전자상거래에 적용 가능한 시간 연관 규칙 탐사 기법†

°서성보, 이준욱, 김선철, 류근호
충북대학교 컴퓨터과학과

Temporal Association Rule Mining on Electronic Commerce

Sung Bo Seo, Jun wook Lee, Sun cheul Kim and Keun Ho Ryu
Dept. of Computer Science, Chungbuk National University
E-mail Address : {sbseo,junux,sckim,khryu}@dblab.chungbuk.ac.kr

요 약

전자상거래가 활성화되었지만 현재의 쇼핑몰은 단순한 상품 구매 역할과 정보 제공, 그리고 정적인 웹 공간의 관리로 이를 통해 인터넷 쇼핑몰 상품의 효율적인 진열, 상품 연관성을 고려한 마케팅 전략, 고객관리와 웹 페이지간의 시간 연관성을 고려한 링크 정보 등과 같은 효율적인 마케팅 전략을 구사하기에는 한계가 있다. 이 논문에서는 전자상거래에 적용 가능한 시간 연관 규칙 탐사 기법을 통해 구매 데이터를 기반으로 상품간의 시간 연관 규칙 탐사와, 웹 서버에서 수집된 웹 로그 파일로부터 연관 규칙을 찾아내기 위한 모델을 제시한다 또한 이를 통해 생성된 규칙을 해석하여 사용자에 따른 웹 공간 구성, 방문 페이지간의 연관성을 고려한 마케팅 전략과 효율적인 광고 전략 등을 위한 방안을 제시한다.

1. 서론

데이터베이스 마케팅이란 고객에 대한 여러 가지 정보를 컴퓨터에 의해 데이터베이스화하고 구축된 고객 데이터베이스를 전략적으로 활용하여, 고객 개개인과의 접촉을 통해 직접적인 반응/판매를 유도하거나 장기적인 일대일 관계를 구축하고자 하는 제반 마케팅활동을 말한다. 인터넷을 통한 전자상거래가 활성화됨에 따라 현재 고객의 특성을 정확히 파악하여, 고객 개인의 관심에 부합하는 개인화 된 정보나 상품 서비스를 제공함으로써 고객만족을 극대화하기 위해 이미 구축된 고객 데이터베이스 및 거래 데이터를 활용한 데이터베이스 마케팅 도구들이 많이 제시되었지만 현재 구축된 전자상거래 시스템 상에서 상품 판매 향상과 직접적으로 관계되어 특성화된 고객 및 상품관리의 연계는 소홀하다 또한 현존하는 대부분의 웹 분석 도구들은 사용자의 웹 접근 활동에 대해 사용자 방문 IP주소, 방문 횟수, 방문시간과 사용자의 URL경로의 기본적인 정보만을 제공한다.

이 논문에서는 다음의 두 가지 방안을 제시하고, 이 방안에 사용할 수 있는 시간 연관 규칙 탐사 기법을 제안한다. 첫째, 상품의 과거 매출 실적에 관한 기록데이터로부터 상품간의 연관성 정도를 측정하여 이를 기반으로 마케팅 전략을 수립하기 위한 방안을 제시한다 둘째 웹 서버의 로그 파일의 정보를 분석하여 사용자의 접근 경향, 웹 공간 구성, 자주 방문하는 페이지와 구매 정보를 바탕으로 제품간의 연관성을 고려한 마케팅 전략 수립과 광고를 효율적으로 할 수 있는 방안을 제시한다

상품간의 연관성과 웹 로그 파일의 분석에 따라 쇼핑몰 상품 진열, 카탈로그 구성과 상품을 패키지화 하는 것은 좋은 예가 될 수 있으며 마케터는 고객과 상품 또는 웹 페이지간의 연관성에 따라 특정 상품을 구입한 고객(군)에 따라 적합한 마케팅 전략을 사용할 수 있다.

† 이 논문은 한국전자통신 연구원의 "통합 우정 물류 실시간 관제 시스템 개발" 사업의 일부 연구비 지원에 의해 수행되었음.

2. 관련연구

연관 규칙을 찾는 문제는 [Agra94]에 소개되어 실제 거래 데이터베이스에 기록된 판매 데이터의 분석에 적용되고 있다. 항목들의 집합으로 구성된 트랜잭션이 주어졌을 때 연관 규칙은 X->Y의 형태로 표현되며, 이는 X와 Y가 항목들의 집합이고 하나의 트랜잭션에 X와 Y가 동시에 존재한다는 것이다. 마이닝 연관 규칙 탐사의 문제는 사용자가 지정한 최소 지지도와 신뢰도를 만족하는 모든 연관 규칙을 찾기 위한 시도로 정의한다[Agra94].

