대용량 개인화 실시간 상품 추천 시스템 설계

김종희*, 심장섭**, 이동하***, 정순기****
*충북대학교 컴퓨터 공학과 e-mail:paper@sunmoon.ac.kr
**정보통신연구진흥원(IITA) 정보화추진팀 e-mail:sjs@iita.re.kr
***(주)넷스루 dongha@nethru.co.kr
****충북대학교 컴퓨터 공학과 e-mail:soonkey@cbnu.ac.kr

Design of a Large Real-Time Personalized Recommendation System

Jong-Hee Kim*, Jang-Sup Shim**, Dong-Ha Lee***, Soon-Key Jung*
*Dept of Computer Engineering, Chungbuk National University
**Information Promotion Team Institute for Information Technology Advancement(IITA)
***Data Mining Laboratory, Nethru Incorporation

요 약

최근 대용량 추천시스템에 대한 필요성이 증가하고 있고, 특히 대규모 인터넷 쇼핑몰을 위한 개인화 추천 시스템 구조에 대한 관심이 높아지고 있다. 본 논문에서는 k-means 클러스터링과 순차 패턴 기법을 이용한 인터넷 쇼핑몰 상품 추천 시스템을 설계 및 구현한다. 사용자 정보의 일괄처리와 카테고리의 계층적 특성을 반영하면서 데이터 마이닝 기법을 활용하여 개인화된 추천 엔진을 대형 시스템에서 동작하도록 설계 하였다. 설계 구현한 시스템의 평가를 위해, 대형 쇼핑몰의 데이터를 이용하여 추천 예측 정확율(PRP: Predictive Recommend Precision), 추천 예측 재현율(PRR: Predictive Recommend Recall), 정확도 인수(PF1 : Predictive Factor One-measure)를 구하였다.

1. 서론

e-비즈니스의 활성화로 인하여 수많은 인터넷 쇼핑몰 업체가 등장하였으며 극심한 경쟁에서 생존하기 위한 고객 관계 관리의 노력도 계속 되고 있다[2]. 또한 고객의 취향이나 관심에 초점을 맞추어 고객에게 상품이나 컨텐츠를 제공하는 고객맞춤(customization) 전략이 온라인 쇼핑몰이나 정보서비스 제공자에게 있어서 성공을 위한 필수적인 요소가 되고 있다[1].

전자상거래의 보편화에 따라 온라인 쇼핑몰에서 구매자가 최선의 구매를 할 수 있도록 선호도를 고려하여 원하는 상품들을 구매하기 위하여 많은 상품 정보를 처리해야만 하는 부담이 발생한다. 이와 같이 고객이 처리해야할 정보 과부하 현상은 고객의 구매 의욕까지 상실 시킬 수 있는 요인이 되고 있다. 이러한 문제점을 해결하기 위하여 고객이 처리해야할 정보의 양을 감소시키고, 고객에게 개인화된 서비스를 제공하는 시스템을 개인화된 상품추천 시스템(personalized product recommendation system)이라고 부른다. 기존 추천 시스템은 이웃고객 세그먼트의 추출과 추천후보 상품의 탐색에 포괄적인 연산 방법을 사용하기 있기 때문에 대형 인터넷 쇼핑몰의 상품 추천 시스템에서는 적용 할 수 없다.

본 논문에서는 대용량에서의 데이터 처리를 위해 순차 패턴과 k-means 클러스터링 기법을 활용하고, 고객이 선호하는 상품들을 정렬하여 지지도가 높은 상품을 추천하기 위하여 Match-Find 알고리즘을 제안하여 적용한 웹로그데이터를 가지고 상품 패턴의 학습 성능, 상품 추천 성능 및 추천 타당성을 평가하며, 평가 결과를 기초로 고객들에게 상품 추천정보를 제공할 수 있는지 검증한다.

본 논문은 2장에서 기존 추천시스템에서 사용하고 있는 데이터 마이닝기법들을 설명하고 3장에서는 VLDB기반의 추천시스템에서 사용되는 데이터들의 구조에 대해서 설명한다. 4장은 제안하는 상품 추천 시스템의 실험환경과 성능평가를 기술하고 5장에서는 결론과 향후 연구과제를 제시한다.

2. 관련 연구

2.1 데이터 마이닝

데이터 마이닝이란 대량의 실제 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다. 데이터 마이닝 기법에는 여러 가지가 있지만 본 논문에서 사용한 기법은 다음과 같다.

