論文
2-7-8

전자상거래에서 연관규칙과 순차패턴을 이용한 온라인 마이닝 (On-Line Mining using Association Rules and Sequential Patterns in Electronic Commerce)

김 성 학*
(Sung-Hark Kim)

요 약

인터넷 사용 인구의 증가로 전자상거래는 새로운 상거래 형태로 빠르게 발전하고 있으며, 대다수 인터넷 쇼핑몰들은 사용자에게 더 많은 정보와 편리한 사용자 인터페이스를 제공함으로써 보다 많은 고객을 확보하려고 노력하고 있다. 편리한 인터페이스 중의 하나는 상품을 추천해주는 서비스이며, 이를 위해서는 쇼핑몰에서의 구매정보, 행동 그리고 장바구니 등 사용자로부터 특정 행동패턴을 추출하고 분석하는 방법이 필요하다. 이러한 방법 중에서 상품간의 연관성 추출을 위하여 주로 연관규칙과 순차패턴이 이용되고 있는데, 대부분의 온라인 전자상거래에서는 사용자의 정보 또는 구매이력을 가지고 카테고리를 중심으로 수행하고 있다. 그러나 이는 단일한 구매패턴에 의한 연관성만을 나타낼 뿐이며, 상품 각각에 대한 연관성을 찾아보기 힘들다. 또한 단일 구매패턴은 계산 비용이 작기는 하지만 사용자의 구매패턴을 정확하게 반영하기 어렵다. 따라서 본 논문에서는 이러한 문제를 해결하기 위하여 카테고리 독립적이고 단일 항목간의 구조화를 통하여 항목간의 연계성을 갖는, 다중 구매패턴을 고려하는 마이닝 방법을 제안한다.

ABSTRACT

In consequence of expansion of internet users, electronic commerce is becoming a new prototype for marketing and sales, and most of electronic commerce sites or internet shopping malls provide a rich source of information and convenient user interfaces about the organizations customers to maintain their patrons. One of the convenient interfaces for users is service to recommend products. To do this, they must exploit methods to extract and analysis specific patterns from purchasing information, behavior and market basket about customers. The methods are association rules and sequential patterns, which are widely used to extract correlation among products, and in most of on-line electronic commerce sites are executed with users information and purchased history by category-oriented. But these can't represent the diverse correlation among products and also hardly reflect users' buying patterns precisely, since the results are simple set of relations for single purchased pattern. In this paper, we propose an efficient mining technique, which allows for multiple purchased patterns that are category-independent and have relationship among items in the linked structure of single pattern items.

* 정회원 : 유한대학 전자계산과 부교수

논문접수 : 2001. 7. 13.
심사완료 : 2001. 7. 21.

※ 본 연구는 2000년도 유한대학 연구지원센타 연구비 지원에 의한 것임.

1. 서론

인터넷을 통한 정보교류와 정보검색이 보편화되면서 정보의 질적양적인 면에서 급격한 증가와 사용자의 폭발적인 증가를 가져오게 되었다. 이와 같은 현실에서는 원하는 정보를 갖고 있는 웹 사이트를 찾아내기가 점차 더 어려워지며, 적절한 시간 안에 원하는 정보의 획득이 힘들고 또한, 제공되는 많은 정보는 그 가치가 저하되는 등의 문제들이 발생하게 된다[4,5].

이러한 문제점들은 인터넷 사용의 증가로 인해 발전하고 있는 전자상거래 분야에 대해서도 동일하게 나타난다고 볼 수 있다. 전자상거래 분야에서는, 기존의 데이터마이닝(data mining) 방법을 웹 환경에 응용하여 이러한 문제를 해결하기 위한 많은 연구가 진행되고 있으며, 이를 응용하여 e-business에서의 다양한 마케팅 전략을 수립할 수 있게 되었다. 전자상거래는 새로운 상거래의 형태로 빠르게 발전하고 있으며, 빠른 변화에 맞추어 각 인터넷 쇼핑몰들은 사용자에게 더 많은 정보를 제공하고 편리한 사용자 인터페이스를 제공함으로써 보다 많은 수의 회원을 확보하려고 노력하고 있다. 편리한 인터페이스 중의 하나는 상품을 추천해주는 서비스이며, 이는 사용자가 쇼핑몰에서 구매한 정보, 행동 그리고 장바구니(market basket) 등 사용자로부터 특정 행동패턴을 추출하여 분석하고 이를 바탕으로 사용자에게 적합한 특정 상품을 구매하도록 유도하는 서비스이다[1,7].

