장에서는 매우 바람직하다고 판단할 수 있다. 그러나 현실적으로 고객에게 추천할 수 있는 상품의 수는 제한되어 있다. 본 논문에서 제안하는 시스템의 경우, 추천 결과 상품수가 10개일 때의 SCR 값이 12%까지 높아졌다. 이것은 추천된 상품을 하반기 15일 동안 접근한 고객 10만 명중 1만 2천 명이 추천 결과에 속한 상품을 한번 이상 선택했다는 것을 의미한다. 이것은 높은 정확도임을 의미한다. 추천된 결과를 대형 쇼핑몰의 개인 로그인 화면을 통하여 고객들에게 노출한다면 SCR값은 비례적으로 높아질 것이다.

5. 결론 및 향후 연구 방향

최근까지 상품 추천시스템에 대한 많은 연구와 이에 대한 상용 시스템이 출시되고 있지만, 대형 온라인 쇼핑몰의 경우, 엄청난 크기의 웹 로그 데이터를 처리 할 수 있는 시스템에 대한 연구는 아직 없을 뿐만 아니라 매일 변화하는 고객의 관심도를 반영하는 추천 시스템에 대한 연구도 거의 전무한 상태이다.

따라서, 본 논문에서는 상품수가 수십만 종류 이상이고, 고객수가 수백만 명 이상인 환경에 적용될 수 있고, 매일 변화하는 고객의 관심도를 반영하며 카테고리와 상품 정보를 활용하는 상품 추천 시스템을 제안하였다. 제안된 상품 추천 시스템은 학습 모듈과 추천모듈의 두 가지 핵심 모듈과, 웹 로그를 전처리하는 기능으로 구성하였다. 학습 모듈은 기간에 따라 변화하는 고객의 관심도를 반영하기 위하여 매일 고객을 재 그룹화하고, 고객의 관심 상품 정보를 학습하였다. 이 과정에서 일정한 과거 데이터를 합산하여 반영하였고, k-means 클러스터링 기법을 사용하여 이웃고객을 미리 계산해두는 방법을 이용하였다. Match-Find 알고리즘

을 제안하여, 고객별 과거 관심 상품 내역과 학습 시스템에서 결정된 순차 패턴을 비교하여 추천 상품을 결정하였다.

실험은 사용자 기준으로 표본화 추출(샘플링)된 30일간의 웹 로그 데이터를 사용하였다. 50만 명의 사용자가 한 달간 접근한 모든 상품과 카테고리 정보를 수집하였으며, 수집된 대형쇼핑몰 데이터에는 100만개의 상품과 5만개의 카테고리가 있었다. 학습 모듈의 경우, 클러스터링을 통한 이웃고객 형성 단계는 고객 수 대비 선형에 가까운 속도 성능을 보였으며, 카테고리 접근 기록이 있는 30만 명의 사용자에 대하여 수분 내에 처리됨을 알 수 있었다. 순차 패턴 추출 과정은 최소 지지도 자동 수정 기능을 통하여 24시간 이내에 수행될 수 있음을 확인하였다. 또한 상품 추천 결과의 처리 성능을 분석한 결과, 실험 환경에서 초당 5천명 이상의 추천을 처리할 수 있음을 보였고, 추천 정확도 인수(PF1)값이 1.4%가 되고 고객 성공 비율(SCR)값이 12%에 접근함에 따라 추천 모듈의 성능이 우수함을 알 수 있었다. 실험에 사용된 데이터가 기존의 연구들과 속성이 다를 뿐만 아니라 데이터의 크기가 대용량인 새로운 환경에서의 실험이었기 때문에 기존 연구들과의 성능 비교를 할 수 없는 아쉬움이 있었다. 본 논문의 실험에서 정확도 검증을 위해 사용한 방법은 전반부 15일의 상품 관심 내역을 학습한 후, 목표 고객에게 추천을 수행하고 추천 결과와 해당 고객의 후반부 15일의 상품 관심 내역을 비교한 것이다. 따라서 실제 대형 온라인 쇼핑몰에서는 개인화 추천 뿐 아니라 캠페인 중인 상품, 신규 상품, 인기 상품 등을 조합하여 추천하는 방식으로 운영하고 있으므로 실제 환경에서는 추천 결과에 대한 성능 값이 높아질 것이다.

향후, 고객에게 전달할 수 있는 추천 상품의 수에 제약이 있는 모바일 환경의 경우, 이들 정보를 통합하여 추천할 수 있는 프레임워크의 연구가 필요하다. 또한 순차 패턴의 수행 속도가 학습 과정의 대부분을 차지하고 있기 때문에, 순차 패턴 수행 속도의 개선도 연구되어야 할 것이다.

참 고 문 헌

[1] 김재경, 안도현, 조윤호, "개인별 상품추천시스템, WebCF-PT: 웹마이닝과 상품계층도를 이용한 협업필터링", 경영정보학연구, 제15권, 제1호, pp.63-79, 2005.

[2] Minos Garofalakis, Rajeev Rastogi and Kyuseok Shim, " Mining Sequential Patterns with Regular Expression Constraints," In the Proceeding of the 2001 ACM SIGMOD International Conference ON Management of Data, 2001.

[3] 오용생, "시계열 데이터로부터 경향성을 이용한 순차 패턴의 탐색", 포항공과대학교 대학원 석사학위논문, 2001.

