데이타 웨어하우스 환경에서의 설명기반 데이타 마이닝

김현수 · 이창호*

Explanation-based Data Mining in Data Warehouse

Hyun Soo Kim, Chang Ho Lee*

요약

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

본 논문의 목적은 이러한 데이타 마이닝에서 추출된 결과를 검증하고 아울러 새로운 지식 탐색 방향을 제시하는 방법론을 정립하는데 있다.

본 논문에서는 데이타 마이닝 기법 중 연관규칙탐사로 얻어진 결과를 설명가능성 여부의 판단을 통해 검증하는 기법을 제안하며, 이를 통해 얻어진 검증된 지식을 토대로 일반화를 통한 새로운 가설을 생성하여 데이타 웨어하우스로부터 연관규칙을 검증하는 일련의 아키텍처(architecture)를 제시하고자 한다.

먼저 데이타 마이닝 결과에 대한 설명의 필요성을 제시하고, 데이타 웨어하우스와 데이타 마이닝 기법들에 대한 간략한 설명과 연관규칙탐사에 대한 정의 및 방법을 보이고, 대상 영역에 대한 데이타 웨어하우스의 스키마를 보였다.

다음으로 도메인 지식(domain knowledge)과 연관규칙탐사를 통해 얻어진 결과를 표현하기위한 지식표현 방법으로 Relational Predicate Logic 을 제안하였다.

연관규칙탐사로 얻어진 결과를 설명하기 위한 방법으로는 연관규칙탐사로 얻어진 연관규칙에 대해 Relational Predicate Logic 으로 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 반복적인 Explanation-based Data Mining Architecture 를 제시하였다.

* 동아대학교 경영정보학과

본 연구의 의의로는 데이타 마이닝을 통한 귀납적 지식생성에 있어 귀납적 오류의 발생을 도메인 지식을 통해 설명가능 함을 보임으로 검증하고 아울러 이러한 설명을 통해 연역적으로 새로운 가설지식을 생성시켜 이를 가설검증방식으로 검증함으로써 귀납적 접근과 연역적 접근의 통합 데이타 마이닝 접근을 제시하였다는데 있다.

Key words: 데이타 마이닝, 데이타 웨어하우스

1. 서론

데이타 마이닝을 통한 지식발견은 의사결정의 주요한 수단으로 부각되고 있다. 그 기본적인 방법론은 이미 자동학습(machine learning), 통계학 등의 분야에서 개발되어 왔으나, 데이타 마이닝이란 용어는 특정한 방법론에 기초 하기 보다 실제 비즈니스 상황에서 경쟁적 우위전략의 획득에 초점을 맞추고 있다. 데이타 마이닝의 전형적인 성공은 주로 마케팅에서 발생하였다. 예를 들어, 고객들에게 광고전단을 발송할 때 어떠한 특성을 가진 고객들이 반응률이 높은지를 파악함으로써 절반 정도의 발송비를 줄일 수 있다면 그 영업적 이득은 쉽게 계산해 낼 수 있다. 데이타 마이닝이란 용어는 결국 이러한 비즈니스적 성공과 경영자들이 납득할 수 있는 성과에 기초하여 태동한 것이라 하겠다. 이러한 이유로 Gartner Group 에 의하면 포춘 1000 대 기업 중 45%가 2000 년까지 자동화된 데이타 마이닝 도구를 사용하겠다고 조사된 바 있다(Stedman, 1997).

데이타 마이닝의 목표는 요약(summarization), 분류(classification), 클러스터링(clustering), 연관규칙 탐사(association), 경향분석(trend analysis)등을 통해 데이타로부터 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.

