전자상거래 머천트 시스템에서의 원투원 마케팅을 위한 데이터마이닝 시스템의 설계 및 구현

김종달*, 홍정희*, 김성민*, 남도원*, 이동하*, 김성훈**, 이전영*
*(lark, manifold, smkim, irene, dongha, jeon)@postech.ac.kr, **saint@etri.re.kr
*포항공과대학교 컴퓨터공학과
**한국전자통신연구원

Design and Implementation of A Data Mining System for One-to-One Marketing in EC Merchant Systems

*Jong-Dal Kim, *Jeong-Hee Hong, *Sung-Min Kim, *Do-Won Nam,
*Dong-Ha Lee, **Sung-Hoon Kim, *Jeon-Young Lee
*Dept. of Computer Science and Engineering, POSTECH
**Electronics and Telecommunications Research Institute

요 약

전자상거래에서 판매 실적을 높이기 위한 효과적인 방법의 하나는 사용자에 따라 개별화된 정보의 제공, 즉 원투원 마케팅의 개념을 도입하는 것이다. 이를 위해서는 사용자의 구매 성향이나 사용자의 특성에 대한 지식베이스가 있어야 한다. 이러한 지식베이스로 데이터마이닝 기법 중의 하나인 연관규칙을 도입하였다. 본 논문에서는 연관규칙을 기본 연산으로 하는 데이터마이닝 시스템의 설계와 구현을 기술하였다. 사용자와 제품간의 연관규칙을 추출하여 동적으로 제공되는 웹 문서를 생성하는데 필요한 지식베이스를 구축하였다. 또한 구축된 데이터마이닝 시스템은 연관규칙 탐사 엔진과 개념 계층 관리기로 구성되어 있으며, 대용량의 데이터를 다루기 위해 기존의 방법과는 다른 파일을 기반으로 한 빈번항목집합 인덱싱 기법을 제시하였다.

1. 서론

전자 상거래 머천트 시스템에서 판매 실적을 높이기 위한 효과적인 방법은 사용자에 따라 개별화된 서비스를 제공하는 원투원마케팅의 개념을 도입하는 것이다. Gartner Group의 보고에 따르면, 마케팅 전략의 추세는 필연적으로 시장중심(market-centric)에서 고객중심(customer-centric) 환경으로 변화할 것이라고 한다[1]. 즉 이러한 고객중심 마케팅의 기본이 되는 전략들이 바로 고객-관계 마케팅(customer-relation marketing), 원투원 마케팅(one-to-one marketing)이고 이를 뒷받침 할 수 있는 기술 중 하나가 바로 데이터마이닝을 사용하는 방법이다. 데이터마이닝이란 대용량의 데이터로부터 기존에 알려지지 않은 유용한 지식을 효과적으로 찾아내는 과정이다[2]. 데이터마이닝 기술을 이용한 원투원마케팅은 데이터베이스 마케팅 분야에서도 신기술로 인식되고 있으며, 초기단계의 몇몇 제품들만 선보이고 있다. 이 중 대표적인 제품인 Thinking Machine Inc. 의 DARWIN은 SOM(Self Organiaing Maps)과 같은 전통적인 인공지능 기술에 기반한 클러스터링(Clustering) 연산을 데이터마이닝 알고리즘으로 사용하고 있고, 그 밖의 제품들은 분류(Classification), 회귀분석 트리(Regression Tree), 신경망(Neural Networks) 등을 사용하여 모델링하고 있다.

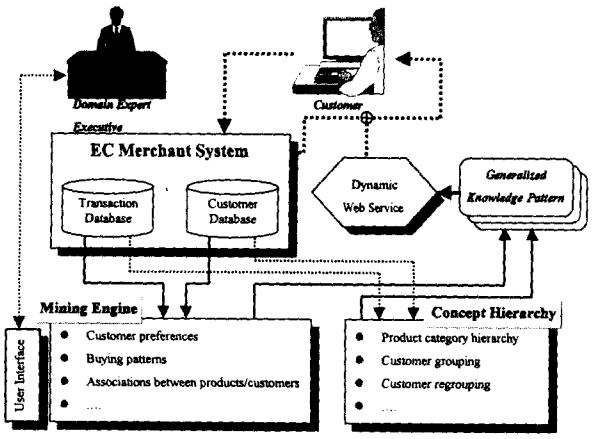
기존의 이러한 제품들과는 달리, 본 연구에서는 원투원마케팅을 지원하기 위해 데이터마이닝의 기본연산 중 하나인 연관규칙을 도입하였다. 연관 규칙은 항목의 집합으로 이루어지는 트랜잭션으로 구성된 데이타베이스에서 X → Y 의 형태로 표현되는 패턴이다. 여기서 X, Y 는 항목들의 집합이다. 이 규칙의 의미는 주어진 데이타베이스에서 X를 포함하는 트랜잭션들이 항목 집합 Y를 포함하는 경향이 있다는 것이다[3].

