▩ 응용논문

연관규칙을 이용한 데이터 분석에 관한 연구

(A Study on the Analysis of Data Using Association Rule)

임 영 문 *
Young-Moon Leem
최 영 두**
Young-Doo Choi

Abstract

In General, data mining is defined as the knowledge discovery or extracting hidden necessary information from large databases. Its technique can be applied into decision making, prediction, and information analysis through analyzing of relationship and pattern among data. One of the most important works is to find association rules in data mining. Association Rule is mainly being used in basket analysis. In addition, it has been used in the analysis of web-log and user-pattern. This paper provides the application method in the fie이 of marketing through the analysis of data using association rule as a technique of data mining.

1. 서 론

데이터마이닝이란 데이터베이스의 양이 방대해짐에 따라 데이터베이스에 저장된 데이터에서 사용자가 기존의 질의 언어(SQL)로는 얻을 수 없는 결과에 대해 필요한 정보를 획득하는 방법에 관한 연구로서, 사용자가 사전에 예측 할 수는 없지만 미리 예정된 일정패턴의 정보를 찾아내는 일에 집중하고, 데이터에 존재하는 일관된 흐름이나 경향 패턴을 파악하는 것이다. 연관규칙은 장바구니 분석에 주로 많이 이용되고 있으며, 최근 웹 로그분석 및 사용자 패턴 분석에도 이용되고 있는 실정이다.

그러나 오류가 있는 데이터에는 아무리 훌륭한 방법을 적용한다고 해도 얻어진 결과에 대해서는 신뢰성을 가질 수 없다. 그러므로, 양질의 데이터가 보장된 후에야 비로소 효과적인 데이터마이닝을 수행 할 수 있다. 이러한 부분들은 데이터웨어하우스가 구축되어 있다면 보다 쉽게 해결될 수 있다.

본 연구에서는 데이터웨어하우스로 잘 구축되어진 FoodMart의 데이터를 이용하여 기본적 통계분석 및 연관규칙을 이용한 분석을 통하여, 마케팅분야에 응용할 수 있는 방법을 제시 해 보고자 한다.

* 강릉대학교 산업공학과 교수
** 강릉대학교 산업공학과 석사과정

1.1. 연구동기 및 목적

현대 산업구조는 저 비용의 고 부가가치 창출을 견지하면서 고도의 정보화 시대로 바뀌어 가고 있는 사회구조의 변화에서 기존에 구축되어 있는 데이터 베이스를 이용한 연구활동이 활발해 지기 시작하였으며, 초기 데이터마이닝은 KDD (Knowledge Discovery in Database)의 활동에 의해 시작되었고, 현재는 데이터마이닝과 KDD에 구별을 두지 않고 사용되고 있다. 정보의 가치가 중요시되면서 데이터마이닝에 대한 성공적인 사례는 마케팅 분야로써, 각각의 마케팅 활동에 대해서 적절하게 차별화된 방법을 사용하는 것이 아마 가장 현명한 방법이 될 것이다.

이와 같이 다양한 규칙을 파악하고 응용하면 많은 이득을 얻을 수 있다는 것은 분명하다. 그러나 데이터마이닝이 마케팅에서만 사용되는 것은 아니며, 터빈유지 보수 데이터를 가진 데이터베이스, 출력 인쇄기로부터의 고장 리포트, 기상관측 데이터의 수정자료등 여러 분야에서 새로운 지식을 유도할 수 있는 풍부한 원천이 될 수 있다. 대규모 데이터가 있는 곳은 어디에라도 유용하고 새로운 응용을 찾을 가능성이 있는것이다.

따라서 본 연구에서는 체인 판매망을 가진 FoodMart체인점의 데이터를 가지고, 데이터마이닝의 기법에서 연관규칙을 이용한 결과분석을 제공하고자 한다.

1.2. 데이터마이닝의 발전

데이터마이닝에 관한 관심이 이 분야의 짧은 역사에도 불구하고 급속히 상승한 것은 다음의 몇 가지 요인에 의하여 부분적인 설명이 가능하다[13].

1. 데이터마이닝 기법은 방대한 데이터베이스에서 최적화된 집단을 편성하거나 의미 있는 규칙성을 찾아낸다. 이에 반하여 SQL은 질의자가 이미 알고 있는 제약 조건하에서 어떤 데이터를 찾는데 도움을 주는 질의 언어이다. 대개 데이터마이닝 알고리즘은 관심의 대상이 되는 데이터베이스의 어느 한 부분을 집중적으로 탐색한다. 대부분의 경우에 데이터마이닝 알고리즘은 많은 SQL 질의를 반복적으로 사용하기도 하며 그 중간 결과를 저장하기도 한다. 물론 이러한 작업을 수작업으로 할 수 있으나 수작업으로 수행되는 작업은 단순 반복적인 지루한 작업이 될 것이며 많은 시간을 소비하게 될 것이다.

2. 네트워크의 사용이 지속적으로 증가함에 따라 여러 데이터베이스에 접속하는 것이 훨씬 쉬워졌다. 따라서 고객정보파일을 인구 통계학적인 파일들과 연결시켜 고객의 특정 집단에 대한 소비 행위를 파악함으로서 이제까지 알 수 없었던 전혀 새로운 지식을 얻을 수 있게 되었다.

3. 과거 수년간에 걸쳐 기계 학습 기법(Machine Learning Technique)이 급속히 발전하였다. 신경망 이론, 유전자 알고리즘(Genetic Algorithm), 및 일반적인 학습 기법 등의 발전으로 데이터베이스에서 지식을 발견하는 것이 쉬워졌다.

4. 클라이언트/서버 구조의 출현으로 개인 지식노동자들은 자신의 책상에서 클라이언트를 통하여 중앙에 있는 정보시스템에 쉽게 접근할 수 있게 되었다. 이에 따라 마케팅 전문가나 기획 전문가 등도 역시 이러한 새로운 기술을 자기 업무에 활용하기를 원하였다.

