□ 기술해설 □

데이타 마이닝 기술 및 연구동향

전남대학교 김정자·이도헌*

1. 서 론

최근 여러 업무현장에서 정보 기술의 발달에 따라 데이타베이스에 저장되는 데이타 량의 현저한 증가, 데이타베이스 시스템 기술의 신속한 발전, 데이타베이스 시스템의 신뢰성이 증가하게 되었다. 이러한 데이타들은 본래의 운영 목적외에 현장의 특성을 반영해주는 실증적 정보원이라는 인식이 확산되면서 데이타 마이닝에 대한 관심이 날로 높아지고 있다.

데이타 마이닝은 대량의 실제 데이타로부터, 이전에 잘 알려지지는 않았지만, 묵시적이고, 잠재적으로 유용한 정보를 추출하는 작업이라 정의한다. '대량의 실제 데이타'란 실제 현장에서 생성되는 수천, 수백만 건 이상의 데이타를 의미하는 것이다. '이전에 잘 알려지지 않은 정보'라는 것은 현장에서 통용되는 상식적인 내용을 탐사대상으로 하는 것이 아니라 새로운 정보를 탐사대상으로 한다는 것을 의미한다. '묵시적'이란 데이타베이스나 시스템 카탈로그에 저장된 명시적 정보가 아닌 숨겨진 정보를 의미하며, '잠재적으로 유용한 정보'란 현장에서 의사결정, 성능 향상의 목적으로 활용할 수 있는 정보를 의미한다.

이와 같이 데이타 마이닝을 통하여 대용량의 데이타베이스에 숨겨져 있는 데이타간의 관계, 패턴의 탐색에 의해, 이를 의미있는 정보로 변환함으로써 기업의 의사결정 과정을 지원하고 그 결과를 예측할 수 있는 것이다.

데이타 마이닝은 기계학습, 통계학, 인공지능을 포함한 다른 연구 결과로부터 발전된 최신의 데이타 분석 기술로서 특히 기계학습과 비교되어 많이 설명된다. 기계학습 분야는 주입식 학습(rote learning), 연역적 학습(deductive learning), 귀납적 학습(inductive learning), 유사성 기준 학습(similarity-based learning)으로 나누어진다[1]. 데이타 마이닝은 구체적 사실인 데이타베이스 레코드들로부터 일반화된 지식인 규칙성을 발견하는 작업이므로, 위의 네가지 분야 중 귀납적 학습기법을 주로 활용하게 된다. 하지만 다음과 같이 적용 데이타의 특성이 상이하므로 그와 같은 차이점을 극복할 수 있는 연구가 필요하다[2].

첫째, 기존의 기계학습 기법은 주로 고정된 데이타 집합에 적용하는데 반해서 데이타 마이닝은 계속적으로 삽입과 삭제가 이루어지는 동적인 데이타 집합에 적용한다. 둘째, 기계학습에 사용되는 데이타는 정제과정을 거친 오류가 없는 데이타임에 반하여 데이타 마이닝에서는 현장에서 발생하는 다량의 데이타를 취급하기 때문에 오류 데이타도 포함된다는 것을 고려해야한다. 셋째, 기계학습 기법은 정확하면서 누락된 데이타의 값이 없어야 되지만 데이타 마이닝은 불확실한 데이타나 데이타 값이 누락되어 있어도 가능하다. 넷째, 기계학습에는 주제와 관련있는 데이타만 존재하지만 데이타 마이닝은 무관한 데이타와도 공재하면서 이들로부터 유용한 정보를 탐사하기도 한다. 다섯째, 기계학습에 사용되는 데이타는 대략 수천개 정도 밖에 되지 않으나 데이타 마이닝은 수십만 개 이상의 대규모이다. 여섯째, 기계학습은 단순 데이타의 집합을 취급하지만 데이타 마이닝은 필드의 집합, 객체의 집합 등과 같은 구조

* 정회원

화된 데이타를 취급한다.

본 논문에서는 현재까지의 데이타 마이닝의 연구 동향을 데이타 마이닝 기법별로 분류하고 현재 수행중인 데이타 마이닝 시스템들의 사례에 대하여 논의한다. 2절에서는 다양한 데이타 마이닝기법을 분류하고 3절에서는 데이타 마이닝에서 가장 활발히 연구되고 있는 연관규칙 탐사를 논의한다.

4절과 5절에서는 자료의 분류와 자료의 요약을 다루겠으며, 6절에서는 현재 수행중인 국외의 데이타 마이닝 시스템을 소개한다. 7절에서는 데이타 마이닝과 데이타 웨어하우스와의 관계를 기술하면서 향후 연구과제에 대하여 논의한다.