현재의 인터넷 쇼핑몰에서 상품을 추천하는 방식은, 각 카테고리별로 사용자의 관심도(profile)를 저장하게 되고, 후에 사용자가 특정 카테고리에 접속하게 되면 관심을 보인 카테고리의 상품이 디스플레이되도록 하고 있다[8]. 이러한 방법은 사용자가 처음에 어떤 항목에 관심이 있는 지를 일일이 알려주어야 하며, 특정 카테고리의 상품에 대해서만 추천을 함으로써 다른 카테고리에 있는 상품에 관한 연관성을 찾아 볼 수 없다는 단점이 있다.

이에 본 논문에서는 상품들 간의 연계성 추출을 위해서 연관규칙(association rules)과 순차패턴(sequential patterns) 방법을 이용하여 데이터베이스에 저장되어 있는 트랜잭션(transaction)에 대해서 마이닝(mining)하고, 관련되는 패턴을 추출하여 다양한 패턴들간의 연관도를 보다 구체적으로 나타낼 수 있도록 하였다. 보통 마이닝을 통해서 추출해 낸 정보는 A→B , 즉 "구매자가 A라는 물건을 사면 B라는 물건을 살 가능성이 높다."라는 형태의 패턴으로 추출된다. 이는 단일 품목간의 구매패턴으로써 여러 품목에 관한 연관성을 찾기가 불가능하며, 또한 사용자의 구매패턴을 정확하게 반영하기 어렵다. 따라서 본 논문에서는 다중 패턴 즉, A → B → C와 같은 패턴을 이용하여 카테고리 독립적이고 단일 항목간의 구조화를 통해서 항목들을 연계하는, 다중 구매패턴을 고려하는 마이닝 방법을 제안하고, 일반적인 테이블 구조에서 나타나는 상품들간의 연관성 표현보다 더욱 효율적으로 다양한 상품들간의 연관도를 추출할 수 있음을 보인다.

2. 관련 연구

이 장에서는 데이터베이스의 트랜잭션으로부터 상품 구매패턴을 추출하기 위하여 필요한 데이터마이닝 기법 중에서 본 논문과 관련이 깊은 연관규칙과 순차패턴에 대해 알아본다[2,10].

2.1 연관규칙

데이터베이스에서 잘 알려져 있지 않은 숨겨진 패턴을 탐사하는 연구 중에서 연관규칙에 대해 가장 많은 연구가 이루어 졌다. 연관규칙은 문자 그대로 한 항목 그룹과 다른 항목 그룹 사이에 존재하는 강한 연관성을 찾아내어 그룹화 하는 클러스터링(clustering)의 일종이다. 또한, 동시에 구매될 가능성이 큰 상품들을 찾아냄으로써 장바구니 분석(market basket analysis)에서 다루는 문제들에 적용할 수 있다. 연관규칙 기법에 적용되는 데이터는 판매 시점에서 기록된 거래와 품목에 관한 정보를 담고 있고, 연관규칙 탐사과정은 크게 두 단계로 진행이 된다. 첫번째는 높은 지지도(support)를 갖는 즉, 항목간의 연관성이 높다고 가정되는 항목의 집합(itemset)인 빈발 항목집합(frequent or large itemsets)을 식별하는 작업이고, 두 번째 단계는 이러한 빈발 항목집합을 이용하여 높은 신뢰도(confidence)를 갖는 연관규칙을 도출하는 작업이다. 여기서 지지도와 신뢰도는 매우 중요한 개념으로써 빈발 항목집합과 연관규칙을 찾아내