1.3. 데이터마이닝관련 기법의 종류

데이터마이닝에 사용되는 기술은 몇 가지로 나열할 수 있는데, 먼저 신경망(Neural Network) 알고리즘은 예측, 군집, 분류에 사용되며 신경망은 복잡하고 불완전한 데이터로부터 의미를 유추하는데 탁월하며 패턴을 추출하거나 매우 복잡한 동향을 탐지하는데 이용될 수 있

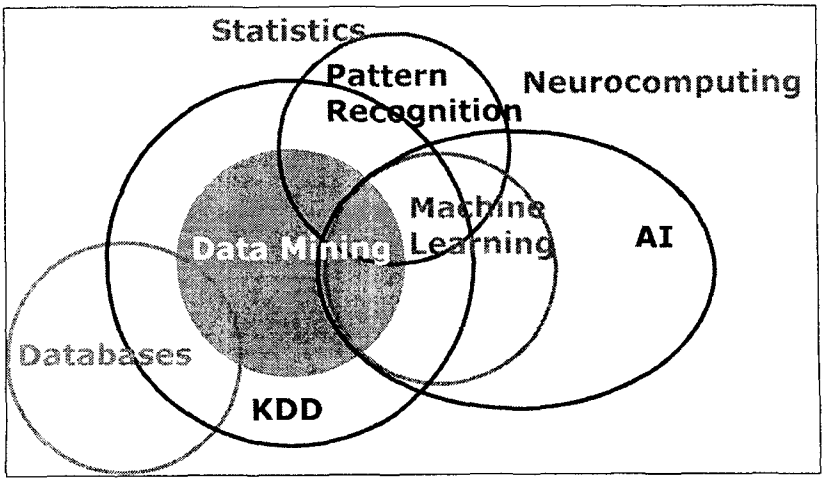
다. 훈련된 신경망은 전문가 수준으로 여겨질 정도이기 때문에 문제에 대해 예측하고 what-if 질문에 응답하는데 이용될 수 있다.

·근접이웃(Nearest Neighbor)은 군집분석과 분류에 가장 적합하다. 우선 레코드를 다차원 속의 한 개의 점으로 나타낸 후, 유사성이 높은 목적속성을 갖고 있는 데이터 점을 한 개의 강하게 결합된 클러스터가 되도록 각 차원의 가중치를 튜업한다. 예를 들면, 은행의 목적차원이 이차저당 신청자라고 가정할 때, 차원은 나이, 일차저당액수, 빚, 수입등 일 수 있다. 알고리즘은 이차저당을 받은 사람이 가장 가까운 이웃으로 모일 수 있을 때까지 각기 차원에 대한 가중치를 증가시킨다.

의사결정나무(Decision Tree)는 한정된 수의 클래스로 예를 분류하는 것으로 군집화와 분류에 적합하다. 군집화는 모집단을 유사한 특성에 따라 세그먼트로 나누는 것이다. 예를 들면, 보험회사는 어떤 고객속성이 높은 클레임과 관계가 깊은지를 알고 싶어한다. 알고리즘이 가장 영향력 있는 속성은 결혼상태라고 결정한다면 모집단을 결혼과 미혼의 두개의 클러스터로 나누고, 그 다음 영향력 있는 나이, 소유차종, 거주지 등으로 나누어 분할한다. 이 알고리즘에서는 예측을 보다 정교하게 하기 위해 각 분할된 클러스터에 통계적인 유의성을 부여한다. 하나의 대상이 어떤 클래스에 속하는지를 결정하기 위해 개념을 나무구조로 표현한다. 일반적으로 효율성을 위해 한 번에 전체 데이터에 대하여 의사결정나무를 구성하는 방법보다는 초기에 랜덤하게 선택된 작은 부분집합을 설정하는 것으로부터 출발하여 전체 데이터로 확대해 나간다. 나무의 각 노드는 속성명칭을, 각 노드에서 분기되는 아크는 이들 속성이 가질 수 있는 값들을 의미하며 최하위 노드는 여러 클래스로 분류되었음을 나타낸다.

유전자 알고리즘(Genetic Algorithm)은 진화이론에서와 같이 진화적인 방법으로 계산을 변형시키는 방법이다. 어떤 문제에 대한 잠재적인 해결책들을 상호 경쟁적인 것으로 간주한다. 최선의 해결책이 선택되고 상호 비교되는데, 승리자의 생존원리와 같이 최선의 해결책이 살아 남게 되는 것이다. 유전자 알고리즘은 신경망과 자주 결합해 사용된다. 예를 들면, 신경망 알고리즘이 몇 개의 가능한 모델을 파생시키면, 유전자 알고리즘이 이중에서 최선의 모델을 선정하게 된다. 이 알고리즘은 양호한 대출신청자를 식별하는 것과 같이 많은 변수를 갖고 있는 어려운 최적화 문제를 해결하는데 적합하다.

연관규칙(Association Rule)은 시장바구니 분석의 단점을 극복하기 위해서 사용된다. 전형적인 시장바구니 분석방식은 여러 항목(시장 바구니의 내용물들)의 구입을 하나의 거래로 처리한다. 이렇게 하는 목적은 자연스러운 구매 패턴을 이해하고 발견하는데 사용될 수 있는 엄청나



[그림 1] 데이터마이닝 구성 분야

게 많은 거래 사이에서의 특정한 경향을 찾아내기 위한 것이다. 연관성 분석은 구입항목의 집합에서 하나의 구입상품 또는 구입상품들 집합의 존재가 또 다른 구입상품의 존재를 암시하는 규칙을 발견하고자 하는 것이다.