2.1.1 순차패턴(Sequential Patterns)

순차 패턴은 한 트랜잭션 안에서 발생하는 항목들 간의 연관규칙에 시간이 변이를 추가한 것이다.[2] 순차 패턴에서는 주어진 트랜잭션 데이터베이스에서 사용자가 정의한 최소 지지도를 만족하는 모든 시퀀스들 사이에서 최대 시퀀스를 찾는 것이다. 여기서 시퀀스는 시간에 따라 정렬된 트랜잭션들의 리스트를 말한다.[7][9]

2.1.2 클러스터링(Clustering)

클러스터링 기법은 임의의 데이터 집합으로부터 서로 유사한 속성을 가지는 데이터의 군집(Cluster) 또는 세그먼트(Segment)를 추출하는 기법을 의미한다,

클러스터링의 대상이 되는 객체(Object)들은 각 객체의 특성을 나타내는 속성을 가지고 있다. 객체들은 클러스터링을 통해서 특정 군집에 속하게 되며, 각군집은 소속 객체들의 속성정보를 소유한다. 객체에 대한 클러스터링 결과를 분석하면 각 군집에 분포된 객체들의 분포도에 대한 정보를 얻을 수 있다. 본 논문에서는 유사한 상품 선호도를 갖는 고객들을 클러스터링하기 위해 k-means 클러스터링 기법을 이용하였으며, 거리 기능 척도로는 맨하탄 거리 척도를 사용 하였다.

2.2 추천시스템

추천 시스템이란 고객이 원하는 상품 또는 컨텐츠를 제공하여 상품에 접근이 용이하게 하는 시스템이다. 현재 협업 필터링과 속성기반 필터링 기법이 일반적으로 이용된다. 하지만 협업 필터링은 가장 널리 쓰이고 있음에도 불구하고, 데이터의 희소성 문제가 있고, neighbor 계산에 따른 소요시간의 문제로 대용량의 환경에 적용이 어렵다. 다른 추천 방법들도 유사하게 대용량 처리에 취약한 점들을 가지고 있다.

이러한 단점을 보완하기 위해, 데이터 마이닝 기법과 혼합되어 사용되기도 했지만, 실제 대용량에서 연구된 예가 드물고, 단순히 association rule을 적용한 것에 불과하다.

3. 대용량 DB기반의 실시간 추천 시스템 설계

본 논문에서 제안하는 추천 시스템의 구조는 학습 모듈, 추천 모듈 및 평가 모듈로 구성 되며 과정은 다음과 같다. 첫째, 학습 모듈에서는 목표 고객에 대한 이웃고객(neighborhood)을 추출한다. 이웃고객은 웹 로그로부터 고객들이 선호하는 상품 및 상품 카테고리 정보를 기준으로 k-means 클러스터링 알고리즘에 적용하여 이웃고객을 추출한다. 둘째, 생성된 이웃고객(고객군집)으로부터 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.

3.1 로그데이터와 상품 카테고리

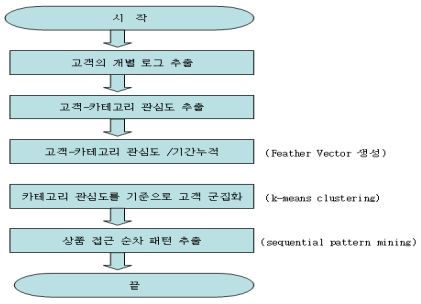
고객이 대형 인터넷 쇼핑몰을 이용하여 상품을 구매할 경우 추천 시스템은 고객의 모든 DB 접근정보를 웹 로그에 기록 및 유지 관리하며, 웹 로그에 저장된 데이터를 분석하여 고객에게 유용한 상품 정보를 제공한다. 추천 시스템은 상품 수가 100만 개 이상, 상품 카테고리 수가 5만개 이상 그리고 대상 고객이 천만 명 이상이 되는 대형 쇼핑몰을 전제로 하여 설계한다. 고객의 신상정보 대신에 상품 카테고리에 대한 고객의 관심도를 이용하여 고객들을 클러스터링한다. 고객들의 클러스터링에는 k-means 클러스터링 기법을 사용하므로 고객군집으로부터 고객이 가장 선호하는 상품들의 순차 패턴을 추출할 때 소요되는 시간 복잡도를 O(N)에 가깝게 감소시킬 수 있다. 제안하는 추천 시스템은 상품과 컨텐츠 모두를 취급한다. 상품들은 특징에 따라 특정한 카테고리에 속하게 되며 상품 카테고리는 단일 계층구조(single hierarchy)의 트리구조를 갖는다.

3.2 상품 선호도 추출

웹 로그 DB의 스키마 는 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. count는 상품 또는 카테고리에 대한 고객 선호도를 나타내며, 이것은 웹 페이지의 힛팅 수를 의미한다. 본 논문에서 제안하는 추천 시스템에서는 특정 상품이 소속된 카테고리의 힛팅 수 계산에 상품과 카테고리 정보를 우선적으로 적용한다.

3.3학습모듈

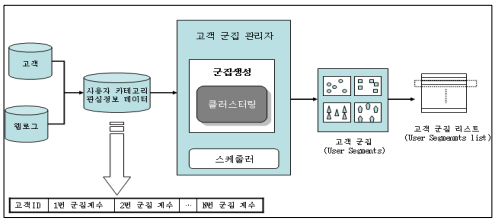
학습 모듈에서 일괄처리(batch processing) 방법으로 처리되는 주요 작업과 수행과정은 (그림 1)과 같다.