Agrawal과 Srikant에 의해 Apriori, AprioriAll, Aprio-Hybrid와 같은 기법 등이 연관 규칙을 탐사하기 위해 제시되었고, 이후 수행향상을 위해 많은 연구가 진행되어져 왔다 [Chen98]에서는 시간 연관 규칙에 대해 기존의 트랜잭션에 관련된 시간 요소 이외에 절대시간과 주기시간을 언급하는 시간 연관 규칙이 소개되었다. 또한 연관 규칙 탐사를 이용한 사용자 패턴 분석 기법에 대한 연구[Cool99,Osma98]는 웹 로그 파일의 순차적인 데이터를 바탕으로 최대 시간 간격으로 트랜잭션을 처리한 후 각 트랜잭션에 연관 규칙 알고리즘을 적용하여 연관 규칙을 생성하게 된다.

[이강태99]에서는 시간 지원 데이터베이스를 대상으로 연관규칙 탐사를 위한 시간 연관 규칙 탐사 기법 및 탐사과정 그리고 제안된 알고리즘을 설명하고 있다.

3. 연관 규칙 탐사 모형과 처리과정

전자 상거래에 적용 가능한 연관 규칙 탐사 기법의 모형과 처리과정은 [그림 1]과 같다. 단계별 처리과정으로 먼저 사용자는 웹 브라우저를 통해 쇼핑몰에 접속한 후 자신의 계정과 비밀번호를 입력하여 인증을 받게 된다. 연관 규칙 분석기는 규칙 관리기와 분석기의 입력 값을 마이닝 탐사 모듈에 입력으로 넘겨주게 되며 탐사 모듈은 고객, 구매와

상품 데이터를 이용해 상품 연관 규칙을 탐사하게 된다 또한 전처리 과정을 거친 웹 로그 파일의 데이터베이스와 실제 구매한 데이터의 연계성으로 각 개인의 특성 및 연관성을 탐사하며 개인에게 적합한 동적인 웹 서비스와 순차적인 웹 사용자 패턴의 결과를 생성하여 웹 상에 제시한다.

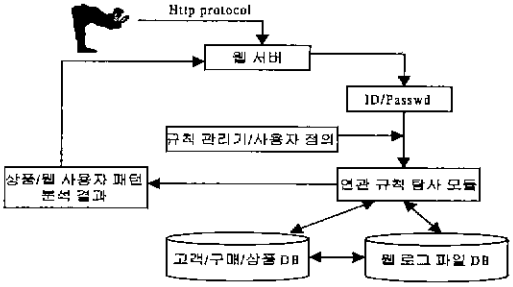

그림 1. 연관 규칙 탐사 모형과 처리과정

4. 시간 연관 규칙을 위한 데이터 모델

이 장에서는 상품간의 연관 규칙 탐사 모델과 시간 연관 규칙을 탐사를 위한 웹 사용자 패턴 분석 모델과 처리과정을 제시한다.

4.1 상품간의 연관 규칙 탐사

연관 규칙 탐사를 위해 T를 기본적인 시간 간격의 집합이라 가정하며 이를 시간영역이라 부르고 $I=(i_1,i_2,...i_m)$는 항목이라 불리는 집합이다. D를 타임 스탬프 트랜잭션의 집합이라 가정할 때 각 타임 스탬프 트랜잭션 S는 3개의 튜플 <tid, itemset,timestamp>로 구성된다. S.tid는 트랜잭션 구별자이고 S.itemset ⊆ I와 같은 항목의 집합이며 S.timestamp는 트랜잭션 S가 S.timestamp ⊆ T에 적용되는 기본시간 간격이다. 항목 집합에 항목의 개수는 이 항목의 크기로서 지정되어 지며 항목 X⊂I는 집합을 항목집합이라 한다. 크기 K의 항목집합을 K-itemset이라 부르며 만약 X⊆S.itemset된다면 트랜잭션 S는 itemset을 포함한다 ρ를 시간 표현의 집합이라 가정하면 $\forall p \in \rho$에 대해 $\Phi(p)=\{p_1,p_2,...p_n\}$의 해석은 시간 간격 집합의 표현이다.