연속성 기반 분석(Sequence Base Analysis)은 전형적인 시장바구니 분석의 문제점인 항목들의 조합을 거래 시점의 일부분으로 다루게 된다. 이 문제의 변수는 연속된 시간선상에서의 구매의 수열(예를 들어 계좌번호, 신용카드 또는 거래가 잦은 바이어의 번호)을 하나로 묶어주는 추가적인 정보가 있을 경우에 생기게 된다. 이 경우 여러 항목이 하나의 트랜잭션에 공존하는 것만이 중요한 것이 아니라 주문된 트랜잭션이나 트랜잭션 사이의 시간적 양으로 나타나는 주문 역시 중요하다. 이러한 관계분석을 통해서 얻어진 규칙은 특정 항목의 연속적인 구매형태를 예측하는 도구로 사용될 수 있다. [그림1]은 데이터마이닝을 사용하는 분야별 구성분야를 나타내고 있다.

2. 연관규칙 정의

2.1 연관규칙 탐사(Association Rule Mining)

연관규칙은 빈발 항목집합에 대한 연관규칙을 생성해 내는 작업으로 나눌 수 있다. 전체 항목집합(상품들의 집합)이라 불리는 문자들의 집합을 $I=\{i_1, i_2, \cdots, i_m\}$ 라고 주어지면, D는 트랜잭션(Transaction ; 판매기록 - 판매테이블의 한 레코드)의 집합이고, 각 트랜잭션 T는 $T \subseteq I$인 항목들의 집합이라고 정의된다. 한 트랜잭션에서 구입하는 항목들의 양에는 상관하지 않고, 각 항목은 구입여부를 나타내는 이진 변수이다. 각 트랜잭션은 TID라 불리는 식별자와 연계되어 있다. X를 항목들의 집합(Itemset)이라 하고 트랜잭션 T가 X를 포함한다면 $X \subseteq T$이라 한다. 연관규칙은 $X \Rightarrow Y$의 형식으로 표시되고, $X \subset I$, $Y \subset I$ 및 $X \cap Y$는 공집합이다. 또한 $X \Rightarrow Y$ 신뢰도(Confidence) c를 가지고 있다는 것은 X를 포함하는 D의 트랜잭션들 중 c%가 Y도 포함하고 있음을 의미한다. 또한 $X \Rightarrow Y$는 트랜잭션 집합 D내에 지지도(Support) s%를 갖는데 이는 D의 트랜잭션들 중 s%는 $X \cup Y$를 포함하고 있음을 뜻한다. 최소 지지도(Minimal Support) 이상을 갖는 항목집합을 빈발항목집합(Large Itemset or Frequently Itemset)이라 한다. K개의 항목들로 이루어진 빈발 항목집합을 빈발 k-항목집합이라고 한다. 빈발 k-항목 집합들의 집합을 Lk라하고 이를 위한 후보 k-항목집합들의 집합을 Ck라고 한다.

연관규칙을 발견하는 과정은 다음 두 단계로 구성된다. 첫째는 빈발항목 집합들을 찾는다. 즉 주어진 최소 지지도 이상의 트랜잭션들의 지지를 갖는 항목 집합들의 모든 집합을 찾는것이며, 둘째는 데이터 베이스에 대한 연관규칙을 찾기 위해 빈발항목집합을 사용한다. 발견된 연관규칙은 지지도 및 신뢰도를 가지는 수치 값으로 정량화 된다[1,2,3,4,5,6].

2.2 Apriori Algorithm

Apriori 알고리즘은 각각의 반복시행에서 빈발 항목 집합들에 대한 후보 집합을 생성하고 각 후보 항목 집합의 빈도수를 계산한 후 주어진 최소 지지도 이상의 빈발 항목 집합들을 결정하게 된다. 최초의 반복시행에서 Apriori는 각 항목에 대한 빈도수를 계산하기 위해 모든 트랜잭션을 읽는다. Apriori 알고리즘은 이제까지 발표된 알고리즘들 중에서 후보 항목집합을 생성하는 아주 독특한 방법을 사용하였으므로[7,8,9,10,11,12], 여기에서 주어진 데이터베이스[표1]에서 빈발항목집합들을 탐사하는 전형적인 알고리즘으로 사용한다. Apriori에서는 각 패스에서 빈발

[표 1] 트랜잭션 데이터베이스 D의 예

TID	Items
100	ACD
200	BCE
300	ABCE
400	BE

항목집합의 후보 집합을 구성하고 난 후에 후보 항목집합의 발생 빈도수를 계산하고, 미리 결정된 최소지지도를 기초로 하여 빈발항목집합을 결정한다. 첫 번째 패스에서 Apriori는 각 항목의 발생빈도수를 세기 위하여 단순히 모든 트랜잭션들을 스캔하여 읽는다. 후보 1-항목집합들의 집합 C1은 [표2]에서와 같이 얻어진다. 최소 트랜잭션 지지도가 2라고 가정하면 (s_{min} = 2), 필요로 하는 최소지지도를 갖는 후보 1-항목집합들로 구성되는 빈발 1-항목집합들의 집합 L_1이 결정될 수 있다. 빈발 2-항목집합들의 집합을 탐사하기 위해서는, 한 빈발항목집합의 한 부분집합이라도 역시 최소지지도를 가져야 한다는 사실에 입각하여 Apriori는 ACG(Apriori Candidate Generation)을 사용하여 후보 항목집합들의 집합 C_2를 생성하기 위해 $L_1 * L_1$을 사용하였다. *는 접속(concatenation)연산자다. C_2는 $\binom{|L_1|}{2}$개의 2-항목집합들로 이루어진다. $|L_1|$이 크게되면 $\binom{|L_1|}{2}$는 급속도로 큰 숫자가 된다. 다음으로 D에 속한 네개의 트랜잭션들이 되어 읽히고 C_2에 속한 각 후보 항목집합의 지지도는 계산된다. [표2]에서 두 번째 행의 가운데 테이블은 C_2에 속한 후보 항목집합의 지지도 계산결과를 나타낸다. 해쉬트리(hash tree)는 보통 바르게 지지도 계산을 위해 사용된다. 빈발 2-항목집합들의 집합 L_2는 C_2에 속한 각 후보 2-항목집합의 지지도에 기초하여 결정된다.