2. 데이타 마이닝 기술의 분류

최근 수년동안 학계, 연구계, 산업계에서 데이타 마이닝에 대한 연구가 이루어져 왔다. 그간 제안된 다양한 데이타 마이닝 기법들은 어떤 형태의 지식을 탐사하고자 하는가, 어떤 종류의 데이타베이스에 적용될 수 있는가, 어떤 분야의 기술에 바탕을 두고 있는가 등의 기준에 의거하여 아래와 같이 분류한다[3].

2.1 탐사될 지식의 형태에 따른 분류

• 특성화(characterization) : 데이타 집합의 일반적 특성을 분석하는 것으로 일반화 및, 세분화 과정에 의한 자료 요약과정을 거쳐 특성규칙을 발견한다.

• 분류화(classification) : 다른 클래스에 대한 차별적인 특성을 도출한다. 이와 같은 차별적인 특성은 소속 클래스를 알 수 없는 미지의 객체가 있을 때, 그 소속 클래스를 결정하는데 활용된다.

• 군집화(clustering) : 유사한 특성을 갖는 데이타들을 묶음 지워주는 것이다. 인공지능 분야에서 분류는 감독학습임에 반해 클러스터링은 비감독 학습으로 불린다. 감독학습이란 감독자가 자료를 집단별로 구분해 놓고 분류기준은 컴퓨터 프로그램이 학습에 의하여 발견하도록 하는 방법이다. 비감독학습은 감독이 없이 컴퓨터 프로그램 스스로가 자료집단의 유사성을 바탕으로 집단을 나누어 나가는 방식이다.

• 연관규칙 탐사(association) : 여러 개의 트랜잭션들 중에서 동시 발생하는 트랜잭션의 연관관계를 발견하는 것이다. 규칙 발견에 사용한 측정값은 연관성의 신뢰요인으로 사용된다.

• 경향분석(trend analysis) : 시계열 데이타(주식, 물가, 판매량, 과학적 실험 데이타)들이 시간 축으로 변하는 전개과정을 특성화하여 동적으로 변화하는 데이타의 분석을 수행한다.

• 패턴 분석(pattern analysis) : 대용량 데이타베이스 내의 명시된 패턴을 찾는 것이다.

2.2 탐사될 데이타베이스의 타입에 따라

탐사될 데이타베이스가 관계형 데이타베이스인지 혹은 객체 지향형, 네트워크형, 계층형인지에 따라서 적용 할 수 있는 데이타 마이닝 기법이 달라지게 된다. 최근 POS 시스템에서 흔히 사용되는 트랜잭션 데이타베이스도 이와 같은 구분에 따라 고려 할 수 있다. 한편 공간 데이타베이스, 멀티 미디어 데이타베이스처럼 문자나 숫자가 아닌 복잡한 자료만을 포함하고 있는 경우에도 새로운 데이타 마이닝 기법이 필요하게 된다.

2.3 적용기술의 종류에 따라

데이타로 사용된 변수의 분포, 상관관계, 탐사된 규칙에 대해 확신하기 위한 수단으로 통계를 이용하고, 논리를 이용한 기호학습 방법도 사용한다. 신경망은 분류화를 위한 기법으로 많이 사용하며, 수행결과를 보다 효과적으로 보여주기 위해 가시화 기법을 사용하기도 한다[14]. 일반적으로 이러한 기법들은 독립적으로 사용되기 보다는 혼합적으로 사용되는 경우가 많다.

3. 연관규칙 탐사

이 절에서는 사건들의 동시 발생성을 규칙의 형태로 표현한 연관규칙에 대하여 살펴본다. 연관규칙에 대한 연구는 현재 IBM Almaden 연구소에서 QUEST project로서 활발히 진행되고 있다[4].

3.1 연관규칙의 정의

연관규칙이란 어떤 사건이 일어나면 다른 사건이 일어나는 관련성을 의미한다. 주어진 트랜잭션의 집합을 {item-1, …, item-n}라 하였을 때 서로 소의 관계인 두개의 부분집합 A= {item-$_{11}$, item-$_{12}$, …, item-$_{1m}$}과 B={item-$_{21}$, item-$_{22}$, …, item-$_{2k}$}의 연관규칙은 "A→B"로 표시된다. 이때 전제 항목들의 집합 A는 결론 항목들의 집합 B를 야기한다고 정의할 수 있다.