는데 있어서 논리적 타당성을 제공하는 큰 역할을 한다. 연관규칙발견 알고리즘으로는 AIS, SETM, Apriori, DHP 알고리즘 등이 연구되었는데, 이들 대다수는 알고리즘에 의해 생성되는 빈발 항목집합 후보들의 수를 감소시키는 것에 초점을 두고 있다. 이는 후보 빈발 항목집합의 발생 빈도를 계산하는 것은 상당량의 프로세싱 시간과 메모리를 요구하기 때문에 알고리즘 성능을 평가하는 중요한 요인이기 때문이다[1,6].

[연관규칙의 정의]

$I = \{i_1, i_2, \cdots, i_m\}$을 항목들의 집합이라 하자. D를 트랜잭션들의 집합이라 부르고, 각 트랜잭션 T는 $T \subseteq I$인 항목들의 집합이다. 각 트랜잭션들은 TID(Transaction IDentifier)를 갖고 있다. X를 항목들의 집합이라고 하면, $X \subseteq T$ 이고 이 때, 트랜잭션 T가 X를 포함한다고 한다. 연관규칙은 $R : X \rightarrow Y$ 형식이고, 이 때 X와 Y는 서로 같은 원소를 갖지 않는 항목집합이다. 즉, $X, Y \subseteq I$ 이고 $X \cap Y = \phi$ 이다. 단, $Y \neq \phi$ 이어야 한다.

만일 한 트랜잭션이 X를 지지한다면, 또한 어떤 확률에 의해 Y도 지지할 것이라는 예측이 연관규칙이다. 이런 확률을 이 규칙의 신뢰도(conf(R))라 한다.

$$\begin{aligned} \text{conf(R)} &= p(Y \subseteq T \mid X \subseteq T) \\ &= \frac{p(Y \subseteq T \wedge X \subseteq T)}{p(X \subseteq T)} \\ &= \frac{sp(X \cup Y)}{sp(X)} \end{aligned}$$

또한 T가 X의 모든 항목들을 포함한다면($X \subseteq T$) T가 집합 X를 지지(support)한다고 한다. X의 지지도를 sp(X)로 정의하며, 이는 X를 지지하는 D에 있는 든 트랜잭션의 개수를 의미한다. 따라서, D에 있는 규칙 R에 대한 지지도는 sp($X \cup Y$)가 된다. 규칙의 신뢰도는 얼마나 자주 적용할 수 있는 지를 나타내는 반면 지지도는 그 규칙 전부가 얼마나 믿을 만한 지를 보여준다. 규칙이 데이터베이스에서 적절해지려면 충분한 지지도와 신뢰도를 가져야 한다[11].

[연관규칙 탐사단계]

(1) 빈발항목집합을 찾는다.
미리 결정된 최소지지도 s_{min} 이상의 트랜잭션 지지도를 갖는 모든 빈발 항목집합들을 찾는다.

(2) 데이터베이스로부터 연관규칙 생성을 위하여 빈발 항목집합을 사용한다.
모든 빈발 항목집합 l에 대하여 l의 모든 공집합이 아닌 부분집합들을 찾는다. 각각의 부분집합 a에 대하여, $sp(a)$에 대한 $sp(l)$의 비율이 적어도 최소신뢰도 c_{min}이상이면, 즉

$$\frac{sp(l)}{sp(a)} \geq c_{min}, \quad a \Rightarrow (l - a)$$

형태의 규칙을 생성한다.
그러므로 어떤 주어진 최소신뢰도 c_{min}와 최소지지도 s_{min}에 대하여 만일 conf(R) $\geq c_{min}$이고 sp(R) $\geq s_{min}$이면 규칙 R은 D에 대하여 성립한다. 규칙이 성립되기 위하여 필요한 조건으로서 규칙의 조건부(antecedent)와 결과부(decendent)가 모두 빈발해야 한다.