후보 항목집합들의 집합 C_3는 L_2에서 다음과 같이 생성된다. L_2에서 첫 항목이 같은 두 개의 빈발 2-항목집합들을 먼저 확인한다. 예를 들면, $\{BC\}$와 $\{BE\}$에서는 B가 동일한 항목이다. 다음으로 Apriori는 $\{BC\}$와 $\{BE\}$의 두 번째 항목들로 구성된 2-항목집합 $\{CE\}$가 빈

[표 2] 후보 항목집합들과 빈발항목집합들의 생성

C_1

Itemset	Sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

SCAN D ===>

L_1

Itemset	Sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{AB}
{AC}
{AE}
{BC}
{BE}
{CE}

C_2

Itemset	Sup
{AB}	1
{AC}	2
{AE}	1
{BC}	2
{BE}	3
{CE}	2

SCAN D ===>

L_2

Itemset	Sup
{AC}	2
{BC}	2
{BE}	3
{CE}	2

C_3

Itemset
{BCE}

C_3

Itemset	Sup
{BCE}	2

SCAN D ===>

L_3

Itemset	Sup
{BCE}	2

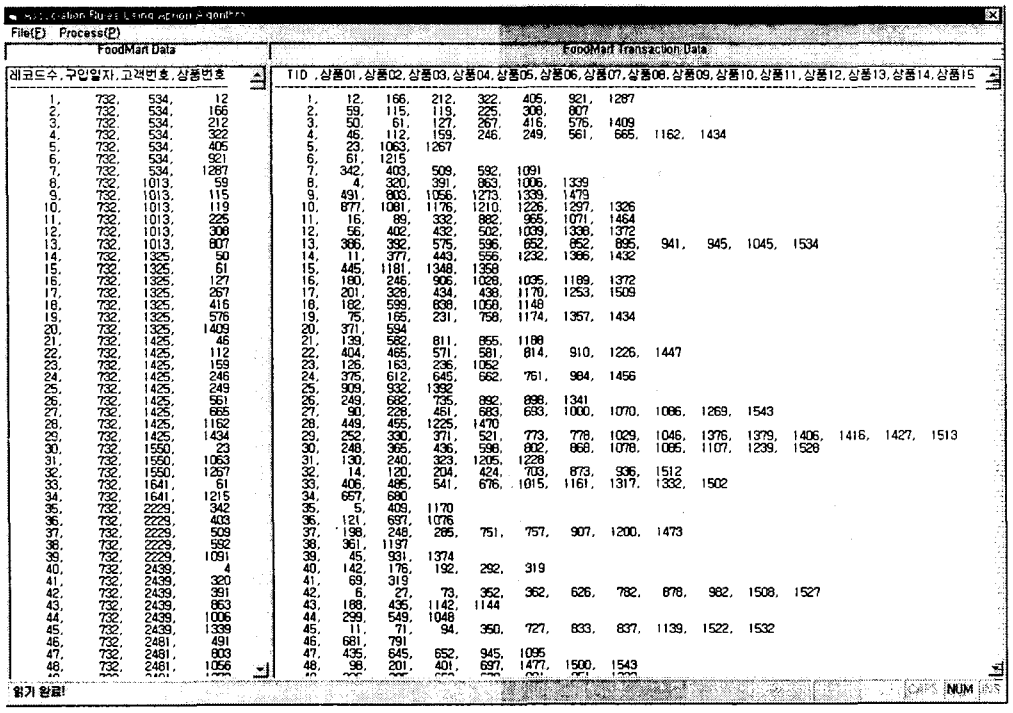
발 2항목집합들에 속하는 지를 검사한다. {*CE*}가 그 자신이 빈발집합이므로, {*BCE*}의 모든 부분집합들은 빈발하다는 것을 알았고 그러므로 {*BCE*}는 후보 3-항목집합이 된다. L_2에서 더 이상의 다른 후보 3-항목집합을 구할 수 없다. 그러면 Apriori는 모든 트랜잭션들을 스캔하면서 [표2]와 같이 빈발 3-항목집합을 발견한다. L_3에서부터 구성될 수 있는 후보 4-항목집합이 없으므로, Apriori는 빈발항목집합을 발견하는 과정을 마친다. 앞에서 본 것처럼, C_i에 속한 각 항목집합의 지지도를 전체 데이터베이스를 스캔하면서 계산해야 하기 때문에 가능한 후보 항목집합들의 원소들의 개수가 줄여지도록 후보 항목집합들을 생성하는 것이 중요하다.

3. 실험 방법

본 연구를 위한 시스템 구성을 위해 Apriori알고리즘을 MS-SQL Server 7.0, Visual Basic 6.0과 ADO(ActiveX Data Object)로 구현하였으며, FoodMart의 데이터로부터 특성을 파악한 뒤, 필요 항목을 SQL 질의를 통하여, Txt파일이나, MDB파일로 트랜잭션 파일을 생성한다.

프로그램의 개략적인 구성은 크게 3가지 단계로 구성할 수 있다. 판매 일자와, 고객에 따른 판매 상품이 순차적으로 저장되어 있는 것을 트랜잭션 데이터 형태로 변환하기 위한 작업이 전처리 단계인 첫 번째 단계([그림2])이고, 두 번째 단계로는 연관규칙의 Apriori알고리즘을 적용하기 위한 단계로 빈발항목과 빈발 후보 항목에 대한 질의를 구성하여 결과를 출력하는 부분이며, 마지막 단계로는 알고리즘 적용 후 생성된 연관규칙을 가지고, 분석하는 단계로 구분할 수 있다.