이와 같이 생성된 연관규칙은 고객 데이타베이스로부터 고객들의 구매품목들간의 관련성을 발견한다. 이는 진열장의 상품들을 진열하기 위한 선반 레이아웃 디자인 등에 이용되고, 우량고객에 대한 상품 카탈로그 발송 등의 다이렉트 메일링에 이용될 수 있다. 이외에도 특정 노드들에서의 경보음의 발생수가 다른 노드의 경보를 야기한다는 연관성을 발견한다면 컴퓨터 네트워크의 고장진단을 할 수 있다. 공용통신서비스에서 고객들의 사용패턴 등을 분석함으로써 기업의 마케팅 전략에 이용될 수 있다.

3.2 연관규칙의 척도

생성된 연관규칙이 트랜잭션들의 상황을 얼만큼 잘 뒷받침해 주는가는 두가지의 척도로서 측정한다.

• 지지도(support degree) : 생성된 연관규칙이 전체 아이템에서 차지하는 비율을 말한다. 즉 데이타베이스에 속한 전체 트랜잭션의 갯수중 그 연관규칙을 지지하는 트랜잭션의 비율을 의미한다.

• 신뢰도(confidence degree) : 연관규칙의 강도를 의미하며 전제부를 만족하는 트랜잭션이 결론부까지를 만족하는 비율을 말한다.

그림 1은 연관규칙 탐사과정의 예를 보인 것이다. 이는 Apriori 알고리즘에 의거한다[4]. 연관규칙 탐사과정은 크게 두 단계로 구성되는데, 첫단계는 높은 지지도를 갖는 아이템의 집합을 식별하는 작업이며 두번째 단계에서는 높은 신뢰도를 갖는 연관규칙을 도출하는 작업이다. 첫단계를 수행하기 위해 데이타베이스의

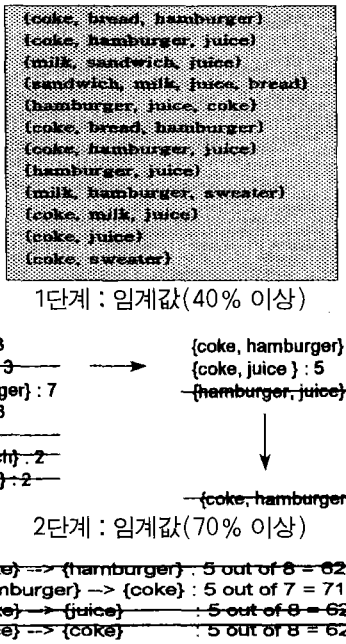


그림 1 Apriori 알고리즘을 이용한 연관규칙 탐사과정

트랜잭션을 조회하여 항목별 빈도수를 구한 다음 최소한의 지지도를 만족하는 항목만을 고른다. 이때 지지도에 대한 임계값이 40%라 주어졌다면 이를 만족하는 아이템들의 집합은 {coke}:8, {juice}:8, {hamburger}:7이 된다. 다음으로 이 항목들의 조합으로 구성된 사건항목 집합에 대해 지지도 이상을 만족하는 항목들을 반복하여 찾는다면 {coke, hamburger):5와 {coke, juice}:5의 결과를 얻는다. 두번째 단계는 신뢰도의 임계값이 70%라 하였을때 생성된 연관규칙은 최종적으로 {hamburger} {coke} : 5/7=71%을 얻게된다.

Apriori 알고리즘은 사건항목 집합의 크기가 하나씩 늘려갈 때마다 전체 데이타베이스를 조회해야 한다는 문제점을 안고 있다. 또한 단순히 카운트 값만을 가지고 연관성을 추측할 수 있느냐의 비판도 인다. 이러한 문제점을 해결하기 위해 현재 GSP, AprioriSome and AprioriAll, AprioriHybrid와 같은 많은 변형 알고리즘들이 제안되고 있다[5, 6, 7].

3.3 연속 패턴 탐사

연속 패턴이란 특정 아이템의 과정이 일련의

CID	Time	Items
1	95/06/25	30
1	95/06/30	90
2	95/06/10	10,20
2	95/06/15	30
2	95/06/20	40,60,70
3	95/06/25	30,50,70
4	95/06/25	30
4	95/06/30	40,70
4	95/07/25	90
5	95/06/12	90

CID	Sequence
1	<(30)(90)>
2	<(10,20)(30)(40,60,70)>
3	<(30,50,70)>
4	<(30)(40,70)(90)>
5	<(90)>

지지도 25%이상인
연속패턴은 :

<(30)(90)>
<(30)(40,70)>

그림 2 연속 패턴 탐사의 예

순서적으로 일어나는 경향을 말한다. 즉 앞서 설명한 연관규칙에 시간적인 관계를 추가한 것으로 이해하면 된다.