정의 1: 시간 연관 규칙은 <R,P>의 쌍이며 R은 항목 X,Y에 대해 X⊂I, Y⊂I 이며 X∩Y=Ø일 때 X->Y의 형태로 표현하며 이때 P∈ρ 되는 시간 표현 집합이다.

정의 2: 지지도는 연관 규칙 X->Y의 지지도 S를 말하며 N을 전체 항목의 개수라 할 때 S= | XUY | /N로 표현하며, 신뢰도 C= | XUY | / | X | 로 표현한다

시간 연관 규칙 탐사는 시간 도메인 T와 시간 표현 P상에서 시간 간격 트랜잭션 집합 D가 주어졌을 때 최소 지지도와 신뢰도를 만족하는 D에 대해 <X->Y(S,C),P>의 모든 규칙을 발견하는 것이다. 연관 규칙 탐사 알고리즘은 Agrawal과 Srikant에 의해 표현된 방법과 유사하게 2단계로 구분된다. 단계 1에서는 데이터베이스로부터 발생하는 항목 중에서 주요 1항목을 찾는 것이며 다른 시간 간격 사이에서 후보 1항목에 대한 발생 빈도수를 계산한 후 주요 1항목으로서 최소 지지도와 신뢰도를 만족하는 항목들을 선택한다. 단계 2에서는 주요 k항목의 집합은 Apriori-gen 알고리즘을 이용하여 k-1의 주요 항목으로부터 최소 지지도와 신뢰도를 만족하는 주요 k항목의 집합을 생성해 내며 이 단계는 새로운 주요항목 집합이 발견되지 않을 때까지 반복한다.

4.2 웹 사용자 패턴 분석

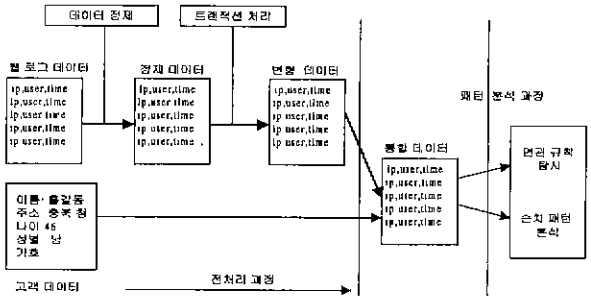

그림 2. 웹 사용자 패턴 분석 모형과 처리 과정

대부분의 웹 접근 로그는 HTTP프로토콜로 규정한 Common Log Format을 따른다. 이 표준에 따르면 로그 항목에는 사용자의 IP주소, 사용자 ID, 접근 시간, 요구방법, 접근된 페이지 URL, 프로토콜 등이 포함되어 진다. 웹 사용자 패턴 분석 과정은 [그림 2]와 같으며 각 단계별 처리과정을 알아보면 먼저 전처리 과정에는 데이터 정제 과정과 트랜잭션 처리 과정으로 분리되어 진다 로그 데이터에 데이터 마이닝 알고리즘을 적용하기 위하여 전처리 과정은 필수적이다. 원시 데이터에는 사운드, 비디오, CGI 실행 파일 등 불필요한 항목이 존재하며 이는 실제적인 마이닝에 적합한 속성이 아니므로 이를 제거하기 위한 정제 과정과, 개별적인 페이지를 연관 규칙과 같은 패턴 분석에 대해 의미있는 트랜잭션 그룹 단위로 구분 짓기 위해 트랜잭션 처리과정을 가진다. 이 처리과정을 위해 로그 엔트리는 IP주소를 기초로 하여 타임 스탬프 간에 다양한 시간 단위(granularity)와 최대 시간 간격을 고려해 처리하였다. L을 서버 접근 로그 엔트리 집합이라고 한다면, 하나의 로그 엔트리 l은 $l \in L$되면 다음과 같은 항목 IP주소: l.ip, 사용자 ID: l.uid, 접근한 페이지 URL: l.url, 접근 시간: l.time을 가진다

정의 3: 하나의 트랜잭션 t는 하나의 로그 엔트리 그룹 g로부터 3개의 항목으로 구성되며 트랜잭션은 다음과 같은 최대 간격 시간 또는 사용자 정의 시간으로 구분한다.