맨 처음 모든 데이터를 구입일자, 고객번호, 상품번호 순으로 질의하여 MS-SQL에 임포트한 후, MS-SQL에서 1-항목집합의 지지도 수를 카운트 한 뒤, 질의 결과를 MDB파일이나,



[그림 2] 전처리과정 후 생성된 데이터와 트랜잭션 데이터

Txt파일로 트랜잭션을 저장한다. 그 후 Apriori의 알고리즘을 적용하여 2-항목집합서부터 k-항목집합까지 구하기 위한 루프를 적용한다. 그후 도출된 연관규칙에 대하여, 관련상품과의 관계를 파악, 분석한다.

4. 데이터 분석

본 연구에서 사용하는 데이터는 FoodMart로 MS-SQL 7.0 OLAP 예제 데이터이며, 다국적 체인형태의 잡화점으로 1997년과 1998년 상품 판매 실적을 기록한 데이터이다. 본 실험의 장바구니 분석을 위한 데이터로는 충분하다고 생각되며, 다만, FoodMart데이터가 인위적으로 만들어졌는지, 실제 판매 데이터에 대한 기록인지에 대한 내용은 파악되지 않았다. 실제 판매 데이터를 구하기는 현실적으로 매우 어려웠으며, 실험에 사용된 FoodMart의 데이터 테이블 구조를 살펴보면 총 15개의 테이블로, 상품테이블, 고객테이블, 재고테이블, 판매테이블, 상품분류항목테이블, 체인점 테이블 등으로 이루어져 있으며, 체인점의 소유형태에 따라 5개 범주로 구성되어 있다. [표3]은 1997년과 1998년 데이터의 특성에 대한 비교를 보여준다.

[표 3] FoodMart 데이터 연도별 비교

	1997년	1998년
레코드수	86,837건	164,558건
트랜잭션수	20,522건	34,015건
평균 상품구매수	4.838개	4.2288개
최대 상품구매수	17개	28개

FoodMart의 상품 판매 데이터는 sales_ fact_1997과 sales_fact_1998 테이블이 있는데, 각각 상품판매 기록으로는 86,837건과, 164,558건이 있으며, 1997년 판매기록의 경우는 20,522건의 트랜잭션 데이터를 얻을 수 있었으며, 평균 4.838개의 상품구매와, 최대 17개의 상품을 구매 한 것으로 나타났다. 1998년의 데이터를 가지고 트랜잭션 데이터를 생성해 본 결과 34,015건의 트랜잭션이 발생되었으며, 이는 한 고객이 평균 4.2288개의 상품을 구매하였고, 최대 28개의 상품구매 한 것으로 나타났다. 등록된 고객은 10,281명이며, 상품은 1,560개에 110개의 범주로 분류된다.

FoodMart의 각 테이블의 관계도를 보면, 각 구성은 관련 코드와 체인점 정보, 고객 정보테이블이 판매테이블에 결합되어 있는 것을 볼 수 있다. 이상으로 FoodMart의 데이터 구조를 살펴보았으며, 상기 자료를 가지고 데이터베이스 내에서 연관규칙을 찾기 위한 절차는 크게 3단계로 나뉘어 지는데, 첫 번째가 전처리 단계이며, 두 번째는 마이너 구성, 세 번째가 결과유도의 과정을 가진다. 여기서는 판매기록 테이블과, 고객테이블, 상품테이블 및 상품 분류 테이블을 가지고 전처리 과정을 거치면, [그림2]의 왼쪽 프레임과 같은 데이터를 얻을 수 있으며, 알고리즘 적용에 필요한 트랜잭션 데이터 형태로 만들면 [그림2]의 오른쪽 프레임과 같은 예제 데이터 베이스 [표1]의 형태를 얻을 수 있다. 여기서 생성된 총 54537건의 트랜잭션에 대해 연관규칙을 적용한 뒤 생성된 연관규칙에 따라, 1997년과 1998년의 데이터에 적용, 연도별 트랜잭션에서 생성된 연관규칙을 가지고 총 트랜잭션에 적용하여 비교 분석하였다.

4.1 고객분석

고객분석은 고객으로 등록되어 있는 10,281명에 대한 기본적인 통계적 분석으로 연관규칙에 의한 결과분석에 도움을 주는 것으로, 분석의 기본적인 과정이다. 통계적 분석을 이용한 수치적 평가를 가진 후, 연관규칙으로 생성되는 결과와 연결하여 최종 결론을 유도 할 수 있다. 고객 관련 데이터로 분석할 수 있는 기본적인 사항에 대한 분석결과이다.

고객자료는 3개국에 걸친 자료로서, 캐나다, 멕시코, 미국으로 나타났으며, 미국에 거주하는 고객의 경우가 71%로 가장 높은 비율을 가진 것으로 나타났다. 거주 주별로 고객자료를 분석해 본 결과, 캐나다 1개주, 멕시코 9개주, 미국 3개주로 나타났으며, 미국의 CA주가 20%로 가장 많은 비율이었으며, 캐나다의 BC주가 17%, 멕시코에서는 Zacatecas주가 2%를 차지하고 있었다. 미국의 CA에서 높은 비율을 나타내는 것은 시별 구성비율을 살펴보면 알 수 있듯이, 많은 도시에 체인점을 가지고 있기 때문이다.

거주 시별로 고객 데이터를 분석해본 결과, 가장 많은 시를 보유한 주는 CA주로서 45개의 시로 구성되어 있으며, 각 도시마다 등록된 평균 고객 수는 84.321명이며, 최대 고객보유시는 캐나다의 BC주 Shawnee시로 나타났다. 따라서 결과를 종합하면 [표4]와 같이 정리할 수 있다.

[표 4] 고객데이터 정리표

국가명	주수(개)	도시수(개)	고객수(명)
캐나다	1	18	1717
멕시코	9	24	1205
미국	3	78	7359

교육 수준별 고객구성비율을 보면, 교육등급은 5등급으로 나뉘어져 있으며, 구분은 Bachelors Degree, Graduate Degree, High School Degree, Partial College, Partial High School로 구분되어 있으며, 전체 구성에서 Partial High School의 고객이 가장 많은 것으로 나타났다. 대졸의 경우는 캐나다가 가장 높은 비율을 가지나, 멕시코와 미국과의 유의한 차이가 있는가의 검정은 이루어지지 않았다.