연속패턴 탐사의 예는 그림 2와 같다.

어떤 상점에서 고객번호와 판매날짜, 품목항목으로 구성된 고객 데이타로부터 고객별로 구입상품 품목에 따른 시퀀스를 조사하였다. 이중 지지도가 25%이상인 연속패턴은 〈(30)(90)〉 항목 3건, 〈(30)(40)(70)〉 항목 2건으로 선택된다. 즉 (30)항목을 사고 나면 (90)항목을 사며, (30)항목을 사면 (40)(70)항목을 사더라는 예가 전체 항목의 25%이상을 차지한다는 것이다.

4. 자료의 분류

자료의 분류는 귀납적 학습문제에서 가장 많이 연구가 되어진 분야로서 각 클래스가 갖는 특징에 근거하여 분류하는 것이다. 분류화를 표현하기 위한 방법으로는 결정 트리 접근(decision tree approach) 방법이 가장 많이 사용된다. 그림 3은 테이블을 결정 트리로 나타낸 것이다.

A1	A2	C
a	d	1
a	e	2
b	f	3
b	g	3

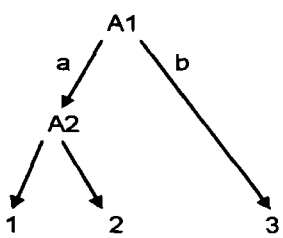


그림 3 결정 트리(Decision Tree)

분류화를 위한 표현 기법은 첫째 정확해야하며, 둘째 효율적이고, 셋째는 대량의 데이타에 적용 가능한 알고리즘이어야 한다. 이외에 신경망에 내재된 수치적인 자료에 의해서 분류가 이루어 지는데 데이타 마이닝 결과가 숫자로만 나타나서 사람이 이해하기 힘들고 학습시간과 비용이 많이 든다.

4.1 애트리뷰트 종속성

자료분류의 일반화된 형태는 속성간 종속성으로 표현할 수 있다. 애트리뷰트 종속성은 애트리뷰트 A1, A2, ..., Am이 주어졌을때 f(A1, A2, ···, Am, 상수의 집합)→g(A1, A2, ···, Am, 상수의 집합)로 표현 할 수 있다. 예를 들면 즉 "A1=c1이고 A2=c2이면 A3=c3이고 A4=c4이다"는 속성간 종속성을 생각할 수 있다. 하지만 f나 g는 임의의 값으로서 주어질 수 있는 함수의 가지 수가 이론적으로 너무 많기 때문에 다루기 쉬운 문제가 아니다. 그래서 실제 도메인에 적용가능 하도록 f, g를 제한해야한다. 예를 들면 f는 단순 술어들의 집합으로, g는 클래스의 레이블의 형태로 제약하게 되면 앞서 언급한 자료 분류 문제로 귀착된다.

5. 자료의 요약

데이타베이스 요약은 방대한 양의 데이타베이스 레코드들을 적은 양의 일반화된 대표적 표현으로 축약시키는 작업을 의미하며 자료 이해도의 증진, 자료 전달 및 저장의 효율화와 같은 목적을 위해 수행된다.

다음은 효과적인 자료의 요약을 위해 ISA 계층에 근거한 상향식과 하향식의 두가지 방식을 논의한다.

5.1 상향식 요약기법[8]

상향식 요약기법은 수십 만개의 레코드를 수개의 일반화된 형태의 데이타로 요약하는 것이며 그림 4는 캐나다의 Simon Fraser대학에서 연구된 DBLEARN project의 내용이다.

데이타베이스 각 투플의 필드값은 주어진 ISA 계층상의 상위 개념으로 대치되며 이렇게

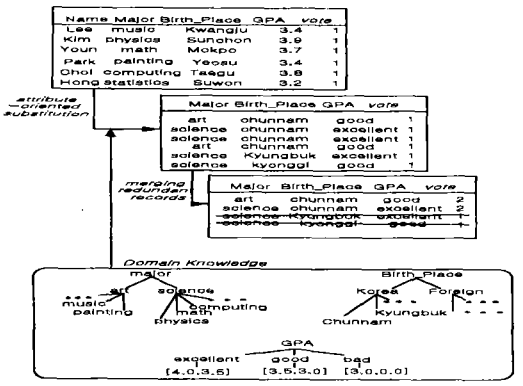


그림 4 상향식 요약방법

얻어진 일반화 투플중에서 동일한 값을 가진 것들을 하나로 합병시키면서 vote에 부가한다. 이와 같은 대치-합병 과정을 반복 후에 사용자가 원하는 수준에 도달하였을 때 중지한다. 이 기법의 장점은 일반화가 진행되가면서 읽어야 할 데이타의 양이 현저하게 줄어든다는 것이다. 그러나 대치 시마다의 테이블의 생성으로 디스크 공간이 많이 소요되며, 일반화하면서 중간단계의 지식이 상실될 우려가 있다. 현재 테이블을 통합 시에 시간, 비용을 줄이는 문제가 연구되고 있다.

