데이터웨어하우스 환경에서의 설명기반 데이터마이닝※

김현수* · 이창호*

Explanation-Based Data Mining in Data Warehouse

Hyun-Soo Kim* · Chang-Ho Lee*

요 약

산업계 전반에 걸친 오랜 정보시스템 운용의 결과로 대용량의 데이터들이 축적되고 있다. 이러한 데이터로부터 유용한 지식을 추출하기 위해 여러 가지 데이터마이닝 기법들이 연구되어왔다. 특히 데이터웨어하우스의 등장은 이러한 데이터마이닝에 있어 필요한 데이터 제공 환경을 주고 있다. 그러나 전문가의 적절한 판단과 해석을 거치지 않은 데이터마이닝의 결과는 당연한 사실이거나, 사실과 다른 가짜이거나 또는 관련성이 없는(Trivial, Spurious and Irrelevant) 내용만 무수히 쏟아낼 수 있다. 그러므로 데이터마이닝의 결과가 비록 통계적 유의성을 가진다 하더라고 그 정당성과 유용성에 대한 검증과정과 방법론의 정립이 필요하다. 데이터마이닝의 가장 어려운 점은 귀납적 오류를 없애기 위해 사람이 직접 그 결과를 해석하고 판단하며 아울러 새로운 탐색 방향을 제시해야 한다는 것이다.

본 논문의 목적은 이러한 데이터마이닝에서 추출된 결과를 검증하고 아울러 새로운 지식 탐색 방향을 제시하는 방법론을 정립하는데 있다. 본 논문에서는 데이터마이닝 기법 중 연관규칙탐사(Associations)로 얻어진 결과를 설명가능성 여부의 판단을 통해 검증하는 기법을 제안하였고, 이를 위해 도메인 지식(Domain Knowledge)과 연관규칙탐사를 통해 얻어진 결과를 표현하기위한 지식표현방법으로 관계형 술어논리(RPL : Relational Predicate Logic)를 개발하였다.

연관규칙탐사로 얻어진 결과를 설명하기 위한 방법으로는 연관규칙탐사로 얻어진 연관규칙에 대해 RPL로 표현된 도메인 지식으로서 설명됨을 보이게 한다. 또한 이러한 설명(Explanation)을 토대로 검증된 지식을 일반화하여 새로운 가설을 연역적으로 생성하고 이를 연관규칙탐사를 통해 검증한 후 새로운 지식을 얻는 설명기반 데이터마이닝 구조(Explanation-based Data Mining Architecture)를 제시하였다.

Key words: 데이터마이닝, 데이터웨어하우스

※ 이 논문은 한국과학재단의 지원으로 연구되었음

* 동아대학교 경영대학 경영정보학과

1. 서 론

데이터마이닝을 통한 지식발견은 의사결정의 주요한 수단으로 부각되고 있다. 그 기본적인 방법론은 이미 자동학습(Machine Learning), 통계학 등의 분야에서 개발되어 왔으나, 데이터마이닝이란 용어는 특정한 방법론에 기초하기 보다 실제 비즈니스 상황에서 경쟁적 우위전략의 획득에 보다 초점을 맞추고 있다. 데이터마이닝의 전형적인 성공은 주로 마케팅에서 발생하였다. 예를 들어, 고객들에게 광고전단을 발송할 때 어떠한 특성을 가진 고객들이 반응률이 높은지를 파악함으로써 그러한 특성을 가진 고객위주로 발송을 하여 전체 고객에게 발송한 경우의 반응률을 얻으면서도 절반 정도로 발송비를 줄일 수 있다면 그 영업적 이득은 쉽게 계산해 낼 수 있다. 데이터마이닝이란 용어는 결국 이러한 비즈니스적 성공과 경영자들이 납득할 수 있는 성과에 기초하여 태동한 것이라 하겠다. 이러한 이유로 Gartner Group에 의하면 포춘 1000대 기업 중 45%가 2000년까지 자동화된 데이터마이닝 도구를 사용하겠다고 조사된 바 있다(Stedman, 1997).

데이터마이닝은 요약(Summarization), 분류(Classification), 클러스터링(Clustering), 연관규칙탐사(Association), 경향분석(Trend Analysis)등을 통해 데이터로부터 조직의 목표달성 및 성과향상과 직결되는 흥미 있는(Interesting) 패턴을 찾아내는 과정이라고 볼 수 있다. 여기서 흥미 있다는 것은 현재까지 발견되지 못했으며(Previously Unknown), 평범하거나 당연하지 않으며(Nontrivial), 동시에 잠재적으로 매우 유용하다는(Potentially Useful) 의미로 해석할 수 있다(Frawley., Piatetsky-Shapiro and Matheus, 1991).

데이터마이닝과 관련한 기존의 연구는 데이터마이닝의 전체적인 과정(Selection, Preprocessing, Transformation, Data Mining, Interpretation/Evaluation)에서 볼 때, 상대적으로 마이닝 기법의 방법론적 연구에 집중되어왔다. 데이터마이닝의 해석(Interpretation)과 검증(Evaluation)과 관련하여서는 마이닝 결과를 측정하는 정량적인 지표[지지도(Supportiveness), 신뢰도(Confidence), 설득력(Conviction), 흥미도(Interestingness), 경이도(Surprisingness), 상관도(Correlation), 유사도(Similarity)] 개발에 치중되어 연구되어왔다(윤종필 외 2인, 1998).