고객의 연봉에 따른 비율을 살펴 본 결과, 최저 연봉인 만불에서 삼만불의 경우 멕시코의 Jalisco주이며, 가장 높은 연봉 구분인 십오만불 이상의 경우에는 멕시코의 Sinaloa주로 나타났다. 연봉이 높고, 인구가 많은 곳은 매출분석에서 높은 수치를 나타내는 것을 알수 있다.

마지막으로 고객의 성비와 결혼여부, 자녀수에 따른 비교를 해보았으며, 이는 구매경향의 성비와, 결혼여부에 따른 매출차이, 자녀수에 따른 비교를 해보고자 한 것이다. 남녀 비율의 차이는 거의 나타나지 않고 있으며, 전체 고객에서는 여성보다 남성의 고객이 조금 더 높은 것으로 나타나고 있다. 결혼여부에서는 미혼과 기혼의 차이는 거의 비슷하였으며, 멕시코의 Yucatan주에서만 기혼자가 미혼자보다 약 9%의 차이를 보이고 있다. 그리고 자녀수에 따른 비교에서는 자녀가 없거나 혹은 1명 가진 경우가 상대적으로 높은 비율을 가지고 있으며, 5명 이상인 자녀를 보유한 지역도 많은 것으로 나타났다.

4.2 광고분석

광고는 제품 판매활동에 없어서는 안될 중요한 비중을 차지하고 있는 구성 중에 하나이다. 아무리 좋은 제품이라도 광고가 부족하여, 인지도가 낮다면, 결코 많은 수익을 만들 수가 없다. 정보화 시대에 있어서 광고의 영향과 효과는 상당히 큰 것으로 효과적 광고분석으로 광고가 매출에 주는 영향 및 소요예산등의 분석을 통해 보다 빠르게 매출 및 소비자의 반응에 대처할 수 있을 것으로 보여 진다. 이것 또한 연관규칙을 위한 기본자료로 사용되어 지고 있다. 먼저 광고의 형태별, 소요예산별 구성을 보고자 한다.

광고 소요비용은 3년간 18,690,519달러의 비용을 소비하였으며, 매체별 광고비용을 보았을 때는 최고 9%에서 최저 5%의 비용을 지불하고 있으며, 13개 매체형태로 나뉘어 광고가 되고 있었다. 생각과는 다르게 순수 TV광고에는 5%의 비율이었고, Radio에는 9%의 비율로 광고가 되었다.

광고 명으로 분석을 해 본 결과 Prince Winners, Sales Winners의 광고에 많은 비용을 소비

하였으며, 매출분석에서 광고에 의한 수입의 범위가 어느 정도인가에 대한 내용을 확인할 수 있다. 대부분의 광고방영은 1주일 미만이며, 최고 29일까지의 광고도 있었고, 2일과 3일간의 광고홍보에 가장 큰비용을 소모한 것으로 나타났다.

4.3 매출분석

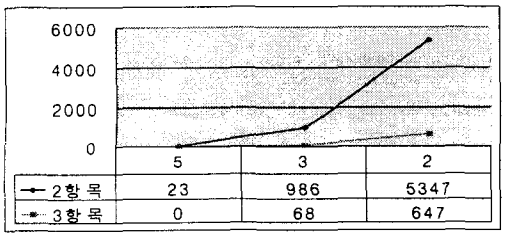
매출분석은 판매테이블에서 각 연도별 자료를 기준으로 연도별, 분기별 판매된 현황과, 매출성과를 비교해 보고자 하는 것으로, 시간에 관련된 매출의 흐름을 분석한다. 1997년 판매데이터에서 Daily Paper, Radio, TV에 동시광고를 한 것이 가장 큰 이익을 주고 있으며, 반면 단순 Radio광고만 한 경우에는 가장 낮은 이익을 가져다주는 것으로 나타났다. 이익률은 거의 비슷한 수준이지만, 특히 식품판매가 보다 많은 이익을 주고 있는 것으로 나타났다. 1997년 판촉활동 시기의 수익은 전체 수익의 26.7%를 차지하고 있지만, 실제 1년 365일에서 일수별 비율을 계산한 뒤에 비교하여 보면 판촉시기에 판매된 이익이 더 많이 나타날 것으로 보인다. 1998년의 경우는 1997년과 동일하게 식품류판매가 많은 이익을 주는 것으로 나타나고 있었으며, Daily Paper, Radio에 동시광고를 한 것이 가장 높은 이익을 주는 것으로 나타나고 있었고 1998년 전체수익의 26.9%를 차지하고 있다. 전년대비 광고수익은 크게 성장하였다고 할 수 없었다. 연도별 분기별 수익은 거의 비슷하였으며 4분기 판매가 상대적으로 매출이 낮은 것으로 나타났다.

각 판촉활동별 제품군에 따른 판매이익은 3개 제품군에서 Cash Register Lottery와 Price Saver의 판촉기간동안 많은 수익이 발생한 것으로 나타났다. 매출분석에서는 기본적인 고객의 자료를 가지고 매출과 관련된 결과를 찾는 것이지만, 연관규칙은 통계적 분석 위에 각 고객의 구매패턴을 찾는 것으로, 이후에 판매될 상품에 대한 예측 및 계획을 수립할 수 있게 해 주는 것이다.

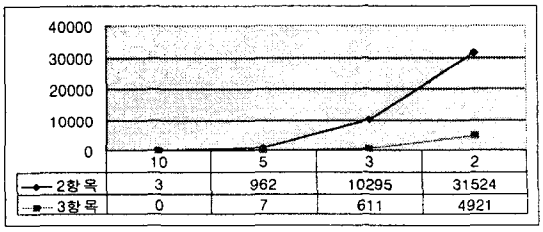
4.4 연관규칙 적용분석

연관규칙은 고객의 물건 구매행위에 대한 경향이나 패턴을 찾는 것으로, 전체 구매데이터에서 고객이 A라는 물건을 구매한 뒤에 B라는 물건을 살 확률을 구하는 것으로, 여러 항목집합과, 판매되는 물건들의 특성 및 그러한 물건을 구매한 고객들의 특성 등을 분석하여, 판매 경향이나, 마케팅 전략 수립에 많은 도움을 줄 것이다.