5.2 하향식 요약기법[9]

이 기법은 가장 일반적인 가정에서 출발하여 사용자가 관심 있는 세부적인 내용으로 전개해 나가는 방식이다. 그림 5는 KAIST DB Lab에서 연구된 CLEVER시스템에서의 하향식 요약에 의한 탐사방법이다. 주어진 ISA 계층상의 최상위 개념을 이용하여 모든 가능한 일반화 투플을 가설로 설정한 후, 각 가설이 데이타베이스 투플을 얼마나 대표하는지를 계산한다. 사용자가 제시한 임계값 이상인 투플들은 다시 가설로 설정되고 위의 과정을 ISA 계층의 깊이만큼 반복한다. 이 기법은 탐사단계 도중에 사용자의 의도를 반영하여 선택적인 자료의 요약이 가능하므로 사용자의 요구를 충분히 반영할 수 있다. 그러나 매 단계마다 데이타베이스의 내용을 읽어야 하는 단점이 있다.

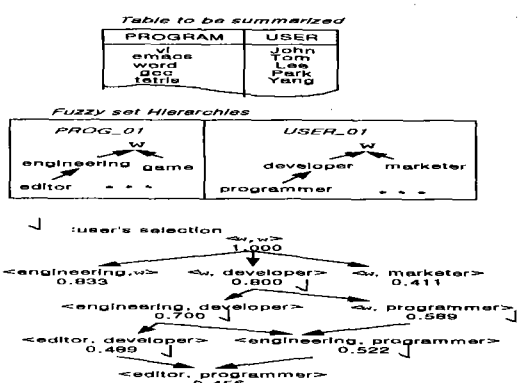


그림 5 하향식 요약방법

6. 데이타 마이닝 시스템 사례

다음은 국외에서 만들어진 데이타 마이닝 시스템을 소개한다. 캐나다의 Simon Fraser대학에서 만들어진 DBMiner/GeoMiner WebMiner와 Silicon graphics사의 MineSet외에 다른 프로젝트들에 대해 논의한다.

6.1 DBMiner[10]

이 시스템은 캐나다의 Simon Fraser대학 DBMiner 연구팀에 의해 만들어졌다. 대형 관계 데이타베이스로부터 다양한 지식의 종류를 발견하고자 데이타 마이닝 기법과 데이타베이스 기술을 통합하여 설계되었다. 대화식 데이타 마이닝을 위해 SQL과 유사한 DMQL(Data Mining Query Language)을 질의어로 사용하고 그래픽 사용자 인터페이스를 제공한다. 효과적인 구현을 위한 자료의 구조로서는 릴레이션과 다차원 큐브 모델을 사용하였다. 또한 데이타 큐브 구축과 처리, 애트리뷰트 지향 유도(attribute-driven induction), 다중 레벨 지향분석, 통계적 데이타 분석, 메타-룰 가이드 마이닝 등의 기법을 사용하였다. 그림 6은 DBMiner의 일반적인 구조를 나타낸 그림이다.