그렇지만 우리는 다음과 같은 의문점을 제기하고자 한다. 데이타 마이닝으로부터 발견되어진 지식이 과연 잘못된 것이 아님을 어떻게 알 수 있겠는가? 우리는 데이타로부터 아무런 사전 지식이 없이 생성된 결과가 귀납적 오류를 범할 가능성이 있음을 안다. 데이타 추출이나 전처리 과정의 잘못이나 잘못된 방법론의 적용은 자칫 데이타 마이닝이 당연한 사실이거나, 사실과 다른 가짜이거나 또는 관련성이 없는(trivial , spurious and irrelevant) 내용만 무수히 쏟아낼 수 있다. 데이타 마이닝의 용어에서도 나타나 있듯이 데이타의 산으로부터 정말 유용하고 값진 황금과 같은 지식을 추출해 내는 것은 쉽지 않은 것이다. 비록 그 결과가 통계적 또는 엔트로피와 같은 기계학습에서 요구되는 선정기준을 통과하여 유의하다고 하더라도 유의성 검증만으로는 그 결과가 맞다는 보증을 할 수가 없다. 어떤 데이타 마이닝의 실행 결과 Standard Poor 500 주식 인덱스가 방글라데시의 버터생산량과 역사적으로 밀접한 상관관계를 가짐을 발견해 내었다. 실제로 전혀 관계 없는 변량들이 통계적 유의성을 통과할 수 있다는 것은 여러 경험으로 알고 있다.

두 번째 제기하는 의문은 데이타 마이닝에 있어서 사람이 가지고 있는 지식을 어떻게 하면 효과적으로 데이타 마이닝 과정에 결합시킬 수 있겠느냐 하는 것이다. 자동학습이든 통계적 기법이든 데이타 마이닝의 실제 구현은 사전적으로 사람의 지식을 반영해야 한다. 예를 들어, 자동학습에 있어서 특성(attribute)이나 통계적 방법론의 변수(variable)의 선정까지 데이타 마이닝 도구가 자동으로 해 주지는 않는다. 또한 그 결과에 대한 해석(interpretation)

과 평가(evaluation)및 채택여부의 결정은 반드시 사람이 해야 한다.

본 연구는 이러한 두 가지 문제점을 제기하면서 이를 해결하기 위해 설명기반 데이타 마이닝(Explanation-based Data Mining)을 제시하고 있다. 또한, 데이타 마이닝을 효과적으로 수행하기 위한 환경으로 데이타 웨어하우스를 전제 조건으로 가정한다. 데이타 웨어하우스는 데이타 마이닝에 필요한 통합적 데이타의 수집과 전처리 과정을 수행하기 때문에 향후 데이타 웨어하우스 기반에서 데이타 마이닝을 수행하는 것이 바람직하다고 볼 수 있다. 본 연구는 이러한 데이타 웨어하우스 기반 위에 데이타 마이닝의 결과를 검증해 주는 수단으로서 사람의 설명을 방법론적으로 통합시켜 데이타 마이닝을 통한 귀납적 지식의 생성 및 가설의 테스트와 사람의 지식을 통한 연역적 가설 설정 및 설명기능을 통합한 방법론을 제시하고, 아울러 이를 위한 지식표현법을 개발하였다.

2. 소매점에서의 데이타 웨어하우스 구조

오류가 있는 데이타로는 데이타 마이닝의 결과가 부정확할 수 밖에 없기 때문에 데이타 마이닝의 적용대상이 되는 데이타가 오류 없이 정제되고 표준화된, 일관성 있는 체계적인 구조로 준비되어 있어야 한다. 데이타 웨어하우스는 이러한 데이타 마이닝의 요구조건을 충족시키고 효율적인 데이타 마이닝을 위한 환경을 제공한다. 본 연구를 위한 데이타 웨어하우스의 대상영역으로는 일반적으로 시장바구니 분석(market basket analysis)에서 많이 다루고 있는 소매 시장(retail market)을 대상으로 하였다.

데이타 웨어하우스는 질의 중심의 시스템(OLAP)이기 때문에 기존의 개체-관계 모형(E-R Model)은 데이타 웨어하우스 설계에 부적합하여 많은 표현방법이 연구되었다. 본 논문에서는 테이블 간의 중요도를 중심으로 스키마(schema)를 구성하는 차원모형(Kimball, 1996)을 사용하여 데이타 웨어하우스를 설계하였다.