FoodMart의 데이터를 분석해 본 결과 1997년은 최소 지지도가 5이상(0.025%)인 경우와, 1998년은 10이상(0.03%)인 경우에 관련 상품은 없는 것으로 나타났으며, 최소 지지도가 5인 경우의 1997년 빈발 2항목은 23건, 3항목은 0건이며, 1998년의 경우 빈발 2항목은 962건, 빈발 3항목은 7건으로 나타났다. [그림3]와 [그림4]에서 이런 차이를 보이는 이유는 동일한 상품(1,560개)에 대한 상품판매 건수 즉 트랜잭션 데이터가 다르기 때문에 빈발 항목의 건수도 차이가



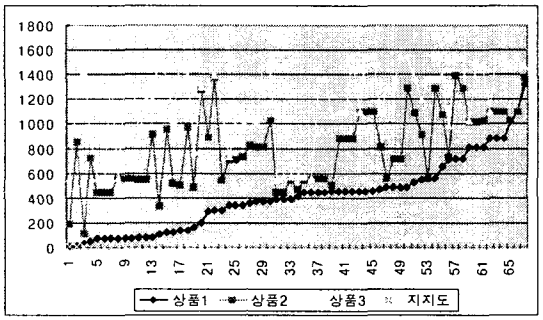
[그림 3] 1997년 지지도별 항목집합 개수



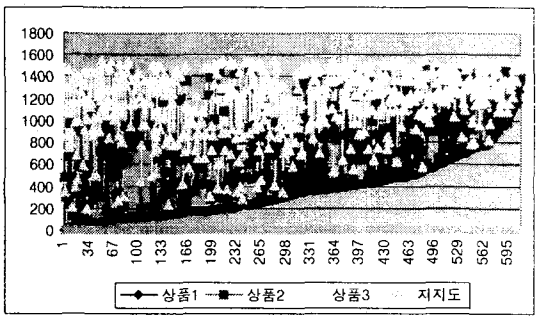
[그림 4] 1998년 지지도별 항목집합 개수

난 것으로 분석된다. 1997년의 경우 최소 지지도가 워낙 낮기 때문에 이 규칙에 의한 실적용에 있어서 우연에 의한 일치에 대한 생각도 고려해야 한다.

[그림5]와 [그림6]은 1997년, 1998년 지지도 3인 경우의 상품데이터를 나타내고 있는 것으로, 상품1을 기준으로 상품2와 상품3에서는 특정상품에 대해 군을 형성하고 있는 것을 볼 수 있다. 1997년과 1998년 지지도3의 결과를 분석한 결과 1997년은 식품류가 67건중 48건으로 72%를 차지하고 있으며, 음료의 경우 1건으로 1%를 차지하고 있고, 1998년은 식품류 611건중 396건으로 65%를 차지하고 있으며, 음료의 경우 33건으로 5%를 차지하고 있다. 이를 보았을 때 1997의 소비패턴이 음료의 경우 4%가 늘었고, 식품류는 7%가 줄었다. 3항목 모두 한 군으로 모여 있으며, 다른 군과의 관련은 없는 것으로 나타났다.

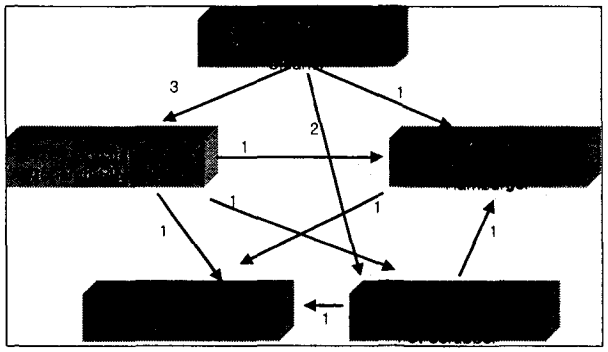


[그림 5] 1997년 3항목집합(지지도 3)



[그림 6] 1998년 3항목집합(지지도 3)

마지막으로 1997년, 1998년 결과에서 상품별 빈도수가 높은 항목을 예를들어 살펴보면 [그림7]과 같이 나타난다. 즉 레드윙사의 클리너를 구입하면, 모나크사의 쌀을 사는 경우와 하이 퀄리티사의 수세미를 사는 사람과, 젠틸사의 햄버거를 사는 경우로 나타나며, 각 상품마다 분기를 숫자로 표시 해 두었다.



[그림 7] 상품코드 454의 분기현황(6건)

5. 결론 및 향후과제

본 연구에서 취급한 FoodMart 데이터를 분석 해 본 결과, 우선 판매량에 있어서 자료와 상품종류가 많을수록, 지지도가 낮을수록 알고리즘을 적용하는데 많은 시간이 소요되었으며, 각 연관규칙에 의한 지지도와 신뢰도를 바탕으로 판매전략에 대한 자료를 제시할 수 있는데 결론을 요약하면 다음과 같다.

1. 국가별 거주 비율에서 미국 거주 고객이 71%를 차지하고 있으며, 거주 주별 고객분포에

서는 미국 캘리포니아주에 41%가 거주하고 있으며, 거주 시별 고객분포에서는 캐나다의 BC주 Shawnee에 111명의 고객이 있는 것으로 나타났으며, 시별 평균 고객수는 84.321명으로 나타났다.

2. 교육수준별 고객구성비율에서 1에 의한 인구밀도가 높은 지역의 교육수준을 분석하였을 때 교육수준은 5개 등급으로 분류되어 있고, 비슷한 비율로 분포되어 있으며, 고학력자의 경우 캐나다에 상대적으로 더 높은 분포가 있는 것으로 나타났다.