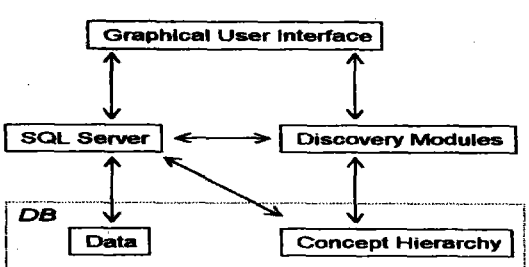


그림 6 DBMiner의 구조

DBMiner는 다음과 같이 8가지의 탐사모듈로 구성되어 있다. 특성화기(Characterizer)에서는 사용자가 명시한 자료 집합의 일반적인 특성을 탐사한다. 특성화 규칙의 예를 들면 "감기에 걸리면 두통과 기침을 한다"이다. 차별화기(Discriminator)는 특정 데이타 집합이 다른 데이타 집합과 대조되는 특성을 분석하는 것이다. 차별화 규칙의 예를 들면 "열이 있으면 백혈구 수가 많아지는 반면 열이 없으면 정상적인 백혈구 수치를 갖는다"이다. 분류화기(Classifier)는 연습 데이타를 분석한 다음 결정 트리의 유도에 의해 객체가 어느 클래스에 속하는지를 검사하여 그 객체의 특성들을 미리 알 수 있도록 해준다. 분류 규칙의 예를 들면 "복통과 오한은 맹장염의 증상이며 외과적인 질환에 속한다"이다. 연관규칙 탐사기(Association rule finder)는 데이타 집합간의 관련성을 나타내 준다. 연관규칙의 예를 들면 "인간의 뇌에 5분이상 산소가 공급되지 않는다면 쇼크 상태가 된다"이다. 메타룰 가이드 마이너(Meta-rule guided finder)는 사용자가 제시한 메타 규칙(패턴)을 처리해주는 부분이다. 전공과 평균평점과의 관계성을 찾기위한 메타 규칙의 예는 "major(s:student, x)∧P(s, y)→gpa(s, z)"로 나타낸다. 예측기(Predictor)는 누락된 자료값에 대해 값의 분포를 알고있는 다른 속성 값을 분석함으로써 가능한 값을 예측할 수 있도록 해준다. 특정 사원의 봉급을 그에 준하는 다른 사원의 봉급으로부터 유추할 수 있을 것이다. 자료 진화 분석기(Data evolution evaluator)는 시간에 따른 데이타의 동향을 분석하기 위한 것이다 예를 들면 "광주백화점 가전제품부의 작년 월별 매출추이 그래프를 작성하라" 등이 이에 속한다. 편차 분석기(Deviation evaluator)는 주요 패턴을 따르지 않는 예외적인 특성을 갖는 지식을 얻기 위한 것이다. 예를 들어 "이번주의 주식동향 경향에 예외적인 주식과 그 원인을 분석하라" 등이 편차 분석에 속한다.

6.2 GeoMiner/WebMiner[11, 12]

실세계를 컴퓨터로 모델링하기 위해서는 많은 문제가 있을 것이다. 특히 실생활이 이루어지는 공간을 모델링 하고자 하는 요구가 증가되면서 공간 데이타베이스의 구축과 공간 데이타 마이닝이 발전되었다. 공간 데이타는 위치, 거리, 위상 등 다양한 형태를 포함하며 실제 비 공간 데이타와 공존한다.

GeoMiner 시스템은 위의 DBMiner system을 확장하여 공간 데이타를 처리할 수 있도록 개발하고 있는 시스템이다. 질의어로서는 GMQL (GEO-Mining Query Language)을 제공하고 마이닝 결과를 테이블, 챠트, 지도 등의 형태로 출력하기 위한 대화식, 그래픽 사용자 인터페이스를 제공해준다.

Geominer에서는 비공간 데이타는 DBMiner에 의해 행해지며 공간데이타 처리를 위한 기능 모듈로서는 Geo-characterizer, Geo-comparator, Geo-associator로 구성된다. 예를 들자면 지역에 따른 캐나다 서부의 일반적인 날씨 패턴은 Geo-characterizer로부터, 브리티쉬 콜럼비아와 알베르타간의 날씨 패턴의 차이는 Geo-comparator에 의해 탐사될 수 있을 것이다. 미 서부 지역은 기온이 따뜻하고 비가 많이 오므로 나무가 잘 자란다는 규칙은 Geo-associator로부터 탐사된다.

WebMiner는 인터넷과 인트라넷을 위한 지식탐사 시스템으로 특정 주제와 관련된 문서의 URL을 인터넷을 통하여 찾아줄 수 있다. 현재 논의가 되는 web traversal pattern discovery의 경우를 살펴보자. 예를 들어 사용자가 특정 웹 홈페이지를 h1→h2→h5를 방문하고 난후 h8→h11로 방문하는 경향이 있다는 사실이 탐사되었다. 이러한 사실은 실질적인 시스템을 구현시에 이 정보를 적용하여 설계함으로써 성능을 증진시킬 수 있고 효과적인 마케팅 결정을 내릴 수 있도록 안내할 수 있을 것이다[13].

6.3 MineSet[14]

Silicon Graphics사에서 개발된 최초의 상업용 시스템이다. 다차원 비쥬얼라이제이션 기술과 지능적인 데이타 마이닝 알고리즘을 결합하여 놀랄 만큼의 데이타 분석능력을 제공한다.