데이터마이닝에 있어서 마이닝 과정을 거쳐 정량적으로 유의성이 확인된 자료라 할 지라도, 과연 잘못된 것이 아닌지를 판단해야 하는 과정이 있어야 한다. 사람의 해석과 판단 과정은 데이터마이닝의 중요한 단계이기도 하다(Piatetsky-Shapiro, Fayyad and Smith, 1996). 그러나, 사람의 해석과 판단을 지원하는 도구에 대한 연구는 마이닝 기법자체의 연구에 비해 취약했던 게 사실이다.

이제, 우리는 다음과 같은 의문점을 제기하고자 한다. 첫째, 데이터마이닝으로부터 발견되어진 지식이 과연 잘못된 것이 아님을 어떻게 알 수 있겠는가? 우리는 데이터로부터 아무런 사전 지식이 없이 생성된 결과가 귀납적 오류를 범할 가능성이 있음을 안다. 데이터 추출이나 전처리 과정(Preprocessing)의 잘못이나 잘못된 방법론의 적용은 자칫 데이터마이닝이 당연한 사실이거나, 사실과 다른 가짜이거나 또는 관련성이 없는(Trivial , Spurious and Irrelevant) 내용만 무수히 쏟아낼 수 있다. 데이터마이닝의 용어에서도 나타나 있듯이 데이터의 산으로부터 정말 유용하고 값진 황금과 같은 지식을 추출해 내는 것

은 쉽지 않은 것이다. 비록 그 결과가 통계적으로, 또는 엔트로피(Entropy)와 같은 기계학습에서 요구되는 선정기준을 통과하여 유의하다고 하더라도 유의성 검증만으로는 그 결과가 맞다는 보증을 할 수가 없다. 어떤 데이터마이닝의 실행결과 S&P 500 인덱스가 방글라데쉬(Bangladesh)의 버터생산량과 역사적으로 밀접한 상관관계를 가짐을 발견해 내었다(Stedman, 1997). 실제로 전혀 관계 없는 변량들이 통계적 유의성을 통과할 수 있다는 것은 여러 경험으로 알고 있다.

두 번째 제기하는 의문은 데이터마이닝에 있어서 사람이 가지고 있는 지식을 어떻게 하면 효과적으로 데이터마이닝 과정에 결합시킬 수 있겠느냐 하는 것이다. 자동학습이든 통계적 기법이든 데이터마이닝의 실제 구현은 사전적으로 사람의 지식을 반영해야 한다. 예를 들어, 자동학습에 있어서 특성(Attribute)이나 통계적 방법론에 있어서의 변수(Variable) 선정까지 데이터마이닝 도구가 자동으로 해 주지는 않는다. 또한 그 결과에 대한 해석(Interpretation)과 평가(Evaluation) 및 채택여부의 결정은 반드시 사람이 해야 한다. 따라서 해당 분야에 전문적인 지식과 경험이 있는 사람의 참여는 데이터마이닝 프로젝트에서 필수적이다.

본 연구는 이러한 두 가지 문제점을 해결하기 위한 목적으로 설명기반 데이터마이닝(Explanation-Based Data Mining)을 제시하고 있다. 또한, 데이터마이닝을 효과적으로 수행하기 위한 환경으로 데이터웨어하우스를 전제 조건으로 가정한다. 데이터웨어하우스는 데이터마이닝에 필요한 통합적 데이터의 수집과 전처리 과정을 수행하기 때문에 데이터웨어하우스 기반에서 데이터마이닝을 수행하는 것이 바람직하다. 왜냐하면 대부분의 데이터마이닝의 시간이 이러한 데이터의 준비과정에서 소모되기 때문이다. 본 연구는 이러한 데이터웨어하우스 기반 위에 데이터로부터의 귀납적 지식생성 방법론인 데이터마이닝과 사람의 사전지식으로부터의 연역적 지식생성방법과의 통합 기법을 개발하는 것을 궁극적 목적으로 한다.

본 연구의 구성은 2장에서 본 연구의 대상이 되는 소매점 분야에서 데이터웨어하우스의 스타조인 스키마(Star Join Schema) 구조를 설명하며, 3장에서는 소매점 분야를 대상으로 많이 쓰이는 데이터마이닝의 한 기법인 연관규칙탐사를 소개하며, 4장에서는 연관규칙을 표현하기 위한 관계형 술어논리(Relational Predicate Logic)를 제안한다. 5장에서는 설명기반 데이터마이닝의 방법론을 설명하고 귀납적 지식생성방법과 연역적 지식생성 방법의 통합 방법론을 제시한다. 마지막으로 6장에서는 결론과 향후연구방향으로 논문을 맺도록 하겠다.

2. 소매점 분야에서의 데이터웨어하우스 구조

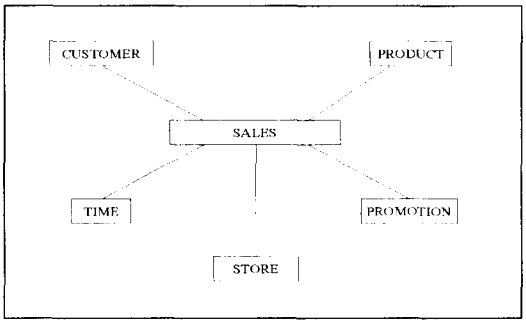
오류가 있는 데이터로는 데이터마이닝의 결과가 부정확할 수 밖에 없기 때문에 데이터마이닝의 적용대상이 되는 데이터가 오류 없이 정제되고 표준화된, 일관성 있는 체계적인 구조로 준비되어 있어야 한다. 데이터웨어하우스는 이러한 데이터마이닝의 요구조건을 충족시키고 효율적인 데이터마이닝을 위한 환경을 제공한다. 본 연구를 위한 데이터웨어하우스의 대상영역으로는 일반적으로 시장바구니 분석(Market Basket Analysis)에서 많이 다루고 있는 소매 시장(Retail Market)을 대상으로 하였다. 특히 슈퍼마켓 형태의 소매점을 체인형태로 각 지역에 두고 있는

대규모 소매점을 대상으로 한다.