차원모형은 사용자의 관점에서 중요한 데이타를 사실 테이블(fact table)에 저장, 사실 테이블을 보조하는 데이타는 차원 테이블(dimension table)에 저장하는 형태로써, 차원은 사용자가 데이타를 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

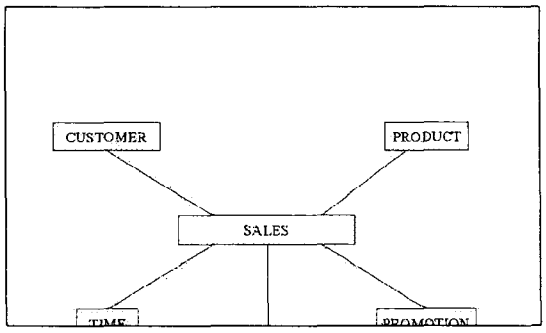

그림 1. 소매시장분야에서의 스타조인 스키마(star join schema)

각 차원테이블 및 사실테이블은 관계형 데이타베이스의 릴레이션(relation)으로 구현될 수 있다. 위의 내용으로 구성된 스타조인 스키마의 대략적인 구조는 그림 1 에 보이고 있고, 세부 필드의 구성은 그림 2 에 보이고 있다.

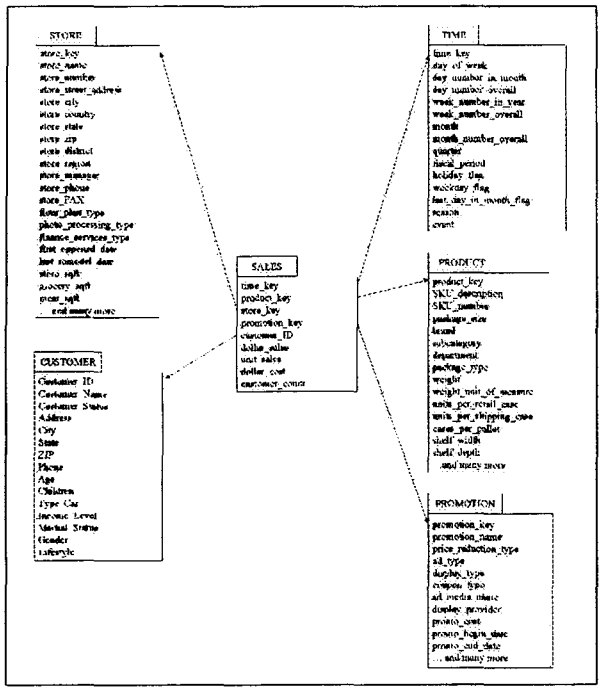

그림 2. 소매시장에서의 스타조인 스키마

3. 연관규칙탐사(Associations)

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

연관 규칙(association rule)은 항목들의 집합으로 표현된 트랜잭션들에서 동시 발생하는 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙이다.(Agrawal, Imielinski, Swami, 1993)

이 규칙은 $X \rightarrow Y$의 형태를 갖는데 여기서 X, Y는 항목들의 집합이다. 연관규칙 $X \rightarrow Y$ 는 "X를 포함하는 트랜잭션들이 Y를 포함하는 경향이 있다"고 해석한다.

전통적인 연관규칙 예제 중 하나인 "기저귀 $\rightarrow$ 맥주[support:10%, confidence: 80%]"는 "전체 트랜잭션의 10%는 기저귀를 포함하고, 기저귀를 구입하는 사람 중 80%는 맥주를 구입한다"라고 해석된다.

위에서 사용된 지지율(support: S)과 신뢰도(confidence: C)가 연관규칙의 유의도를 나타내는 수치로 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. 즉, 지지도는 규칙이 갖는 통계적 유의도를 나타내며, 신뢰도는 규칙 자체의 강도를 의미한다고 볼 수 있다.

연관규칙탐사는 주어진 데이타베이스에서 최소지지도와 최소 신뢰도를 초과하는 모든 연관규칙을 찾는 것이다. 연관규칙을 탐색하는 기법은 여러 가지가 소개되었으나 가장 전형적인 방법론은 Apriori(Agrawal, Srikant, 1994) 알고리즘이다.

탐색절차는 기본적으로 다음의 두 단계를 거치게 된다(Agrawal, Srikant, 1994).