시스템의 탐사 기능 모듈은 다음과 같다. 연관규칙 생성기(Association rule generator)는 데이타 엔티티간의 관련성과 빈도를 나타내

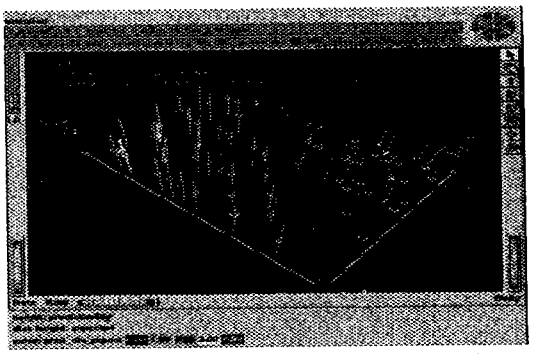


그림 7 MineSet의 규칙 비쥬어라이져

주며 이는 규칙 비쥬어라이져에(Rule visualizer) 의해 연관규칙의 두가지 척도와 함께 도식적으로 표현해준다. 그림 7은 규칙 비쥬어라이져의 실행 화면이다. 그림에서 바(bar)의 높이는 지지도를 나타내고, 링(ring)의 높이는 신뢰도를 나타내는데 링의 높이가 높을수록 신뢰도가 높다.

그 외에도 분류화를 위한 도구는 MLC^{++}을 이용한다. MLC^{++}은 특정 데이타 집합에서 서로 다른 알고리즘 사용을 비교하기 쉽게 하고, 주어진 데이타 집합에 적절한 분류 알고리즘을 선택, 새로운 알고리즘의 개발 등에 도움을 주는 C^{++} 클래스드들의 라이브러리를 제공해 준다[15]. 이외에도 결정 트리 유도기(Decision tree inducer)와 선택적 트리유도기(Optiontree inducer), evidence classifier inducer, decision table inducer가 있다. tree visualizer는 생성된 decision tree를 탐사하기 위해 사용된다. map visualizer는 3차원 지도상에 그래픽 컬한 요소들로서 공간 데이타를 분석하게 한다. 그 외 클러스터링 모듈이 있으며 regression tree inducer는 알려지지 않은 데이타 값의 예견에 이용된다.

이와 같이 MineSet은 여러 가지 분석 도구를 이용하여 데이타를 직관적으로 이해하게 하며, MLC^{++}과 같은 병렬적인 마이닝 알고리즘을 제공함으로써 고객의 의사결정처리에 혁신을 꾀하고 있다. 또한 Mineset의 중요한 인자와 그들의 상호 작용을 묘사하기 위한 새로운 비쥬얼 도구에서는 3차원 "fly-through", drill-up, drill-down, rotation, animation같은 다차원 네비게이션 기법을 사용함으로써 데이타의 분석을 용이하게 한다. 이러한 비쥬얼 패러다임은 변증적인 OLAP 연산자를 만들기 위한 비즈니스와 기술적인 면에 중점을 두는 사용자들을 위해 구현되었다.

6.4 QUEST project[16]

미국 IBM Almaden 연구소에서는 수년 전부터 POS(Point-of-Sales)시스템에서 소매품목간의 상호연관관계(association), 데이타 요약, 증권관련 정보의 시계열패턴, 각종 분류규칙등을 체계적으로 발견하기 위한 기법들을 개발해오고 있다. 또한 IBM사의 주력 데이타베이스 관리시스템 제품인 DB/2에 이와 같은 데이타 마이닝 엔진을 장착하기 위한 시도가 진행중이다.

6.5 KDW(Knowledge Discovery

Workbench[17] 미국 GTE연구소에서는 데이타 마이닝 시스템을 구축하기 위한 시스템 구조적인 측면을 집중적으로 연구하고, 데이타 클러스터링, 분류, 요약, 왜곡검출, 종속성 분석과 같은 작업을 종합적으로 수행할 수 있는 KDW 시스템을 구축한 바 있다. 아울러 동 연구소에서 최근 보건 데이타를 관리하기 위한 KEFIR 시스템을 발표한 바 있다.

6.6 IMACS(Intelligent Market Analysis and Classification System)[18]

미국 AT&T Bell연구소에서는 분류규칙 등의 지식을 발견하기 위한 데이타 마이닝 시스템에 사용자의 상호작용 측면을 접목하여 IMACS라는 시스템을 개발하고 있다. 특히, 동 연구소에서는 처음으로 데이타 고고학(data archaeology)라는 용어를 제안한 바 있다.

6.7 CoverStory : Information Resources incorporated

소매점 유통현황을 분석하기 위한 상품으로 슈퍼마켓 데이타상의 자료의 요약을 다루었다.