데이터웨어하우스는 OLAP(Online Analytical Process)등을 위한 질의 중심의 시스템이기 때문에 기존의 정규형(Normal Form)에 근간을 두어서는 데이터웨어하우스 설계에 부적합하다. 다차원 분석을 위해서 데이터웨어하우스는 차원모형에 바탕을 두고 있다. 차원모형은 사용자의 관점에서 핵심 데이터를 사실 테이블(Fact Table)에 저장하고, 사실 테이블을 보조하는 데이터는 차원 테이블(Dimension Table)에 저장하는 형태로써, 차원은 사용자가 데이터를 분석할 때의 주요분석요인을 의미한다. 차원모형의 스키마 표현법을 스타조인 스키마(Star Join Schema)라 부른다(Red Brick System, 1996)

본 논문에서 다루게 되는 소매상 도메인(Retail Domain)에서 사실 테이블, 차원구축은 다음과 같이 구성하였다.

- 사실테이블: SALES
- 차원테이블: CUSTOMER, PRODUCT, TIME, STORE, PROMOTION



〔그림 1〕 소매시장분야에서의 스타조인 스키마(Star Join Schema)

[그림 1]은 소매시장 분야를 위한 데이터웨어하우스의 스타조인 스키마의 기본 구조를 나타내고 있다. 여기서, CUSTOMER 차원은 고객의 특성에 대해서, TIME차원은 거래가 발생한 시점에 대해서, STORE차원은 거래가 성립한 지점과 관련된 자료에 대해서, PRODUCT는 제품의 특성, PROMOTION은 구매촉진 캠페인 및 광고 등과 관련된 차원을 나타낸다. 각 차원테이블 및 사실테이블은 각기 관계형 데이터베이스의 릴레이션(Relation)으로 구현될 수 있다. 세부 필드의 구성은 [그림 2]에 보이고 있다.

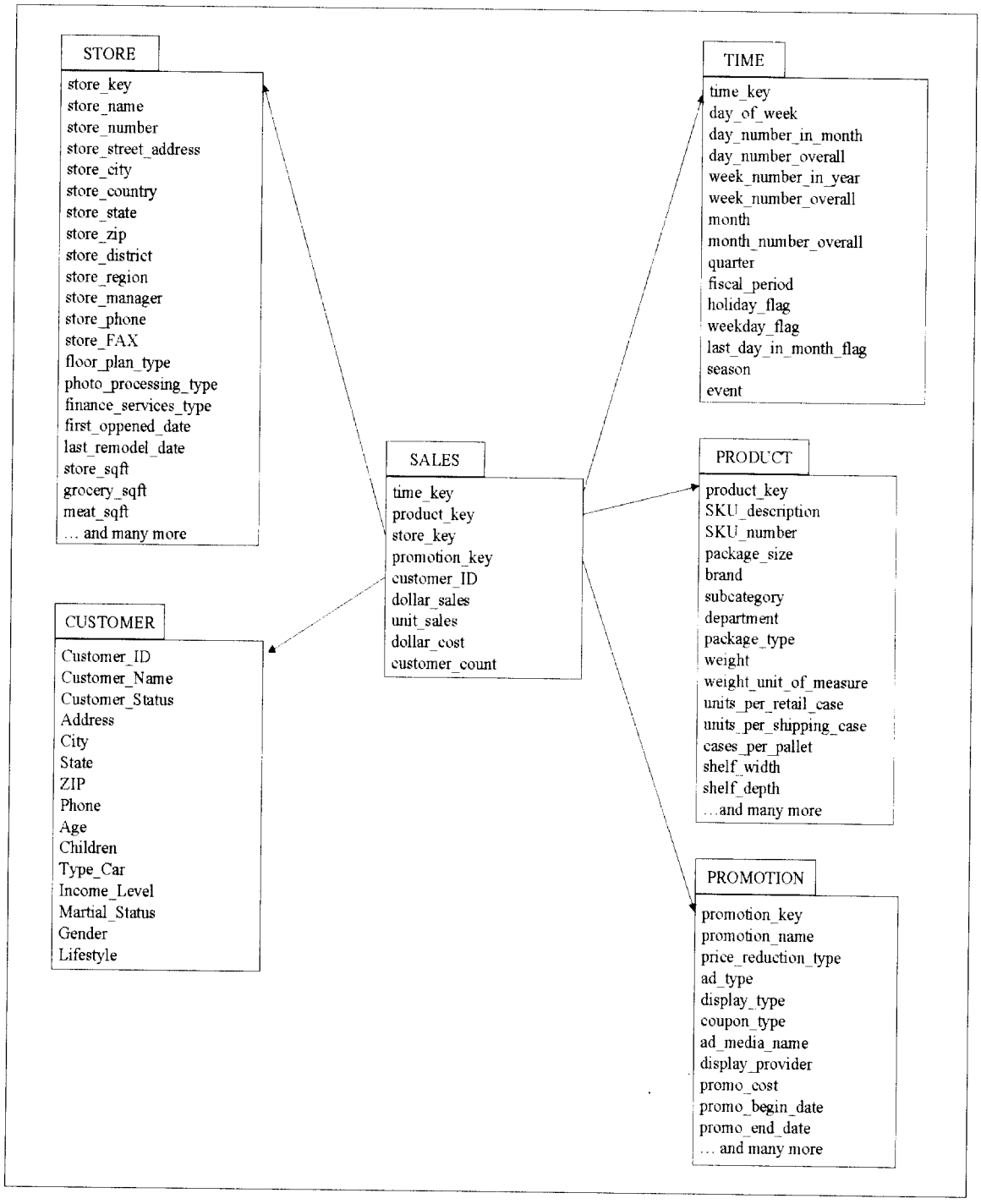
3. 연관규칙탐사(Associations)

데이터마이닝을 위한 구체적인 방법론으로는 여러 가지가 있지만, 본 논문에서는 시장바구니 분석에서 많이 쓰이는 연관규칙탐사를 대상으로 하였다.