- 단계 1: 최소지지도를 만족하는 모든 항목들의 집합(itemsets)을 찾는다. 이러한 항목들의 집합(itemsets)을 빈발항목집합(large itemsets)이라 하고, 나머지를 small itemsets 이라 한다
- 단계 2: 빈발항목집합을 사용하여 요구되는 규칙(desired rules)을 찾는다.

Apriori 알고리즘에서는 먼저 항목 하나로 이루어진 빈발항목집합을 데이타베이스에서 찾고, 이를 Apriori-gen 전략을 사용하여 두개의 항목으로 이루어진 새로운 후보항목집합을 만든다. 그리고 다시 데이타베이스를 스캔하여 후보항목집합중 최소지지도를 만족하는 것을 찾고 Apriori-gen 전략을 사용하여 세 개의 항목들로 구성된 후보항목집합을 만드는 일련의 반복적인 과정을 거친다. 최종적으로 더 이상의 후보항목집합을 생성할 수 없을 때까지 계속된다.

4. 데이타 웨어하우스 환경에서의 연관규칙의 표현방법

연관규칙탐사에 의해서 얻어진 규칙들을 데이

타 웨어하우스의 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 Relational Predicate Logic 을 개발하였다.

먼저 일차 술어 논리를 설명하면, 일차 술어 논리는 명제 논리의 표현력 한계를 극복하고자 논리학자들에 의해 개발되어 인공지능 분야, 전문가 시스템 분야에서 많이 사용되고 있다(Patterson, 1990). 이것은 명제 논리를 보다 일반화 시킨 것이다.

일차 술어 논리의 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. (n 개(n≥0)의 항(terms)을 가질 수 있다. P(t1, t2, t3, ..., tn)로 표기된다.)

일차 술어 논리의 표현 예를 위해, 다음과 같은 문장(statement)이 있다고 하자.

- E1: All employees earning $1400 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) for x is an employee.
- P(x) for x is president.
- i(x) for the income of x.(function)
- GE(u,v) for u is greater than or equal to v.
- S(x) for x is sick today.
- T(x) for 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 연관규칙을 기존의 일차 술어 논리로 표현시에 릴레이션을 표현하지 못하는 점을 보완하고자 일차술어논리를 확장하고 Relational Predicate Logic 이라고 명명했다.

주요 변경사항은 다음과 같다.

- 변수로써 데이타베이스의 릴레이션을 사용할 수 있다. 대문자로 표현하였다.
- 릴레이션의 특정 필드를 지칭할 때 '.' 기호 다음 해당 필드를 나타낼 수 있다. 소문자로 표현한다.
- 전체정량자는 모든 트랜잭션에 대해 연관규칙을 만족한다는 의미로 해석되는데 실제로 데이타베이스에서 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 어느정도의 지지율을 가져야 하므로 단지 그러한 연관성을

갖는 트랜잭션이 존재한다고 하는 것은 큰 의미가 없다. 대신 연관규칙으로 성립하기 위한 최소한 지지율과 신뢰도 이상을 갖는다는 의미의 정량자가 필요하다. 그러나 연관규칙으로 나타내어지는 모든 규칙은 이러한 정량조건을 만족하는 규칙만을 다룬다고 보면 이 정량자는 생략해도 된다.

이제 Relational Predicate Logic 으로 표현한 연관규칙의 예를 들어보겠다. 다음은 슈퍼마켓 체인점들의 판매 데이타 웨어하우스를 대상으로 파악하는 연관규칙들의 예이다.

- E1: 하단동 지점 에서는 생필품이 잘 팔린다.
- E2: 고객이 빵을 구입하면 우유도 구입한다.
- E3: 저녁시간대에 기저귀를 사면 맥주도 구입한다.

위의 문장을 기존의 연관규칙형태로 표현하면 다음과 같다.