3. 연봉수준별 고객구성비율은 8등급으로 최저 3만 달러에서 최고 15만달러 이상의 소득자로 구분되어 있으며, 최저 등급에 속하는 고객은 멕시코의 Jalisco주이며 가장 높은 등급은 멕시코의 Sinaloa주로 나타났다. 연봉이 높고, 인구가 많은 곳에서 상대적으로 높은 매출성과가 있는 것으로 나타났다.

4. 전체고객과 국가별 성비중 여성의 비율이 조금 더 높았으나 차이가 거의 없었고, 멕시코의 Yucatan주에서만 기혼자가 미혼자보다 약 9% 높게 나타났고, 보유자녀수에 따른 고객구분에서 자녀가 없거나 한 명인 고객이 전체의 66.9%로 나타났다.

5. 판촉활동을 위한 광고 소요비용은 연간 평균 6,230,173달러의 비용이 소비되었고, 매체별 광고비용에서는 TV광고에 5%, Radio광고에 9%의 비율로 소비되었다. 광고방영은 1주일 미만에서, 최고 29일까지의 광고 기간을 두고 있었으며, Price Winners와 Sale Winners기간에 가장 많은 소요비용이 발생하였다.

6. 매체별 수익분석에서 Daily Paper, Radio, TV에 동시 광고한 상품이 상대적으로 높은 이익을 도출하였으며, 5%와 9%의 광고비용을 지불한 TV, Radio 광고상품은 상대적으로 낮은 이익을 초래한 것으로 나타났으며, 광고대 비광고 상품의 판매이익 비율은 6:4의 비율로 광고할 경우가 더 높은 판매 이익효과가 있는 것으로 나타났다. 상품 군별 판매이익은 식품류에서 월등히 높은 비율이 있는 것으로 나타났다.

7. 연관규칙 적용에 있어서 낮은 지지도(0.03% - 0.005%)를 나타냈으며, 전체 지지도가 낮은 이유는 상품판매 건수 즉 트랜잭션에 비하여 빈도수가 낮기 때문이며, 최소지지도는 1997년의 경우 5(0.025%)이며, 1998년의 경우 10(0.03%)으로 나타났으며, 항목집합 리스트는 2항목과 3항목에만 나타나고 있다.

8. 데이터를 분석한 결과, 통계적 기본자료도 중요하지만, 대규모 데이터베이스에서 데이터간의 연관성에 대해 조사해 보는 것에도 큰 의의가 있었으며, 데이터마이닝의 연관규칙도 확률의 개념에서 나온 것으로 어디까지나 의사결정의 보조수단으로 사용되어져야 하며, 결과를 여러 각도에서 이해, 해석하는 것이 필요하다고 생각된다.

데이터 마케팅은 어디까지나 의사결정 보조 수단으로 사용되어져야 하며, 의사 결정자는 분석된 결과를 잘 이해하고 분석할 수 있어야만 효과의 극대화를 얻을 수 있다. 또한 효율적 데이터베이스 마케팅을 위하여 여러 가지 알고리즘의 혼합적 분석기법과, 선행 연구된 알고리즘에 대한 정확도 및 수행시간에 대한 비교분석, 최적 알고리즘의 선정, 연관규칙을 위한 웹 마이닝에 관한 연구 등이 필요하다고 보여진다.

6. 감사의 글

이 논문은 2000년도 두뇌한국 21사업에 의하여 지원되었음.

7. 참고문헌

[1] A. Muller, "Fast sequential and parallel algorithms for association rulemining: a

comparison", University of Maryland-College Park CS Technical Report, CS-TR-3515, 76p, Auguest, 1995.

[2] A. Savasere, E. Omiencinsky, and S. Navathe, "An efficient algorithm for mining association rules in large databases", In Proceedings of the 21st VLDB Conference, pp. 432-444, Zurich, Swizerland, 1995.

[3] Aitao Chen, "A Comparison of Regression, Neural Net, and Pattern Revognition," CIKM 98 Bethesda MD USA, pp. 140-147, 1998.

[4] J.S. Park, M.-S. Chen, and P.S. Yu, "An effective hash-based algorithm for mining associatin rules", In Proceedings of ACM SIGMOD Conference on Management of Data, pp. 175-186, San Jose, California, May, 1995.

[5] J.S. Park, P.S. Yu, and M.-S. Chen, "Mining Association Rules with Adjustable Accuracy", In Proceeding s of ACM CIKM 97, pp. 151-160, Las Vegas, Nevada, November, 1997.

[6] Jong Soo Park, Philip S. Yu, Ming-Syan Chen, "Mining Association Rules with Adjustable Accuracy," CIKM 97 LasVegas Nevada USA, pp151-160, 1997.

[7] R. Agrawal and R. Srikant, "Mining sequential patterns", In Proceedings of the 11th International Conferenct on Data Engineering, Taipei, Taiwan, March, 1995.

[8] R. Agrawal, and R.Srikant, "Fast algorithms for mining association rules". In Proceeding s of the 20th VLDB Congerence, Santiago, Chille, Sept., 1994.

[9] R. Agrawal, T. Imielinski, and A. Ssami, "Mining association rules in large databases", In Proceeding of ACM SIGMOD Conference on Management of Data, Washington D.C., pp. 207-216, May, 1933.

[10] R. Agrawal, T. Imielinski, and A. Swami, "Database minig: a performance perspective", IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 6, pp. 914-925, Dec., 1993.

[11] R. Srikant and R. Agrawal, "Mining Generalized Association Rules", In Proceedings of the 21st VLDB Congerence, Zurich, Swizerland, 1995.

[12] R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvements", In Proc. of the Fifth Int'l Congerence on Extending Database Technology(EDBT), Avignon, France, March, 1996.

[13] 용환승, 데이터마이닝, 그린출판사, 1998.