$t = < ip_g, uid_g, \{ l^g_1.url, ..., l^g_m.url \} >$ $1 \le k \le m$일 때 최대 시간 간격 Δt는 $l^g_{k+1}.time - l^g_k.time \le \Delta t$로 정의한다.

전처리 과정 후 도메인 이름을 주 키로 하여 이후 사용자 별, 항목별로 각각을 다시 정수형 코드로 1:1 매핑 하여 [표 1]과 같은 두 개의 매핑 테이블과 이를 토대로 트랜잭션 처리와 타임 스탬프 속성을 추가하여 일반화된 마이닝 알고리즘에 적용 가능한 형태의 파일을 형성한다. 재구성된 로그 파일은 [정의 1]과 [정의 2]의 표현과 앞에서 제시된 연관 규칙 탐사 알고리즘 단계를 수행하여 최소 지지도와 신뢰도를 만족하는 웹 페이지간의 연관성을 찾게 된다.

표 1 매핑 테이블과 재구성 된 로그 파일

매핑 코드	사용자 IP
1	210.104.41.10
2	203.245.8.22
3	210.104.41.11

매핑 코드	웹 페이지 항목
1	/store/food/a.html
2	/store/food/b.html
3	/store/cloth/c.html

트랜잭션 ID	사용자 식별자	웹 페이지 식별자	시간
100	10	20	03/Apr/99:12:36:40
100	10	47	03/Apr/99:12:37:20
101	12	30	04/Apr/99:09:20:10

5. 시간 연관 규칙 탐사기 구현

이 논문에서는 시간 연관규칙 탐사를 위해 Agrawal의 Apriori와 AprioriAll을 변형 구현하였다. 실험 데이터로는 상품 거래 데이터와 인터넷 쇼핑몰 서버의 웹 로그 파일을 적용하였다. 자바 서블릿을 이용해 웹 환경에서 구현하였고 오라클 데이터베이스와 연결을 위해 2-tier JDBC 드라이버를 이용하였다.

5.1 상품 연관 규칙 탐사기

상품간의 연관성 탐사를 위한 데이터 셋을 생성하기 위해 고객과 상품, 구매 테이블을 구축하였으며, 구매 테이블을 이용하여 고객의 판매 형태에 대한 자료를 지지도 15%와 신뢰도 50%로 분석하여, 아래 [그림 3]과 같은 결과를 얻을 수 있었다.

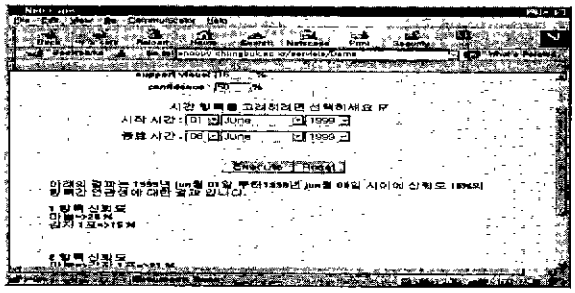

그림 3. 상품 연관 규칙 탐사 결과

상품의 과거 매출 실적에 관한 기록 데이터로부터 상품간의 연관성 정도를 측정하여 이를 기반으로 마케팅 전략을 수립할 수 있으며, 상품과 관련하여 분석가는 다음과 같은 사항에 관심을 가질 수 있다.

a) "쇼핑몰의 상품을 어떻게 배치할 것인가 ?"
b) "계절 및 특정 시간대의 카탈로그를 구성 효율적인 DM 발송 방법은 없는가 ?"
c) "어떤 상품을 판매하는 곳에 적합한 광고 형태는 무엇인가 ?"

위 물음에 대해 규칙 탐사기는 지지도와 신뢰도가 높은 상품의 진열 위치를 짧게 함으로서 동시 구매가 가능하도록 유도할 수 있으며, 특정 시간이나 계절 기간에 대한 연관 규칙을 찾아내 높은 연관성을 갖는 상품을 패키지화 하거나 카탈로그를 작성해 소비자에게 DM이나 E-mail을 발송하여 적은 비용으로 높은 효과를 낼 수 있다.