본 논문의 구성은 다음과 같다. 먼저, 2장에서는 설계된 전체 시스템의 구조를 설명하고, 3장에서는 각 모듈의 기능 및 요소 기술을 설명한다. 4장에서는 시스템 구현에 관하여 설명하고, 5장에서 결론을 제시한다.

2. 데이터마이닝 시스템의 설계

이 장에서는 전자상거래 머천트 시스템에서 원투원마케팅을 지원하기 위한 데이터마이닝 시스템의 구조를 살펴보고자 한다. 먼저, 머천트 시스템과 데이터마이닝 시스템과의 연계를 고려한 전체 시스템의 구조를 살펴보면 <그림1>과 같다.

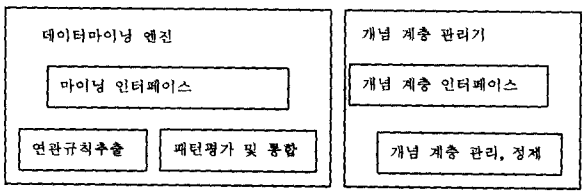
즉, 머천트 시스템의 고객 데이터베이스와 트랜잭션 데이터베이스를 대상 데이터베이스(Target Database)로 하여 데이터



<그림1> 전체시스템의 구조

마이닝 엔진으로부터 고객의 성향, 구매패턴, 상품과 고객간의 상관성 등을 연관규칙으로 추출한다. 이때, 상품의 카테고리, 고객의 그룹화 등을 개념 계층(Concept Hierarchy)으로 표시하고, 이를 이용하여 보다 일반화된 지식 패턴을 얻는다. 이렇게 추출된 지식을 사용하여 동적인 웹서비스를 통해 사용자에게 원투원마케팅을 하는 것이 전체 시나리오이다. 즉, 추출된 연관규칙을 사용하여 머천트 시스템에서는 동적인 사용자 인터페이스의 구성(예를 들면, 구매자가 다음으로 관심 있어할 쪽의 링크를 페이지 상단에 동적으로 구성), 개별 마케팅(구매자의 패턴에 따른 할인율 변경 및 이벤트), 이메일 발송(발송회수는 줄이고, hit ratio는 높게), 사용자 그룹화(구매 패턴에 따른 사용자의 재그룹화 및 세분화)등의 응용이 가능하다. 이러한 응용은 머천트 시스템을 운영하는 곳의 마케팅 전략에 따르게 된다. 따라서, 머천트 시스템에 데이터마이닝 시스템을 연계함으로써 원투원 마케팅을 위한 지식베이스를 구축할 수 있다.

이를 위한 데이터마이닝 시스템의 구조는 <그림2>와 같다.



<그림 2>. 데이터마이닝 시스템 구성도

데이터마이닝 엔진과 개념 계층 관리기의 두 부분으로 되어있으며, 마이닝 엔진에는 연관규칙 추출의 기본 모듈과 추출된 패턴의 통합기와 인터페이스가 있다. 개념 계층 관리기는 고객이나 상품 세부 정보의 값들간의 계층적 정보를 저장하고 마이닝 엔진과의 인터페이스를 담당하는 부분이다. 개념 계층을 사용하여 보다 일반화된 패턴을 추출할 수 있게 된다.

3. 각 모듈별 기능 및 요소기술

이 장에서는 2장에서 설명한 데이터마이닝 시스템을 데이터마이닝 엔진 부분과 개념 계층 관리기 부분으로 나누어서 살펴보도록 하겠다. 이 장에서 언급하는 연관규칙 추출과정은 [3][4]에서 찾아 볼 수 있다.

3.1 데이터마이닝 엔진

데이터마이닝 엔진을 구성하는 프로세스는 데이터베이스 인터페이스, 빈번항목추출, 패턴평가 및 통합, 사용자인터페이스의 네 가지 부분이다.

데이터베이스 인터페이스는 전자상거래 데이터를 가지고 있는 로-데이터베이스(raw-database)와의 연결을 담당한다. 로-데이터베이스는 한번 완전히 읽어들여지며, 세 가지 데이터를 준비하게 된다. 데이터베이스의 항목을 코드화하고, 1-빈번항목집합을 결정하며, 이후의 데이터 마이닝 엔진이 참조할 대상 데이터베이스(target database)를 구축하게 된다.