연관규칙(Association Rule)은 항목들의 집합으로 표현된 트랜잭션(Transaction)들에서 동시 발생하는 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙이다(Agrawal, Imielinski and Swami, 1993). 이 규칙은 X → Y의 형태를 갖는데 여기서 X, Y는 항목들의 집합이다. 앞의 데이터웨어하우스의 차원모형에서 각 항목은 차원 테이블의 필드에 해당한다. 연관규칙 X → Y 는 "X를 포함하는 트랜잭션들이 Y를 포함하는 경향이 있다"고 해석한다.

전통적인 연관규칙 예제 중 하나인 "기저귀 → 맥주[Support:10%, Confidence: 80%]"는 "전체 트랜잭션의 10%는 기저귀를 포함하고, 기저귀를 구입하는 사람 중 80%는 맥주를 구입한다"라고 해석된다.

위에서 사용된 지지도(support: S)와 신뢰도(confidence: C)가 연관규칙의 유의도를 나타내는



〔그림 2〕 소매시장에서의 스타조인 스키마

수치로 사용된다. 즉, 지지도는 규칙이 갖는 통계적 유의도를 나타내며, 신뢰도는 규칙 자체의 강도를 의미한다고 볼 수 있다. 사용자가 미리 정의한 최소 지지도(Minimum Transactional Support: S_{min}), 최소 신뢰도(Minimum Confidence: C_{min})에 대해서 $S \geq S_{min}$, $C \geq C_{min}$하면 연관규칙 $X \rightarrow Y$ 은 전체 트랜잭션 집합에 대하여 성립한다.

연관규칙탐사는 주어진 데이터베이스에서 최소지지도와 최소 신뢰도를 초과하는 모든 연관규칙을 찾는 것이다. 연관규칙을 탐색하는 기법은 여러 가지가 소개되었으나 가장 전형적인 방법론은 Apriori(Agrawal and Srikant, 1994) 알고리즘이다. 이 방법론에서 탐색절차는 기본적으로 다음의 두 단계를 거치게 된다.

- 단계1: 최소지지도를 만족하는 모든 항목들의 집합을 찾는다. 이러한 항목들의 집합을 빈발항목집합(Large Itemsets)이라 하고, 그 외 모든 항목집합들은 작은 항목집합들(Small Itemsets)이라 한다
- 단계2: 빈발항목집합을 사용하여 최소신뢰도를 만족하는 규칙(Desired Rules)을 찾는다.

단계 1에서 빈발항목집합이 발견되면 단계 2는 쉽게 유도될 수 있다. 빈발항목집합은 2개 이상의 항목끼리의 항목집합의 지지도를 계산하는 과정을 통해서 유도할 수 있다. Apriori 알고리즘에서는 먼저 항목 하나로 이루어진 빈발항목집합을 데이터베이스에서 찾고, 이를 Apriori-gen 전략을 사용하여 두개의 항목으로 이루어진 새로운 후보항목집합을 만든다. 그리고 다시 데이터베이스를 스캔하여 후보항목집합 중 최소지지도를 만족하는 것을 찾고 계속 Apriori-gen 전략을 사용하여 세 개의 항목들로 구성된 후보항목집합을 만드는 일련의 반복적인 과정을 거친다. 최종적으로 더 이상의 후보항목집합을 생성할 수 없을 때 까지 계속된다. 자세한 내용은 (Agrawal and Srikant, 1994)를 참조하고, 본 논문에서는 생략한다.

4. 데이터웨어하우스 환경에서의 연관규칙의 표현방법 : RPL(Relational Predicate Logic)의 개발

연관규칙탐사에 의해서 얻어진 규칙들을 데이터웨어하우스의 차원 테이블 및 사실 테이블과 연결시키고 아울러 사람이 가지고 있는 사전지식이나 판단지식을 함께 표현하기 위해서는 기존의 단순한 연관규칙 표현법 보다는 새로운 지식표현 방법이 필요하다. 본 논문에서는 연관규칙을 표현하기 위해 일차술어논리(FOPL: First Order Predicate Logic)를 연관규칙 표현에 맞게 변경한 관계형 술어논리(RPL : Relational Predicate Logic)를 개발하였다.

먼저, 일차 술어논리를 설명하면, 일차 술어논리는 명제 논리(Propositional Logic)의 표현력 한계를 극복하고자 논리학자들에 의해 개발되어 인공지능 분야, 전문가시스템 분야에서 많이 사용되고 있다(Patterson, 1990). 이것은 명제 논리를 보다 일반화시킨 것이다.

일차 술어논리의 구문은 다음과 같다.

- 연결자(connectives): ~(not), ∧(and), ∨(or), →(implication), ↔(equivalence)
- 정량자(quantifiers): ∀(universal qualification), ∃(existential qualification)
- 상수(constants): 주어진 정의역에서 고정값을 갖는 항.