2.2 순차패턴

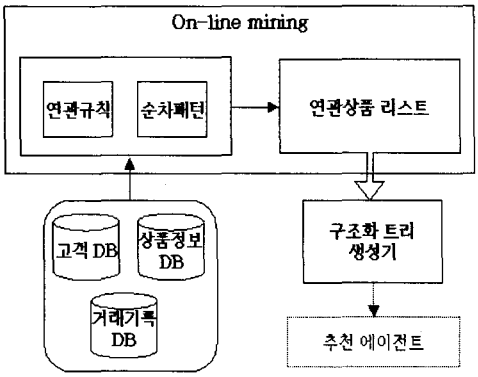
순차패턴탐사는 한 트랜잭션 안에서 발생하는 항목들간의 연관규칙에 시간의 변이를 추가한 것이다. 즉, 연관규칙은 트랜잭션 안에서 어떤 항목을 함께 구입하는 가에 관한 문제로 트랜잭션 내의 문제인 반면, 순차패턴을 발견하는 것은 트랜잭션 상호간의 문제인 것이다[3]. 각 트랜잭션은 고객 ID와 트랜잭션 시간과 그 시간에 구매된 항목들로 구성되고, 같은 고객들에 대해서는 같은 시간에 두개 이상의 트랜잭션은 존재하지 않으며 또한 항목의 수량을 고려하지 않는다고 가정한다. 이렇게 구성된 각 고객에 대한 각각의 시퀀스 집합(트랜잭션 데이터베이스)에서 다른 시퀀스에 포함되지 않는 시퀀스 -최대 시퀀스(maximal sequences)를 순차패턴이라 부르며 최소지지도를 만족하는 시퀀스를 빈발 시퀀스(large sequences)라 한다. 시퀀스에 대한 지지도의 정의는 시퀀스를 지지하는 전체 고객들의 수이다. 빈발 시퀀스는 항목집합 목록의 형태로 나타나며, 그 항목집합들은 반드시 최소지지도를 만족해

야 한다. 주어진 고객에 대한 트랜잭션 데이터베이스에서 순차패턴 탐사는 사용자가 정의한 최소지지도를 만족하는 모든 빈발 시퀀스들 사이에서의 최대시퀀스를 찾는 것이며, 이것이 연관규칙이 된다.

즉, 순차패턴은 동시에 구매될 가능성이 큰 상품군을 찾아내는 연관규칙에, 시간의 개념이 포함되어 순차적인 구매 가능성이 큰 상품 군을 찾아내는 방법이다. 순차패턴에서는 연관규칙 A→B는 "상품 A가 구매되면 일정 시간이 경과한 다음 상품 B가 구매된다."라고 해석된다. 즉, 순차패턴은 구매 순서가 고려되어 상품간의 연관성이 측정되고, 이에 따라 유용한 연관규칙을 찾는 기법이다.

3. 시스템 구조

본 논문에서 제안하고 있는 온라인 마이닝의 개략도는 [그림 1]과 같다. 사용자가 인터넷 쇼핑몰에서 구매를 하게 되면, 구매 내역이 '거래기록 DB'에 남게 된다. 이 정보를 연관규칙과 순차패턴 방법을 이용하여 연관상품 리스트(Rule)를 추출하고, 상품들간의 구조화된 트리를 생성한다. 추천 에이전트는 생성된 구조화 트리를 이용하여 추천할 상품을 결정해서, 이를 사용자에게 추천하게 된다.



[그림 1] 온라인 마이닝의 개략도
[Fig. 1] The schematic diagram of on-line mining

3.1 구매 패턴의 추출 및 구조화

연관규칙과 순차패턴을 발견하기 위한 방법은, Apriori를 개선한 알고리즘으로써 후보 항목집합들을 효율적으로 생성하고, 트랜잭션 데이터베이스의 크기를 효과적으로 줄여 탐색시간이 빠른 것으로 알려져 있는 DHP(standing for Direct Hashing and Pruning) 알고리즘[9]을 사용하였고, 추출된 패턴의 구조화는 다음의 4단계를 통해 이루어진다.