(그림 1) 학습모듈의 작업실행 순서

3.3.1 이웃고객 추출

웹로그로부터 이웃고객을 추출하는 과정은 그림2와 같다.



(그림 2) 웹 로그로부터 이웃고객 추출

3.3.2 순차 패턴 추출

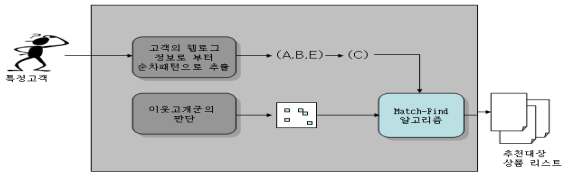
순차 패턴 마이닝 기법을 이용하여 고객 군집으로부터 접근 빈도수가 높은 상품의 순차 패턴을 추출한다. 순차 패턴 마이닝 기법을 이용하여 고객군집으로부터 상품접근 순차 패턴을 추출하는 알고리즘은 (그림 3)과 같다.

```
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 패턴을 추출하여 SeqDB에 저장한다.
end.
```

(그림 3) 순차패턴 알고리즘

3.4 추천모듈

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 상품 접근 패턴과 고객이 소속된 고객군집으로부터 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. 추천 모듈에서 상품의 추천과정은 (그림 4)와 같다.



(그림 4) 추천 모듈에서 상품의 추천과정

크러스터링 과 순차패턴을 활용하여 각 고객 군집별로 추출된 패턴실행 시간을 활용하여 실시간 개인화 추천을 수행하는 Match-Find 알고리즘은 (그림 5)와 같다.

```
입력 : 고객 u
출력 : p 개의 상품
begin
1. 고객 u의 상품 접근 패턴 Su를 웹 로그로부터 추출한다.
2. 고객 u가 속하는 고객 군집 i를 결정한다.
3. for(SeqDBi 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

상품 추천 시스템의 실험 환경과 성능 평가를 기술한다.

4.1 실험 환경

논문에서 제안한 추천 시스템의 실험을 위한 주요 모듈 구현 환경은 <표 1>과 같다.

<표 1> 모듈의 구현 환경

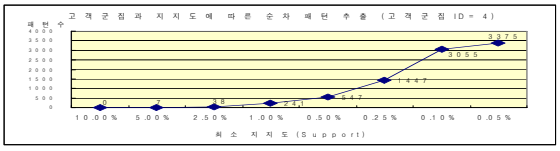
전처리 모듈	학습모듈	추천모듈	평가모듈
WiseLog Premium™, SQL Server 2000™	Visual C++™	Visual C++™	SQL Server 2000™

4) 실험 데이터 : 현재 운영중인 대형 인터넷

4.2 학습 성능평가

학습 모듈에서 실험대상 고객들로부터 고객군집을 추출할 때 소요되는 처리 시간을 나타낸다. 실험을 통하여 20개 이하의 고객군집 수 결정은 10초 이내에 처리 되었고, 30여만 명 정도의 고객에 대한 군집 수 결정은 2분 이내에 처리되었다.

순차 패턴의 경우, 최소 지지도(minimum support)에 따라 수행 속도가 매우 심하게 변하는 특성을 가지고 있어서 실제 응용환경에 적용이 어렵고 (그림 7)와 같이 순차 패턴의 패턴 수를 기준으로 최소 지지도를 자동 조절하여, 수행 속도를 선형 이하로 수행되도록 최적화하였다.



(그림 7) 고객군집과 지지도에 따른 순차 패턴 추출

실험결과 추천의 속도에 대한 결과는 다음과 같다.

<표 6> 고객 수와 추천 결과 상품 수의 변경에 따른 추천 실행 속도

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

추천의 결과가 낮은 값을 보이고는 있으나, 이것을 일반적인 추천 시스템과는 다른 척도임을 고려해야한다. 추천이 고객 행동의 변화에 미친 영향을 분석한 것이 아니라, 추천 결과가 실제 사용자의 행동을 얼마나 예측했는가의 척도이기 때문이다. 12% 의 고객이 추천 결과 중 하나 이상 의 상품에 관심을 보였다는 것은 상당한 예측력이라고 판단된다

5.결론

성능 평가 결과 학습 모듈의 경우, 클러스터링을 통한 이웃고객 형성 단계는 고객 수 대비 선형에 가까운 속도 성능을 보였으며, 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. 향후, 고객에게 전달할 수 있는 추천 상품의 수에 제약이 있는 모바일 환경의 경우, 이들 정보를 통합하여 추천할 수 있는 프레임워크의 연구가 필요하다.

참 고 문 헌

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

[2] 김경재, 김병국, "데이터 마이닝을 이용한 인터넷 쇼핑몰 상품 추천 시스템," 한국 지능 정보 시스템학회논문지 제11권 제1호, pp191-205, 2004

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] "Hyun-Wha Choi, Dong-Ha Lee and Jeon-Young Lee, "Multi-Level Linear Location Tree for Efficient Sequential Pattern Mining," Key Engineering Materials, 2005, Volumes. 277-279 (2005), Part1, page 369-374"