- A1: 하단동 지점 → 생필품
- A2: 빵 → 우유
- A3: 저녁시간대 ∧ 기저귀 → 맥주

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

만약 그림 2 의 데이타 웨어하우스의 구조를 가지고 있을 때 위의 연관규칙을 Relational Predicate Logic 으로 표현하면

- A1' : Eq(STORE.street_name, '하단동') → Eq(PRODUCT.subcategory, '생필품')
- A2' : Eq(PRODUCT.subcategory, '빵') → Eq(PRODUCT.subcategory, '우유')
- A3' : Eq(PRODUCT.subcategory, '기저귀') ∧ Between(18, TIME.hour, 22) → Eq(PRODUCT.subcategory, '맥주')

여기서 PRODUCT, STORE 는 데이타 웨어하우스의 차원 테이블이며, 그 중 street_name, subcategory 필드를 나타낸다. 즉, 해당 차원 테이블의 특정 필드의 값으로 구성된 튜플의 집합을 나타낸다.

단, 여기서 → 의 의미를 연관성에 기초한 implication 이라고 본다.

5. 설명기반 데이타 마이닝

5.1 설명을 통한 연관규칙의 검증

연관규칙에 있어서의 문제점은 첫째 연관규칙 탐사과정을 거쳐 나온 연관규칙들이 과연 옳은 것인지를 판단해야 한다. 그러나 통계적으로 이미 지지율과 신뢰도를 가지고 데이타 웨어하우스로부터 생성된 것이기 때문에 그 규칙의 옳고 그름은 결국 사람이 판단하게 된다. 사람의 해석과 판단 과정은 데이타 마이닝의 중요한 단계이기도 하다.

본 논문에서는 연관규칙 탐사를 통해 데이타베이스로부터 귀납적으로 형성된 연관규칙 A → B 의 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. 비슷한 원리로 사람이 가지고 있는 도메인에 대한 지식으로써 데이타베이스로부터 생성된 연관규칙을 설명함으로써 해당 연관규칙을 새로운 지식으로 확증하며 이러한 설명이 있을 때까지 해당 규칙은 가설의 지위만을 가지게 된다.

앞 절에서 예시된 A1 ~ A3 의 연관규칙을 살펴보자. 왜 하단동지점에는 생필품이 잘 팔리는 것인가? 왜 기저귀를 사는 사람이 동시에 맥주도 사는 것인가? 만약 여기에 대한 설명을 할 수 없다면 해당 연관규칙은 잘못 생성된 것이라고 본다. 마치 Standard Poor 500 주식 인덱스와 방글라데시의 버터 생산량이 통계적으로 유의한 상관관계를 가지고 있지만 그 이유에 대한 설명을 할 수 없으므로 해당

규칙을 폐기시키는 것과 같다.

관리자의 입장에서도 데이타마이닝(연관규칙생성)으로부터 나온 규칙이 관리자가 도저히 납득할 수 없는 규칙이라면 해당 규칙을 구현시킨다는 것을 기대하기란 어렵다.

이제 앞의 연관규칙에 대한 설명을 앞에서 정의한 Relational Predicate Logic 을 통해 표현한다. 이것은 사람이 가지고 있는 사전적 지식을 나타내는 것이고, Predicate Logic 은 이러한 지식을 표현하는데 현재까지 매우 유용하고 널리 알려진 방법이므로 이 표현법을 이용하여 사람이 설명을 하게된다.

설명의 방법은 크게 두 가지로 제시한다.

첫째, 두개의 항목이 서로 연관성을 가지는 것은 공통인자가 있어서 동시에 두개의 항목에 영향을 주기 때문이다. 따라서 연관규칙에 포함된 항목들의 공통인자를 찾아내는 것이 설명하는 방법이 된다.

둘째, 연관되는 두개의 항목간에 매개 요소를 찾아내는 것이다. 이 매개요소가 항목간의 연관성을 중개한다.

이러한 설명원리를 가지고 앞의 연관규칙은 다음과 같이 설명된다.