[4] Cho, Yoon Ho, Jae Kyeong Kim, and Soung Hie Kim, "A Personalized Recommender System Based on Web Usage Mining and Decision Tree Induction," Expert Systems with Applications, Vol.23, No.3, pp.329-342, 2002.

[5] Sarwar.B, Karypis.G, Konstan,J and Riedl.J, "Application of dimensionality reduction in recommender system - a case study," In Proceedings of ACM WebKDD-2000 Workshop. pp.285-295, 2000.

[6] Mulvenna.M.D, S.S.Anand, A.G.Buchner, "Personalization on the Net Using Web Mining," Communication of the ACM, Vol.43, No.8, pp.122-125, August, 2000.

[7] Kim,Jong Woo, Byung Hun Lee, Michael J.Shaw, Hsin-Lu Chang, Mathew Nelson, "Application of Decision Tree Induction Techniques to Personalized Advertisments on the Internet Storefront," International Journal of Electronic Commerce, Vol.5, No.3, pp.45-62, Spring, 2001.

[8] Mobasher, Bamshad, Robert Cooley, and Jaideep Srivastava, "Automatic Personalization Based on Web Usage Mining," Communication of the ACM, Vol.43, No.3, pp.142-151, 2000.

[9] 황병연, "개선된 추천을 위해 클러스터링을 이용한 협동적 필터링 에이전트 시스템의 성능", 정보처리논문지, 제7권, 제5S호, pp.1599-1608, 2000.

[10] J.B. Schafer, J.A. Konstan, and J. Riedl , "Meta-recommendation Systems: User-controlled Integration of Diverse Recommendations," In Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM 2002), McLean, VA, pp.43-51, November, 2002.

[11] F. Masseglia, P. Poncelet and M. Teisseire, "Incremental Mining of Sequential Patterns in Large Databases," Actes des 16imes Journes Bases de Donnes Avances (BDA'00), Blois, France, October, 2000.

[12] 이경희, 한정혜, 임춘성, "지수적 가중치를 적용한 협력적 상품추천시스템", 정보처리학회지, pp.625-632, 2001.

[13] Feng-Hsu Wang, Hsiu-Mei Shao, "Effective personalized recommendation based on time-framed navigation clustering and association mining," Expert Systems with Applications, pp.365-377, 2004.

[14] Soe-Tsyr Yuan, Chiahsin Cheng, "Ontology-based personalized couple clustering for heterogeneous product recommendation in mobile marketing," Expert Systems with Applications 26, pp.461-476, 2004.

심 장 섭

e-mail : sjs@iita.re.kr

1986년 한양대학교 전자공학과(학사)

1992년 한양대학교 전자공학과(석사)

2005년 충북대학교 컴퓨터공학과 (박사)

2001년 데이콤 종합연구소 지능망개발팀장

2003년 ㈜위즈정보기술 연구소장/이사

2003년~현재 정보통신연구진흥원 정보화추진팀장

관심분야 : DBMS, Recommender Systems, 임베디드 SW, 통신 프로토콜 등

우 선 미

e-mail : smwoo@chonbuk.ac.kr
1991년 서남대학교 전자계산학과(이학사)
1995년 전북대학교 전산통계학과 (이학석사)
2001년 전북대학교 전산통계학과 (이학박사)
2001년 3월~2003년 11월 전북대학교 중앙도서관 전산실 조교
2003년 12월~2006년 2월 전북대학교 BK21전자정보사업단 기금교수
2006년 3월~현재 전북대학교 전자정보공학부 시간강사
관심분야 : 사용자 위주의 정보검색, 문서순위결정, 정보 필터링, XML 응용, 디지털 도서관 등

김 용 성

e-mail : yskim@chonbuk.ac.kr
1978년 고려대학교 수학과(이학사)
1984년 광운대학교 전산학과(이학석사)
1992년 광운대학교 전산학과(이학박사)
1985년~현재 전북대학교 전자정보공학부 교수
1996년~1998년 1월 한국학술진흥재단 전문위원
관심분야 : XML, 사용자 중심의 정보검색, 다중 사용자 인터페이스, W3C 웹 서비스 등

이 동 하

e-mail : dongha@nethru.co.kr
1991년 한국과학기술원 전자전산학과(학사)
1993년 포항공과대학교 전자계산학과(석사)
2000년 포항공과대학교 전자계산학과(박사)
1996년 3월~2001년 6월 포항공과대학교 정보통신 연구소 위촉연구원
2001년 7월~현재 (주) 넷스루 데이터 마이닝 연구소 연구소장
관심분야 : Recommender Systems, Web Business Intelligence, Web Log Mining, Efficiency Issues and Applications of Association Rules/Sequential Pattern

정 순 기

e-mail : soonkey@cbnu.ac.kr
1982년 Dortmund대학 전산학과, Dipl.Inf.
1994년 Groningen대학 전산학과, Dr.
1985년~현재 충북대학교 전기전자컴퓨터공학부 교수
1994년 충북대 전자계산소장
1998년 한국과학재단 한독기초과학협력위원회 정보분과위원장
2000년 충북대학교 도서관장
2005년~현재 KISTI 지역자문교수 겸 협의회위원
관심분야 : DBS, 실시간 시스템, 소프트웨어공학