- 변수(variables): 주어진 정의역에서 다른 값을 가정할 수 있는 항.
- 함수(functions): 정의역에서 정의된 관계.
- 술어(predicates): 정의역으로부터 참 또는 거짓으로 사상되는 관계 또는 함수를 나타내는 기호.
  (n개($n \geq 0$)의 항(terms)을 가질 수 있다. P(t1, t2, t3, ..., tn)로 표기된다.)

일차 술어논리의 표현 예를 위해, 다음과 같은 문장(Statement)이 있다고 하자(Patterson, 1990).

- E1: All employees earning $1,400 or more per year pay taxes.
- E2: Some employees are sick today.
- E3: No employee earns more than the president.

먼저 이를 표현할 술어와 함수를 정의하면, 다음과 같다.

- E(x): x is an employee.
- P(x): x is president.
- i(x): the income of x(function).
- GE(u,v): u is greater than or equal to v.
- S(x): x is sick today.
- T(x): x pays taxes.

정의된 술어와 함수를 가지고 문장을 일차술어논리식으로 표현하면, 아래와 같이 표현될 수 있다.

- E1': $\forall x((E(x) \wedge GE(i(x), 1400)) \rightarrow T(x))$
- E2': $\exists y(E(y) \rightarrow S(y))$
- E3': $\forall xy((E(x) \wedge P(y)) \rightarrow \sim GE(i(x), i(y)))$

본 논문에서는 연관규칙과 이를 설명할 사람의 지식을 동일한 구조로 표현하고, 이를 이용하여 설명과정을 표현하고자 관계형 술어논리(RPL: Relational Predicate Logic)를 개발하였다. 관계형 술어논리는 기존의 일차 술어논리가 데이터베이스의 릴레이션을 표현하지 못하는 점을 보완하여 일차술어논리의 구문을 확장한 형태로 개발하였다. 주요 변경사항은 다음과 같다.

- 릴레이션의 표현: 변수로써 데이터베이스의 릴레이션을 사용할 수 있다. 대문자로 표현한다.
  Ex) 소매점 데이터웨어하우스 스키마의 Customer(고객) 릴레이션의 표현: CUSTOMER
- 필드의 표현: 릴레이션의 특정 필드를 지칭할 때 '.' 기호 다음 해당 필드를 나타낼 수 있다. 소문자로 표현한다.
  Ex) Customer(고객) 릴레이션의 Age 필드의 표현: CUSTOMER.age
- 정량자 제거: 전체정량자는 모든 트랜잭션에 대해 연관규칙을 만족한다는 의미로 해석되는데 실제로 데이터베이스에서 그렇게 모든 트랜잭션에 대해 연관성을 갖는 규칙은 거의 없기 때문에 큰 의미가 없다. 마찬가지로 존재정량자의 경우 해당 연관성을 나타내는 트랜잭션이 하나라도 있을 때 성립하나 연관규칙은 어느 정도의 지지도를 가져야 하므로 단지 그러한 연관성을 갖는 트랜잭션이 존재한다고 하는 것은 큰 의미가 없다. 대신 연관규칙으로 성립하기 위한 최소한 지지도와 신뢰도 이상을 갖는다는

의미의 정량자가 필요하다. 그러나 연관규칙으로 나타내어지는 모든 규칙은 이러한 정량조건을 만족하는 규칙만을 다룬다고 보면 이 정량자는 의미가 없다.

그 외, 연결자, 상수, 함수, 술어등은 기존의 일차 술어논리와 동등한 방법으로 사용할 수 있다. 계속해서 실제 연관규칙을 RPL로 표현하면서 설명하겠다. 다음은 슈퍼마켓 체인점들의 판매 데이터웨어하우스를 대상으로 파악하는 연관규칙들의 예이다.

- F1: 하단동 지점에서는 생필품이 잘 팔린다.
- F2: 고객이 빵을 구입하면 우유도 구입한다.
- F3: 저녁시간대에 기저귀를 사면 맥주도 구입한다.

위의 문장을 기존의 연관규칙형태로 표현하면 다음과 같다.

- F1′: 하단동 지점 → 생필품
- F2′: 빵 → 우유
- F3′: 저녁시간대 ∧ 기저귀 → 맥주

위의 연관규칙표현은 직관적으로 이해할 수는 있으나 데이터웨어하우스의 구조와는 별개의 표현법을 쓰고 있다.

만약 위의 연관규칙이 추출된 데이터가 저장된 데이터웨어하우스가 [그림 2]의 구조를 가지고 있을 때 위의 연관규칙을 RPL로 표현하면 다음과 같다.

- F1′: Eq(STORE.street_name, '하단동') → Eq(PRODUCT.subcategory, '생필품')
- F2′: Eq(PRODUCT.subcategory, '빵') → Eq (PRODUCT.subcategory, '우유')
- F3′: Eq(PRODUCT.subcategory, '기저귀') ∧ Between(18, TIME.hour, 22) → Eq(PRODUCT.subcategory, '맥주')

여기서 PRODUCT, STORE는 데이터웨어하우스의 차원테이블(릴레이션)이며, 그 중 street_name, subcategory는 필드를 나타낸다. 즉, 해당 차원테이블의 특정 필드의 값으로 구성된 튜플의 집합을 나타낸다. Eq는 두개의 항이 같다는 술어이며, Between은 3개의 항을 가지는 술어이다. 여기서 → 의 의미는 연관성을 나타내는 "if-then"문장의 의미라고 본다.

RPL의 장점은 SQL이 관계형 데이터베이스를 조작, 정의할 수 있는 언어이듯이, RPL은 관계형 데이터웨어하우스와 연결된 데이터마이닝 엔진으로부터 도출된 연관규칙을 표현할 수 있고, 사람의 논리적 지식까지도 나타낸다. 아울러 사람들이 연역적으로 도출한 연관규칙을 데이터웨어하우스에서 검증하도록 해당 차원 테이블을 표현할 수 있다.