1) 사용자 DB에서 트랜잭션을 추출
2) 사용자 트랜잭션을 모든 부분집합으로 분리
3) 생성된 단일 패턴의 분포도 계산
4) 분포도에 의거하여 구조화된 트리 생성

<표 1> 구매자의 원시패턴
<Table 1> Source patterns of purchaser

Rule 1:	House → Car → TV → DVD
Rule 2:	TV → DVD → Audio
Rule 3:	House → Car → Audio → DVD
Rule 4:	House → TV → DVD
Rule 5:	TV → Audio

구매자의 원시패턴을 연관규칙을 이용하여 단일 패턴으로 나눈다. 단일 패턴으로 나누기 위하여 고려해야 할 조합은, 예로 든 <표 1>의 Rule 1과 같은 경우는 {House, Car}, {House, TV}, {House, DVD}, {Car, TV}, {Car, DVD}, {TV, DVD} 등의 6개($_4C_2$)이다. 이와 같은 방법으로 구매자의 원시 패턴에 대한 모든 조합을 구한다. 이때, {House, Car}와 {Car, House}는 다르게 취급된다. 장바구니 분석에는 위와 같은 패턴이 같은 항목으로 취급 되지만, 제안된 패턴의 구조화에서는 패턴에 방향성이 고려되기 때문이다. 모든 패턴이 조합되었다면, 다음과 같이 <표 2>를 생성할 수 있다.

<표 2> 항목들간의 빈발도 테이블
<Table 2> Frequency table among items

House			
House → Car	House → TV	House → DVD	House → Audio
2 (10.5%)	2 (10.5%)	3 (15.7%)	1 (5.2%)

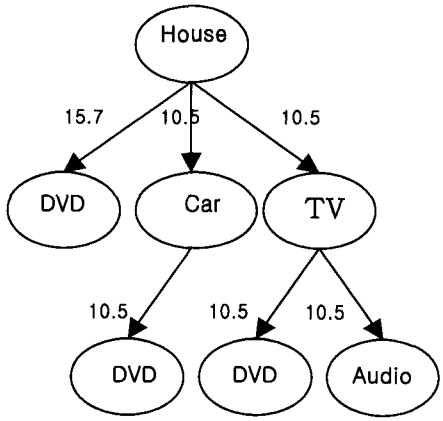
Car			
Car → House	Car → TV	Car → DVD	Car → Audio
0 (0%)	1 (5.2%)	2 (10.5%)	1 (5.2%)

TV			
TV → House	TV → Car	TV → DVD	TV → Audio
0 (0%)	0 (0%)	2 (10.5%)	2 (10.5%)

DVD			
DVD → House	DVD → Car	DVD → TV	DVD → Audio
0 (0%)	0 (0%)	0 (0%)	1 (5.2%)

Audio			
Audio → House	Audio → Car	Audio → TV	Audio → DVD
0 (0%)	0 (0%)	0 (0%)	1 (5.2%)

이렇게 단일 항목으로 추출된 패턴들은 각 항 목과 얼마만큼의 연관성을 가지고 있는 지를 나타내게 된다. 이 빈도수 테이블(지지도 6%이상)에 의거하여 트리를 구성한다. 트리를 구성함에 있어서 가장 빈도수가 높은 항목은 다른 상품과의 연관도가 가장 높다고 할 수 있다. 이 항목을 중심으로 트리를 구성해 나간다. 하위 노드의 확장은 빈도수가 가장 높은 항목을 첫 번째 하위 노드, 두 번째로 높은 항목을 그 다음 노드, 이러한 방식으로 확장 노드를 생성해 나간다. 그 수행결과를 [그림 2]에서 보이고 있다.



[그림 2] 구조화된 트리(지지도 6% 이상)
[Fig. 2] Structured tree (above 6% support)

<표 2>에 따라 빈도수가 높은 항목을 우선으로 하여 하위 노드의 왼쪽에 두면서 점진적으로 트리를 구성해간다. 이때, 지지도 6% 미만인 연 관 항목들은 빈도수가 저조하므로 자동적으로 트리의 구조화에 반영되지 않도록 한다. 이와 같이 일정한 임계치(threshold)를 두어서 많이 선택되지 않는 제품들을 걸러내는 작업이 필요하다. 이 값은 일정한 수치가 정해져 있는 것이 아니라, 통계치에 의해서 쇼핑몰의 운영자가 선택해야 할 사항이다.