빈번항목추출 프로세스는 매번 k-빈번항목집합과 대상 데이터베이스를 한번 완전히 읽어들여, k+1-빈번항목집합을 생성해내는 작업을 수행한다. 또한 이 과정에서 생성된 후보빈번항목집합의 개수를 줄이기 위해, 해쉬테이블을 이용한다[7]. 또한 다음 단계를 위한 대상 데이터베이스도 재구성하여 크기를 줄여나간다. 빈번항목추출 프로세스는 데이터마이닝 엔진의 가장 중요한 부분이다.

패턴추출 프로세스는 빈번항목추출이후에 연관규칙이나 순차패턴을 추출하는 프로세스이고, 사용자 인터페이스는 결과로 얻어진 연관규칙이나 순차패턴을 정렬하거나 선택하는 등 후처리를 담당한다. 후처리를 위한 단순한 프로그램으로 볼 수도 있다.

3.2 개념 계층 관리기

개념 계층 관리기는 개념 계층의 생성과 이용을 담당한다. 개념 계층의 생성은 파일 형태로 주어지는 개념 계층을 코드화해서 데이터베이스에 저장한다. 이러한 작업은 데이터 마이닝 엔진의 코드화 사전과는 다른 작업이며, 개념 계층에서 코드화하는 것은 후에 검색을 빠르게 하려는데 목적이 있다. 데이터 마이닝 엔진에서 이용할 때는 항목을 하나 주고, 이의 상위개념들을 순서화해서 추출해준다. 개념 계층의 관리를 위한 인터페이스에는, 현재 쓰고 있는 개념 계층을 평가하거나, 여러 가지의 개념 계층이 존재할 경우, 이용할 개념 계층을 지정하거나, 관리자가 개념 계층의 일부를 직접 수정하는 등의 처리를 담당한다.

4. 시스템 구현

4.1 파일을 기반으로 한 빈번항목집합 인덱싱

먼저 각 i-빈번항목집합에 대해 각각 다른 파일에 저장하며, level 당 하나의 정렬된 리스트에 저장한다. 후보빈번집합도 같은 구조를 가진다. 대상 데이터베이스를 스캔한 뒤에 지지도를 판정하여 빈번항목집합이 결정되면, 파일을 복사하여 빈번항목집합으로 만들고, 후보빈번항목집합의 파일을 삭제한다.

빈번항목집합은 세 가지로 접근된다. 먼저, k+1-후보빈번집합을 생성할 때의 죠인 연산은, k-빈번항목집합은 두개의 포인터를 이용하여 순차적으로 접근하야만 한다.

다음으로는 대상 데이터베이스를 스캔하며, 후보빈번집합의 발생 순서를 세는 작업이다. 이때, 대상 데이터베이스의 하나의 트랜잭션에 대해, 이 트랜잭션에 포함된 모든 후보빈번집합을 찾는 과정이 필요하다. 이를 위해 binary트리의 검색시, 동시에 검색하는 알고리즘으로 수정한다. 수정된 알고리즘은 트랜잭션

의 최대 크기가 정해진 경우(보통 20 이하) 최대 크기만큼의 파일 포인터를 유지할 필요가 있다.

그 다음으로는 빈번항목집합이 다 찾아진 뒤에 연관규칙 추출 알고리즘에서 참조하는 경우이다. 이때는 임의 접근 방법(random access)로 접근된다.

세가지 접근 방법을 고려할 때, 동시 검색을 지원하는 파일 기반 Trie로 빈번항목집합과 후보빈번항목을 구성하는 것을 선택했다.

크기가 1인 후보항목 집합의 경우, 파일 포인터를 이용한 binary 트리로 생성한 다음, 빈번항목을 결정한 다음에는 위와 같은 파일 기반 Trie 에 항목의 동시 검색을 허용하는 알고리즘을 이용하여 활용한다.

4.2 사전(Dictionary) 구축

로-데이터베이스의 항목들을 실제 프로세싱 할 때는 각 항목을 4비트의 코드로 인코딩하여 사용한다. 이는 빠른 연산과 효율적인 프로세싱을 위한 작업이다. 이를 위해 로-데이터베이스로부터 항목을 읽어들이면서 (코드, 항목)의 형태로 이루어진 사전을 구축하고 이와 동시에 1-후보빈번항목집합에서 항목의 개수를 센다. 이 작업을 마친 후에 각 항목의 빈번한 정도(frequency)를 참조하여 적게 나타나는 항목이 사전에서 앞에 높이도록 재배열한다. 이때 사전은 코드와 항목 두 가지 모두를 키로 가지도록 하여야 한다.