5. 설명기반 데이터마이닝

5.1 설명을 통한 연관규칙의 검증

연관규칙탐사에 있어서의 중요한 점은 첫째, 연관규칙탐사과정을 거쳐 나온 연관규칙들이 과연 잘못된 것이 아닌지를 판단해야 한다는 것이다. 그러나 통계적으로 이미 지지도와 신뢰도를 가지고 데이터웨어하우스로부터 생성된 것이기 때문에 그 규칙의 옳고 그름은 결국 사람이 판단

하게 된다. 사람의 해석과 판단 과정은 데이터마이닝의 중요한 단계이기도 하다(Piatetsky-Shapiro, Fayyad and Smith, 1996).

본 논문에서는 연관규칙탐사를 통해 데이터베이스로부터 귀납적으로 형성된 연관규칙 A → B의 검증은 사람이 해당 규칙의 성립에 대한 이유를 설명할 수 있을 때 검증되었다고 본다.

설명기반 자동학습(Explanation-Based Learning)에 있어서 사전지식을 통한 설명이 새로운 지식을 생성하고 있다(DeJong, 1997). 비슷한 원리로 사람이 가지고 있는 도메인에 대한 지식으로써 데이터베이스로부터 생성된 연관규칙을 설명함으로써 해당 연관규칙을 새로운 지식으로 확증하며 이러한 설명이 있기 전까지 해당 규칙은 가설(Hypothesis)로서의 지위만을 가지게 된다. 실제로 Piatetsky-Shapiro등이 제시한 지식발견과정에서도 데이터마이닝의 다음과정에서 사람의 해석과 평가를 거친 후 지식으로 형성되고 있음을 알 수 있다.

앞 절에서 예시된 F1′ ~ F3′의 연관규칙을 살펴보자. 왜 하단동지점에는 생필품이 잘 팔리는 것인가? 왜 저녁무렵 기저귀를 사는 사람이 동시에 맥주도 사는 것인가? 만약 여기에 대한 설명을 할 수 없다면 해당 연관규칙은 잘못 생성된 것이라고 본다. 마치 S&P 500 인덱스와 방글라데쉬의 버터 생산량이 통계적으로 유의한 상관관계를 가지고 있지만 그 이유에 대한 설명을 할 수 없으므로 해당 규칙을 폐기시키는 것과 같다. 또한 경영자 입장에서도 데이터마이닝부서로부터 보고된 연관규칙이 경영자가 도저히 납득할 수 없는 결과라면 해당 규칙을 구현시킨다는 것을 기대하기란 어렵다.

이제 앞의 연관규칙에 대한 설명을 앞에서 정의한 RPL을 통해 표현한다. 설명을 하기 위해서는 사람이 가지고 있는 사전적 지식을 나타내야 하는데, 술어논리는 사람의 지식을 표현하는데 현재까지 매우 유용하고 널리 알려진 방법이다.

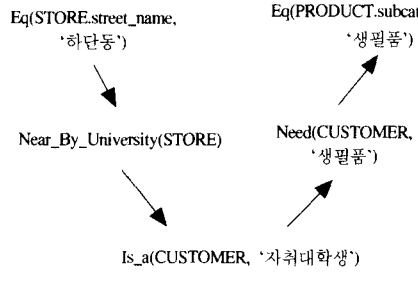
그렇다면, 설명의 개념과 설명의 방법론에 대해 알아본다. 설명이란 해당 연관규칙이 성립하는 이유를 밝히는 것이다. 설명하는 방법으로써 크게 두 가지를 제시한다.

첫째, 두개의 항목이 서로 연관성을 가지는 것은 동시에 두개의 항목에 영향을 주는 공통영향인자(Common Affecting Factors)집합이 있어서이기 때문이다. 따라서 연관규칙에 포함된 항목들의 공통영향인자를 찾아내는 것이 설명하는 방법이 된다. 예를 들어 A항목과 B항목이 직접적인 관계는 없지만 빈발하게 연관성을 가지는 있는 이유는 사실, C라는 공통영향인자가 A와 B에 동시에 영향을 주기 때문이다. 예를 들어, 기저귀와 맥주의 소비는 이러한 공통영향인자의 발견으로 설명되어 질 수 있다.

둘째, 연관되는 두개의 항목간에 매개인자집합을 찾아내는 것이다. 이 매개인자가 항목간의 연관성을 중개한다. 즉, A와 B간의 연관성은 A로 말미암아 C가 영향을 받고 다시 C가 B의 발생에 영향을 주기 때문에 생기는 것이다.

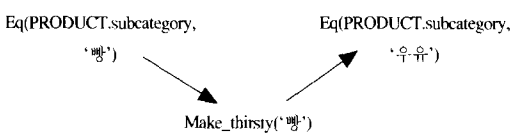
이러한 설명원리를 가지고 앞의 연관규칙은 다음과 같이 설명된다.

· F1′′의 설명:



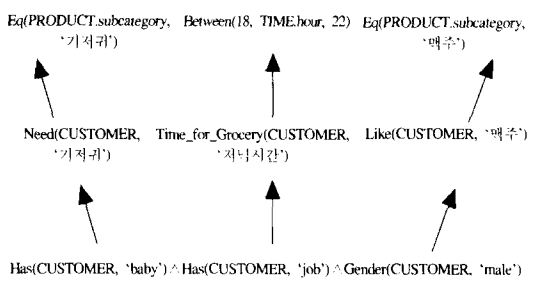
하단동 지점이 생필품이 많이 팔리는 이유는 하단동에 대학이 위치해 있고, 대학근처에는 자취하는 대학생들이 많이 살며 따라서 이들이 생필품을 필요로 하기 때문이다.