3.2 마이닝 결과에 따른 상품 추천

위에서 상품 구매정보에 대해서 마이닝을 수행하여 얻어진 연관 항목들에 빈도수 측정을 하여 특정 항목과 관련이 깊은 항목을 찾아내었으며, 빈도수를 가중치로 하는 트리로 구조화 하여 연관 항목들에 대한 구조화를 수행 하였다.

이렇게 구조화된 트리를 사용하여, 추천 에이전트는 고객의 구매 정보를 감시하다가 특정 항목을 구매하게 되면, 그 즉시 구조화된 트리를 탐색하여 연관

되는 항목을 추천하게 된다. 기본적으로 탐색은 pre-order 방식으로 넓이우선탐색(breath-first search) 방법을 수행하게 되며, 탐색을 시작하게 될 루트 노드(root node)는 고객이 처음 구매한 항목이 된다. 관련상품 추천은 탐색된 하위 노드들 중에서 지지도 값이 가장 큰 노드를 추천하게 된다. 추천한 품목을 구매하게 되면 이는 다시 고객의 구매 정보 DB에 저장되며, 다시 트리의 탐색을 통해 다음 품목을 추천하게 된다. 만약 루트 노드를 정할 때, 트리에 여러 개의 동일 노드가 있다면 하위 노드의 지지도 값이 가장 큰 노드를 루트 노드로 결정한다.

3.3 구조화된 트리의 수정

고객의 구매 데이터가 점점 쌓여 갈수록 사용자들의 구매패턴을 트리에 반영해주어야 보다 정확하며 신뢰성있는 결과를 얻을 수 있게 된다. 따라서 본 논문에서는 항상 최신의 패턴을 반영하기 위하여 고객의 구매 데이터가 특정한 임계치를 초과할 때마다 트리가 수정되도록 하였다.

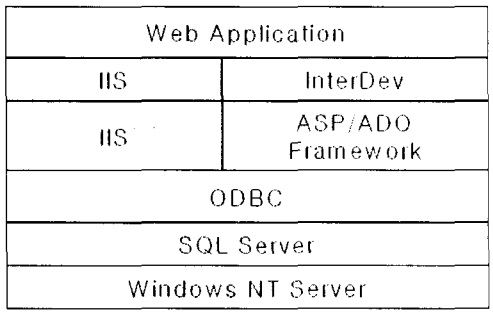
예로써, 새로운 패턴에서 house → car의 지지도가 16.4% 로 변경되었다면 house → car에는 $|sp_{old} - sp_{new}|$를 추가하고 같은 레벨의 나머지 노드에는 $|sp_{old} - sp_{new}| / (동일\ 레벨\ 노드의\ 개수-1)$ 만큼 각각 감소시켜 준다. 이와 같은 방식으로 사용자의 구매 패턴에 맞추어 트리를 수정해 나간다면 사용자의 구매패턴을 보다 정확하고 효과적으로 반영할 수 있을 것이다.

3.4 도태된 항목의 삭제

시간이 지날수록 점점 많이 팔리는 항목이 있는가 하면 점차 팔리지 않아서 구매 데이터베이스에서 찾아볼 수 없는 항목들이 생기게 된다. 이러한 항목들에 대해서는, 일정한 간격마다 한번씩 지지도가 특정한 임계치를 넘지 않는 항목들을 트리에서 삭제하면, 점차 사라져가는 항목에 대해서도 트리의 구조화에 반영할 수 있게 되어 보다 신뢰성있는 결과를 얻게 된다.

4. 실험 및 고찰

본 연구의 실험을 위하여 가상의 인터넷 쇼핑몰, CPS(Computer Part Store)를 구축하여 수행한다. CPS는 운영체제로서 마이크로소프트사의 윈도우 NT 서버 4.0(Windows NT Server)과 IIS 4.0 웹 서버(Internet Information Server), ASP (Active Server Page) Component, SQL Server 7.0, FrontPage98, Visual InterDev 6.0 등을 사용하여 구현하였다.