5.2 웹 사용자 패턴 분석기

웹 사용자 패턴 분석기는 서버에 접속하는 클라이언트의 다양한 페이지 사이에 상관 관계를 찾는 것이다. [그림 4]는 쇼핑몰 서버의 실제 웹 로그 파일에 대해 전처리 과정을 거쳤으며 결과 값을 얻기 위해 실험 값에 크게 영향을 주지 않는 범위에서 수정된 데이터를 입력으로 패턴 분석기를 구현했다. 결과 데이터를 통해 다음과 같은 유용성을 유추해 낼 수 있었다

a) "특정 기간에 스포츠 쇼핑몰에 방문한 사용자의 60%는 테니스 용품 페이지를 또한 방문한다."
b) "VTR과 관련된 페이지를 접속한 40%의 사람은 비디오 테입과 관련된 페이지를 방문한다."
c) "특별한 페이지를 방문한 사용자의 30%는 상품 A에 대한 실질적인 구매가 이루어진다."

위와 같이 웹 로그 파일의 분석을 통해 높은 신뢰도를 갖는 페이지들은 서로간에 밀접한 관련이 있음을 알게 되었고, 각 페이지간의 특정 시간이나 계절 별로 연관성이 높은 페이지에 대해, 웹 디자이너는 홈페이지 재구성에 대한 정보를 얻을 수 있다. 또한 실제 구매한 데이터베이스와 연계하여 어떤 웹 페이지 통해 쇼핑이 이루어지며, 어떤 시점에서 실제 구매가 일어나는지 파악하여 개인화 된 동적인 웹 서비스나 광고 배치에 적용하는 등의 효용성을 얻을 수 있다.

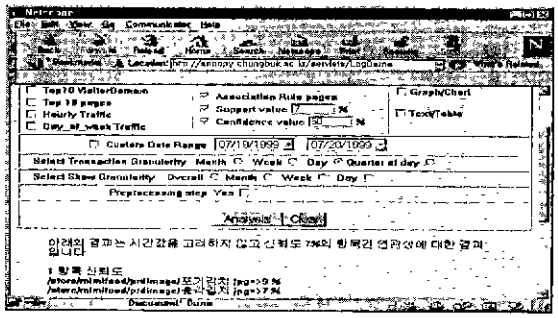

그림 4 웹 사용자 패턴 분석 결과

6. 결론 및 향후 연구

이 논문에서는 인터넷 DB 마케팅 전략에 데이터 마이닝의 기법을 응용하여 부가적인 서비스를 지원 할 수 있는 기법으로서 시간 연관 규칙 탐사 기법을 제안했고 그 세부 내용으로 상품 연관 규칙 탐사와 연관 규칙 탐사를 위한 웹 사용자 패턴 분석 기법에 대해 기술하였다.

상품 연관 규칙 탐사에서는 상품의 과거 매출 실적에 관한 기록 데이터로부터 상품간의 연관성 정도를 측정하여 이를 기반으로 마케팅 전략을 수립할 수 있었으며, 웹 사용자 패턴 분석에서는 웹 로그 파일의 분석을 통해 높은 신뢰도를 갖는 페이지들은 서로간에 밀접한 관련이 있음을 알게 되었고 각 페이지간의 특정 시간이나 계절 별로 연관성이 높은 페이지에 대해 웹 디자이너는 홈페이지 재구성에 대한 정보와 마케터는 효율적인 마케팅 전략을 얻을 수 있다.

향후 과제로는 특정 시간에서의 상품간의 연관성뿐만 아니라 시간의 흐름에 따른 순회 패턴과 경향분석 등의 연구가 진행되어질 필요가 있고, 웹 로그 파일의 사용자 트랜잭션 처리와 관련해 단순한 최대 시간 간격의 속성이나 사용자 정의 방법이외에 이를 효과적으로 처리하기 위한 연구가 진행 되어져야한다.

7. 참고 문헌

[Agra94] R Agrawal and R Srikant. "Fast Algorithms for Mining Association Rules," In Proc. of the 20th International Conference on Very Large Data Bases 1994.

[Chen98] X.Chen and Petrounias. "Discovering Temporal Association Rules in Temporal Databases," In Proc. of the International Workshop on Issues and Applications of Database Technology, 1998.

[Osma98] O.R.Zaane and J.Han "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on the web Logs," In Proc. of the ADL, 1998.

[Cool99] R.Cooley and J Srivastava. "Data Preparation for Mining World Wide Web Browsing Patterns," In Journal of Knowledge and Information Systems, Vol. 1, No. 1, 1999.

[이강태99] 이강태. "시간 연관규칙 탐사 기법," 충북대학교 석사학위논문, 1999.