· F2''의 설명:



빵을 사는 사람이 우유를 사는 것은 빵이 목을 마르게 하기 때문이다. 때로, 하나의 연관성에 대해 여러가지 중첩된 설명이 가능하다. 위의 경우, 우유를 사는 또 다른 설명을 첨가할 수 있다.

· F3''의 설명:



아기를 가지고, 직장을 가진, 남성이라는 공통 영향요인 집합이 기저귀와 저녁시간대, 그리고 맥주라는 세가지 항목간에 연관성을 야기하고 있다.

설명 F1''과 F2''는 두번째 설명전략으로 설명한 것이며, F3''는 첫번째 설명전략으로 설명하였다. 설명에 쓰인 지식은 단지 귀납적으로 데이터마이닝을 통해 생성된 연관규칙의 검증이라는 기능만을 하는 것은 아니고 5.2절에서 설명할 새로운 가설의 생성을 유도하는데 이용된다.

5.2 설명으로부터 가설의 생성과 가설의 검증

데이터마이닝을 위해서 아무런 사전 지식이 없이 오직 데이터로부터 연관규칙탐색을 시작할 수 있지만 만약, 사전 지식이 있다면 그 사전 지식으로부터 가설을 세우고 이 가설을 검증하기 위해 데이터베이스에서 해당 가설과 결부된 항목간의 연관성을 검색할 것이다. 이 때 사전 지식으로 부터의 가설의 설정과정은 사람의 연역적 추론과정의 일부라고 할 수 있다.

이제 앞 절에서 연관규칙에 대한 설명으로부터 새로운 연관규칙 가설을 성립할 수 있다. 이러한 가설은 다시 데이터웨어하우스에서 지지도와 신뢰도 검증, 즉 가설검증과정을 거쳐 지식으로 확증하게 된다.

설명으로부터 새로운 가설의 생성은 설명에서 쓰인 상수를 치환(Substitution)하는 것이다.

예를 들어 F1''에서 쓰인 Eq(STORE.street_name, '하단동')에서 '하단동'을 '대신동'으로 대체 시킬 수 있다. 왜냐하면 '대신동'으로 대체해도

· Eq(STORE.street_name, '대신동') → Near_By_University(STORE)

라는 설명은 계속 성립하기 때문이다(여기서 하단동이나 대신동이나 모두 대학가에 위치한 곳으로 가정한다). 이렇게 기존의 설명구조를 계속 유지시키되 설명에 쓰인 항목을 다른 내용으로 치환함으로써 다음과 같은 새로운 가설이 형성된다.

· G1 : Eq(STORE.street_name, '대신동') → Eq(PRODUCT.subcategory, '생필품')

· G2 : Eq(PRODUCT.subcategory, '빵') →
Eq(PRODUCT.subcategory, '쥬스')

· G3 : Eq(PRODUCT.subcategory, '아기쥬스')
∧ Between(18, TIME.hour, 22) →
Eq(PRODUCT.subcategory, '맥주')

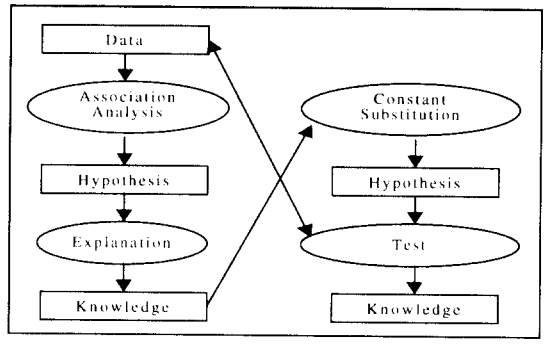
G1 가설은 대신동도 대학주변의 지역이기 때문이고, G2는 빵이 목을 마르게 하므로 갈증해소라는 이유로 우유를 선택한다면 쥬스도 빵과 연관성에 있어서 대안이 될 수 있기 때문이다. 또한 아기를 가졌기 때문에 기저귀를 찾았다면 아기쥬스도 아기를 가진 사람이 필요로 할 것이고 따라서 아기쥬스와도 맥주가 연관성이 있을 수 있기 때문에 G3가설이 만들어 질 수 있다.

이러한 가설이 정확한 지식인지는 데이터웨어하우스에서 해당 연관규칙의 지지도와 신뢰도를 계산함으로써 검증되게 된다. 이것은 가설검증(Hypothesis Test)의 기능이 된다.

5.3 지식생성에 있어서 귀납적 방법론과 연역적 방법론의 통합 모형

5.2절의 설명기반 데이터마이닝은 아무런 사전 지식없이 주어진 데이터로부터 귀납적으로 가설을 생성하는 방법과 사전지식에 바탕을 두고 연역적으로 새로운 가설을 생성하는 방법론간의 통합모형이 될 수 있다. 이를 그림을 나타내면 [그림 3]과 같다.

[그림 3]의 좌측축은 데이터로부터 귀납적으로 추출된 가설을 설명으로 검증함으로써 최종 지식으로 확증되는 과정을 나타내고 있고, 우측축은 주어진 확증된 지식의 내용을 설명하는 항목내용의 치환을 통해 새로운 가설을 생성하고 이를 데이터에서 검증함으로써 지식으로 확정하게 된다. 따라서 연역적 기법과 귀납적 기법의 통합화가 설명에 기반을 둔 데이터마이닝 방법론과 사람의 지식과 데이터웨어하우스를 연계할 수 있는 지식표현방법인 RPL을 통해 구현할 수 있다.