[그림 3] 실험 환경

[Fig. 3] Experimental environments

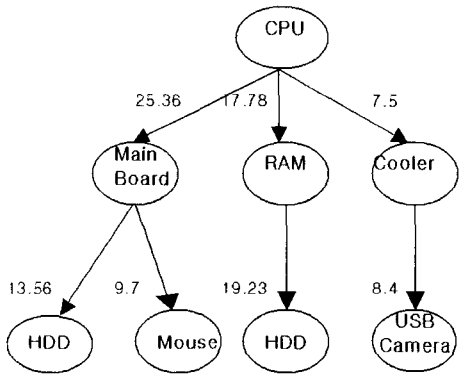
이러한 환경에서의 실험은, 구매자의 원시패턴에 대해 단일 항목으로 분해하고, 일반적인 테이블구조를 사용하는 방법과 본 논문에서 제안하는 구조화된 트리를 사용했을 때, 테이블 구조보다 구조화된 트리가 항목간에 연관성을 더욱 구체적으로 반영해줄 수 있음을 보인다. 1), 2)에서 실험에서의 주요 부분을 표현하고 있다.

1) 상품 구매 데이터베이스에서 트랜잭션을 추출하여 각 항목간의 연관성을 테스트한다. 연관 테이블을 구성할 항목은 지지도 6% 이상으로 작성하였다. 그 결과를 <표 3>에서 보이고 있다.
2) 데이터베이스에서 추출한 데이터에 근거하여 [그림 4]와 같은 트리를 작성할 수 있다. 이 트리는 CPU 항목과 나머지 Main Board, RAM, Cooler 항목들 간의 연관도를 나타내고 있다.

<표 3> 단일 항목집합의 빈발도

<Table 3> Frequency table of single itemsets

연관 항목	지지도
CPU, Main Board	25.36%
CPU, Cooler	7.5%
CPU, RAM	17.78%
Main Board, HDD	13.56%
Main Board, Mouse	9.7%
RAM, HDD	19.23%
Cooler, USB Camera	8.4%



[그림 4] 구조화된 트리

[Fig. 4] Structured tree

단일 항목으로 연관 테이블을 작성하였을 경우, CPU와 RAM 간의 항목은 연관 지을 수 있지만, 데이터에서 항목간에 연관도가 높은 RAM과 HDD에 관해서는 CPU에 대한 연관성을 추출해 낼 수 없다. 또, 마찬가지로 Main Board와 Mouse는 연관 테이블로도 찾아 낼 수가 있지만, 연관도를 무시할 수 없는 HDD는 CPU와 연관 지을 수 없다. 따라서 단일 항목간의 연관도만으로는 다중 상품 간의 연관성을 결정지을 수 없다.

또한, 구매 패턴에서의 지지도 값에 대한 변경이 있는 경우 대다수의 테이블 구조를 사용하는 방법은, 각 관련되는 상품들 모두에 이러한 변동 사항을 적용할 수 없으나 제안하고 있는 구조는 지지도 값의 변경이 발생하면 관련되는 노드간의 가중치 값에 반영되어, 최신의 구매패턴의 형태를 유지할 수 있었다. 본 실험을 통하여 단일 항목간이라도 구성 방법에 따라서 다량의 구매 데이터에 숨겨진 연관성을 찾아 볼 수 있었다.