- E1: Eq(STORE.street_name, '하단동') →
  Near_By_University(STORE) →
  Is_a(CUSTOMER, '자취대학생') →
  Need(CUSTOMER, '생필품') →
  Eq(PRODUCT.subcategory, '생필품')
- E2: Eq(PRODUCT.subcategory, '빵') →
  Make_thirsty('빵') →
  Eq(PRODUCT.subcategory, '우유')
- E3: Has(CUSTOMER,'baby') →
  Need(CUSTOMER,'기저귀') →
  Eq(PRODUCT.subcategory, '기저귀');
  Has(CUSTOMER, job) →
  Between(18, TIME.hour, 22);
  Gender(CUSTOMER,male) →
  Eq(PRODUCT.subcategory, '맥주');

E1 과 E2 는 두번째 설명전략으로 설명한 것이며, E3 는 첫번째 설명전략으로 설명하였다. 즉, 아기와 직장을 가진 남성이라는 공통인자가 맥주와 기저귀의 연관성을 야기한 것이다.

5.2 설명으로부터 가설의 생성과 가설 검정

이제 역으로 주어진 설명으로부터 새로운 연관규칙 가설을 성립할 수 있다. 이러한 가설은 다시 데이타 웨어하우스에서 통계적 검정 즉, 가설검정을 거쳐 지식으로 확증하게 된다.

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

예를 들어 E1 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 설명을 계속 유지시키되 다른 항목으로 치환함으로써 다음과 같은 새로운 가설이 형성된다.

- E1' : Eq(STORE.street_name, '대신동') →
  Eq(PRODUCT.subcategory, '생필품')
- E2' : Eq(PRODUCT.subcategory, '빵') →
  Eq(PRODUCT.subcategory, '주스')
- E3' : Eq(PRODUCT.subcategory, '아기쥬스') ∧
  Between(18, TIME.hour, 22) →
  Eq(PRODUCT.subcategory, '맥주')

이러한 가설이 정확한 지식인지는 데이타 웨어하우스에서 지지도와 신뢰도를 계산함으로써 연관규칙으로 검증되게 된다.

5.3 지식생성에 있어서 귀납적 방법론과 연역적 방법론의 통합 모형

설명기반 데이타마이닝은 아무런 사전 지식없

이 주어진 데이타로부터 귀납적으로 가설을 생성하는 방법과 이에 대한 사람의 사전지식에 바탕을 둔 설명을 토대로 연역적으로 새로운 가설을 생성하는 두 가지 방법론의 통합모형이 될 수 있다.

이를 그림으로 나타내면 다음과 같다.

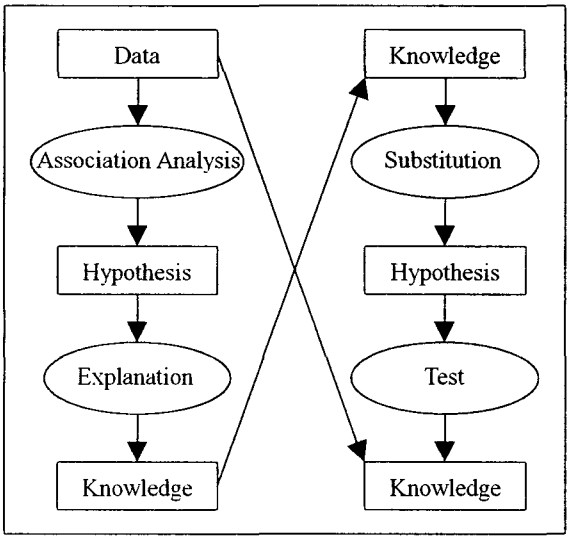

그림 3. 설명기반 데이터 마이닝의 구조

6. 결론

본 연구에서는 데이타 마이닝에 있어서 도출된 지식의 검증 방법론과 이를 위한 지식표현기법을 데이타 웨어하우스와 차원 테이블과 연계하여 개발하였다. 아울러 데이타로부터의 귀납적 지식생성과 사전지식으로부터 연역적으로 가설을 도출하여 이를 데이타에서 검증받는 두 가지 접근법의 통합모형을 제시하였다.

참고문헌

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", In 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", In Proceedings of the 20th VLDB Conference, Santiago, Chile, Sept., 1994.

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

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.

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

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*, pp. 109-136, March 1997.

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

Pearl, Judea, Causal diagrams for empirical research, *Biometrika*, 82(4), pp. 669-710, 1995.

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

Stedman, Craig, Data mining for fool's gold,
Computerworld, p.28, Dec. 1, 1997.