[그림 3] 설명기반 데이터마이닝의 전체 흐름도

6. 결론

데이터마이닝이 효과적으로 수행되기 위해서는 데이터마이닝의 알고리즘개발도 필요하지만 추출된 지식을 저장, 조직하고 추론할 수 있게 함으로써 데이터마이닝을 수행하는 사람의 지식수준 또한 향상되지 않으면 데이터마이닝은 성공하기 어렵다고 본다. 본 연구는 단순히 사람을 배제하고 자동화한다는 기계적인 접근방법을 지양하고, 데이터마이닝의 수행과정에서 데이터마이닝의 결과의 해석과 판단여부는 사람이 할 수 밖에 없는 상황에서 지식의 관리와 새로운 가설의 생성과정을 도와줄 수 있는 방법론과 이에 필요한 지식표현기법의 필요성에서 출발하였다. 또한 단순히 데이터에만 의존하여 지식을 추출함으로써 생기는 귀납적 오류를 방지하고, 사회과학분야의 연구방법론에 있어서 가설검증

(Hypothesis Test)과정을 데이터마이닝 관점에서 흡수하고자 하였다.

본 연구에서는 데이터마이닝에 있어서 도출된 지식의 검증 방법론과 이를 위한 지식표현기법을 데이터웨어하우스의 차원 테이블과 연계하여 개발하였다. 아울러 데이터로부터의 생성된 지식과 사람의 설명지식으로부터, 새로운 가설을 도출하여 이를 데이터에서 검증 받게 함으로 귀납적 지식추출 방법론과 연역적 가설생성의 통합모형을 제시하였다.

본 연구의 의의로는 데이터마이닝을 통한 귀납적 지식생성에 있어 귀납적 오류의 발생 여부를 도메인 지식을 통해 설명가능 함을 보임으로써 검증하고 아울러 이러한 설명을 통해 연역적으로 새로운 가설을 생성시켜 이를 가설검증방식으로 검증함으로써 귀납적 접근과 연역적 접근의 통합 데이터마이닝 접근법을 제시하였다는 데 있다.

향후 연구과제는 본 연구의 아이디어와 지식표현방법을 토대로 연관규칙생성 알고리즘과 통합한 설명기반 데이터마이닝 도구의 구현과 이의 실제문제의 적용이라 하겠다.

참 고 문 헌

김정자, 이도헌, 데이터마이닝의 기술 및 연구동향, 정보과학회지 제 16권 제 9호, 1998.9, 6-14.

박종수, 유원경, 홍기형, 연관규칙탐사와 그 응용, 정보과학회지, 제 16권 제 9호, 1998.9, 37-44.

윤종필, 김희숙 최옥주, 데이터마이닝의 유용성, 정보과학회지, 제 16권, 제 9호, 1998.9, 15-23.

Adrianns, Pieter and Dolf Zantinge, *Data Mining*, Addison Wesley Longman, England, 1996

Agrawal, R., T. Imielinski and A. Swami, "Mining association rules between sets of items in large database", *Proceedings of ACM SIGMOD Conference on Management of Data*, Washington D.C., (1993), 207-216.

Agrawal, R. and R. Srikant, "Fast algorithms for mining association rules", *Proceedings of the 20th VLDB Conference*, Santiago, Chile, Sept., 1994.

Berson, Alex, Stephen J. Smith, *Data Warehousing, Data Mining, and Olap*, McGraw-Hill, 1997.

Blum, Robert L., "Discovery, Confirmation, and Incorporation of Causal Relationships from a Large Time-Oriented Clinical Data Base: The RX Project," *Computers and Biomedical Research* 15, (1982), 164-187.

Chen, M. S., J. Han, and P.S. Yu, "Data Mining: An Overview from a Database Perspective", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6 (Dec, 1996), 866-883.

DeJong, G., "Explanation-Based Learning", In Allen B. Jr. Tucker and Allen B. Tucker Jr. The Computer Science and Engineering Handbook, CRC Press, Inc., 1997.

Frawley, W. J., G. Piatetsky-Shapiro and C. J. Matheus, "Knowledge Discovery in Databases: An Overview", In G. Piatetsky-Shapiro and W. J. Frawley(Eds.), *Knowledge Discovery in Databases*, AAAI/MIT Press,1991.

Inmon, W.H., and R.D. Hackathon, *Using the Data Warehouse*, John Wiley and Sons, New York, 1992.

Kimball, R., *Data Warehouse Toolkit*, John Wiley & Sons, 1996.

Lee, Byungtae, Anitesh Barua and Andrew B.

Whinston, "Discovery and Representation of Causal Relationships in MIS Research: A Methodological Framework," *MIS Quarterly*, (Mar. 1997), 109-136.

Orr, Ken, "Data Warehousing: Phase 2", *DCI's Data Warehouse World Conference Proceedings*, (Aug. 1996), C31-1 ~ C31-50.

Patterson, Dan W., *Introduction to Artificial Intelligence and Expert Systems*, Prentice-hall.,1990.

Piatetsky-Shapiro, G., U. Fayyad and P. Smith, "From Data Mining to Knowledge Discovery: An Overview," In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy(Eds.), Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press,1996.

Pearl, Judea, Causal Diagrams for Empirical Research, *Biometrika*, Vol 82, No. 4 (1995), 669-710.

Red Brick System, *Star Schemas and STAR join Technology*, Red Brick Systems, White Paper, 1996.

Stedman, Craig, "Data mining for Fool's Gold", *Computerworld*, (Dec. 1, 1997), p.28.