5. 결론

본 논문에서는 인터넷 쇼핑몰에서 고객의 구매패턴을 이용하여 상품들간의 다양한 연관도를 표현할 수 있는 마이닝 방법을 제안하였다. 이는 다양한 상품군간의 간섭을 최대한으로 줄이면서 효과적으로 상품간에 연관도를 측정할 수 있는 방법이다. 기존의 쇼핑몰에서는 단일 항목간의 연관도만 반영하기 때문에, 고객의 구매 데이터에 내재된 숨겨진 다중 항목간의 연관도를 밝혀낼 수 없었다. 그러나 제안된 구조에서는 각각 떨어져 있는 단일 항목간에도 항목간의 구조화를 통하여, 다중 항목간의 연관도를 밝혀냄으로써 보다 더 개선된 상품간의 연관성을 알 수 있게 된다. 또한 거래기록 데이터베이스의 구매 정보에 새롭게 일정량의 구매 데이터가 발생되어 추가될 때마다, 구조화된 트리를 수정하여 보다 정확하고 신뢰성 높은 연관도를 구할 수 있게 하였다. 단일 항목간의 연계성을 갖는 트리는, 일반 단일 항목이 가지지 못하는 다중 상품들간의 연계성이 시각적으로 표현되기 때문에 상품들간의 연관도를 구체적으로 알 수 있고 이를 토대로 인터넷 쇼핑몰에서의 마케팅 전략수립에 도움을 줄 수 있을 것이다.

향후 과제로는 본 연구의 확장으로서 트리 수정방법의 개선을 위하여 단위시간에 많이 팔리는 품목같은 인기상품을 구조화 트리에 반영하기 위한 가중치 필터링의 연구와, 트랜잭션의 각 항목에 대한 조상(ancestor)들을 그 트랜잭션에 추가하여 알고리즘을 수행하는 일반화된 연관규칙 탐사와 웹 사용자들의 특성 파악이 효과적인 순회패턴 등을 개선하여 적용하는 것이다.

※ 참고문헌

[1] Agrawal R., Imielinski T., and Swami A., "Database Mining: A Performance Perspective", IEEE Tran. on Knowledge and Data Engineering, Vol. 5, No. 6, pp. 914-925, 1993.

[2] Agrawal R, and Srikant R., "Fast Algorithms for Mining Association Rules in Large Databases", In Proc. Of the 20th Int. Conf. on Very Large Databases, 1994.

[3] Bettini C., Wang X.S, and Jajodia, "Mining Temporal Relationships with Multiple Granularities in Time Sequences", Data Engineering Bulletin, 21:32-38, 1998.

[4] Büchner A.G., Baumgarten M., Mulvenna M.D., Anand S.S, and Hughes J.G., "Navigation Pattern Discovery from Internet Data", WebKDD '99, 1999.

[5] Cooley R., Mobasher R., and Srivastava J, "Web Mining: Information and Pattern Discovery on World Wide Web, In Proc. 9th IEEE Int. Conf. On Tools with Artificial Intelligence, 1997.

[6] Han J.,Pei J., and Yin Y., "Mining Frequent Patterns without Candidate Generation", SIGMOD '00, pp. 1-12, Dallas, TX., May 2000.

[7] Ling C.X., and Li C., "Data Mining for Direct Marketing: Problems and solutions", In Proc. 4th Int. Conf. On KDD, pp. 73-79, 1998.

[8] Michael J. A. Berry Gordon Linoff, , "Data Mining Techniques For Marketing, Sales, and Customer Support", WILEY COMPUTER PUBLISHING, 1997.

[9] Park J.S., Chen M.S., and Yu P.S., "An Effective Hash-Based Algorithm for Mining Association Rules", In Proc. Of ACM SIGMOD, pp. 175-186, 1995.

[10] Srikant R., and Agrawal R., "Mining Sequential Patterns: Generalizations and Performance Improvements", In Proc. of 5th Int. Conf. on Extending Database Technology, pp. 3-17, 1996.

[11] 박종수, "연관규칙 탐사 알고리즘에 대한 조사", July, 1998, http://cs.sungshin.ac.kr/

김 성 학

1985 건국대학교 수학과(이학사)

1987 건국대학교 대학원 전자계산학과(공학석사)

1992-현재 건국대학교 대학원 컴퓨터공학과 박사과정

1987 삼성종합기술원 정보시스템연구소 연구원

1989~ 현재 유한대학 전자계산과 부교수

관심분야: 인공지능, 데이터마이닝, machine learning