4.3 대상 데이터베이스(Target Database)

대상 데이터베이스는 처음에는 로-데이터베이스로부터 생성된다. 데이터베이스를 마이닝 엔진에서 지속적으로 접근하는 것은 성능이 떨어지므로, 대상 데이터베이스는 순차접근만 허용하면 된다. DHP[4] 기법에서 다음 번의 스캔을 위해 대상 데이터베이스를 축소시키는 기법이 있으므로, 매번 대상 데이터베이스는 다른 파일에 생성되고, 이전 단계의 대상 데이터베이스는 지워진다.

대상 데이터베이스의 포맷은 한 트랜잭션의 개수를 지정하는 부분과 고정 길이의 항목 코드의 나열로 구성된 트랜잭션의 배열로 표현된다. 항목 코드의 길이가 4바이트이고, 항목의 개수를 2바이트로 지정하는 경우, 항목 네 개로 이루어지는 트랜잭션은 20바이트로 구성되는 것이다.

4.4 파일처리

데이터마이닝 시스템의 특성으로 대용량의 데이터를 처리하기 위해, 모든 데이터는 파일에 저장될 수 있는 것으로 가정을 한다. 대상데이터베이스, 빈번항목집합, 추출된 연관 규칙 등이 파일로 저장된다. 그러나 메모리안에서 처리가 가능한 부분을 파일에서 수행할 경우, 성능저하가 우려된다. 따라서 각 프로세스는 적당한 단계에서 파일에서 수행하던 작업을 메모리에서 수행하는 것을 결정해야 한다.

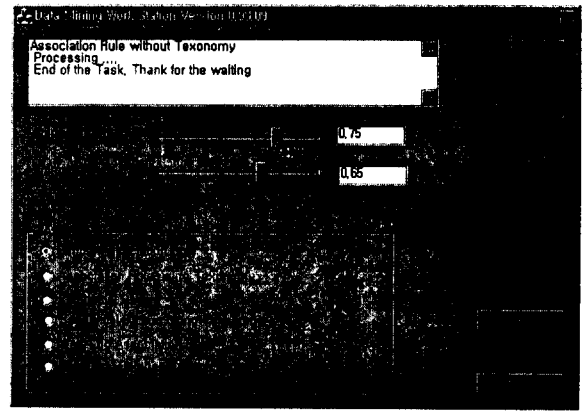
또, 빈번항목집합 탐사의 과정에서 하나의 파일을 두개의 파일포인터를 이용해 스캔해야하는 경우가 있다. 이를 위하여 파일의 동시접근을 처리해야한다. 이 과정이 메모리상에서 수행될 경우는 별도의 고려가 필요 없다.

4.5 개념 계층의 구축

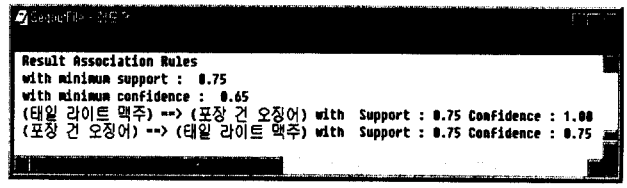
일반화된 규칙을 얻기 위하여 개념 계층을 사용하는데, 머천트 시스템의 상품 카테고리 정보를 개념 계층으로 맵핑하여 각 상품 항목에 대해 상위 카테고리의 개념을 리스트로 유지하고, 사용자 그룹정보 역시 개념 계층으로 맵핑하여 사용자 그룹이 각 사용자의 상위 개념에 해당하는 관계의 리스트로 표현한다.

5. 결론

실제 구현된 데이터마이닝 엔진과 샘플데이터로 수행된 결과를 <그림3>과 <그림4>에서 보여주고 있다.



<그림 3> 마이닝 엔진 실행 화면



<그림 4> 연관규칙을 결과를 파일로 출력한 화면

본 연구에서는 전자상거래 머천트 시스템에서 원투원마케팅을 지원하기 위해 연관규칙을 연산으로 하는 데이터마이닝 시스템을 설계, 구현하였다. 특히 머천트 시스템의 상품 카테고리 정보와 사용자 그룹 정보를 이용하여 개념 계층 트리를 전자상거래 머천트 시스템과 결합시켰고, 파일을 기반으로 한 빈번항목집합 인덱싱 기법은 메모리에 제한 없이, 대용량의 데이터베이스에서 연관규칙 추출을 가능하게 하였다.

참고문헌

[1] Margo Komenar, Electronic Marketing, 83-84pages, Wiley Computer Publishing

[2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. Of the 20th VLDB Conference, pages 487-499, Santiago, Chile, 1994. Expanded version available as IBM Research Report RJ9839, June 1994.

[4] Jong Soo Park, Ming Syan Chen, and Philip S. Yu, An effective hash based algorithm for mining association rules, the ACM-SIGMOD Conference on Management of Data, San Jose, California, May 1995