7. 결 론

데이타 웨어하우징은 기존의 온라인 트랜잭

션 처리(OLTP) 지향으로부터 온라인 분석 처리(OLAP) 지향으로 데이타 관리 전략을 전환하여, 데이타의 효용성을 극대화하기 위한 일련의 기법을 의미한다. 기존의 온라인 트랜잭션 처리는 현시점의 업무별로 필요한 데이타를 저장, 관리하여 업무상 발생하는 자료의 삽입, 갱신, 수정, 검색 등을 신속하고 안정성 있게 처리하는 기법이다. 한편 온라인 분석 처리라는 것은 업무처리를 위해 저장관리하는 데이타를 의사결정에 필요한 주제 지향적이며, 통합적이고, 시간적인 데이타를 다루며, 비휘발성인 데이타로 가공하여 의사결정자가 유용한 분석결과를 손쉽게 획득할 수 있도록 해주는 작업을 의미한다. 따라서 기존의 데이타 베이스들로부터 데이타를 추출하여 분석에 적합한 정보로서 가공한 다음 이를 저장하기 위한 데이타 웨어하우스 구축이 필요하며 그 핵심이 바로 데이타 마이닝 기술이다. 점점 더 복잡해져 가는 의사결정 사안들의 추세로 보아, 앞으로는 온라인 분석 처리와 이에 대한 관심이 더욱 더 높아질 것이고 이를 위한 데이타 웨어 하우스의 구축은 필수적이라 할 것이다. 그림 8은 데이타 웨어하우스에서 데이타 마이닝이 차지하는 위치를 도식적으로 보여주고 있다.

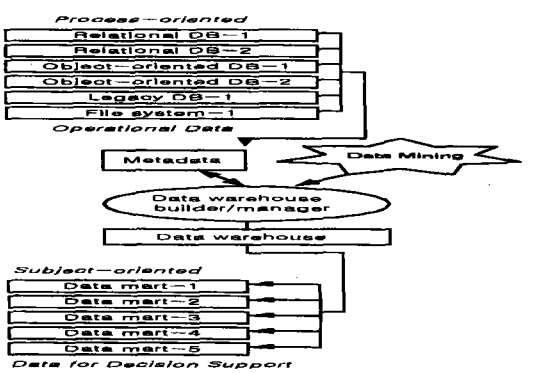


그림 8 데이타 웨어하우스와 데이타 마이닝

데이타 마이닝 분야에는 앞으로도 많은 연구과제들이 남아있다. 먼저 유용한 마이닝 주제를 찾아야 할 것이다. 연관규칙분야에서는 규칙의 흥미도나 중요도 측정문제, 일반화되고 다단계 연관규칙과 같은 연관규칙의 변형문제가 논의되어야 할 것이다. 연관규칙 탐사에 대한 성능증가 문제들이 연구되어야 할 것이고 효율적인 마이닝 시스템을 구축하기 위해서는 공통 연산자가 정의되어야 한다. 다양한 데이타 마이닝 작업을 위한 공통적인 DBMS 요구조건을 유도하여야 하며 탐사지식을 보다 시각적으로 표현해주는 문제 또한 필요할 것이다. 확장된 질의어, 사용자와의 인터페이스 또한 남아있는 문제이다.

참고문헌

[1] P. Cohen and E. Feigenbaum, The Handbook of Artificial Intelligence, Vol. 3, Willial Kaufmann Inc, 1982.

[2] G. Piatetsky-Shapiro and W. Frawley, "Knowledge Discovery in Database: An Overview," *Knowledge Discovery in Databases*, AAAI/MIT Press., pp. 1-27, 1991.

[3] M.-S. Chen, J. Han and P. Yu, "Data Mining : An Overview from Database Perspective", *IEEE Trans. on Knowledge and Data Engineering*, 1997.

[4] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Database", *Proc. ACM SIGMOD*, pp. 207-216, 1993.

[5] R. Srikant, and R. Agrawal, "Mining Sequential Patterns : Generalizations and Performance Improvements", *Proc. the 5th EDBT* 1996.

[6] R. Agrawal and R. Srikant, "Mining Sequential Patterns", *Proc. ICDE, March* 1995.

[7] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules", *Proc. VLDB*, pp. 487-499, 1994.

[8] J. Han, Y. Cai and N. Cercone, "Knowledge Discovery in Databases : An Attribute-Oriented Approach", *Proc. VLDB*, pp. 547-559, 1992.

[9] D. H. Lee and M.H. Kim, "Database Summarization Using Fuzzy ISA Hier-

archies", *IEEE Transactions on Systems, Man and Cybernetics*, 27 (4), pp. 671-680, August 1997.

[10] J. Han, Y. Fu, W. Wang et. Al, "DBMiner : A System for Mining Knowledge in Large Relational Databases", *Proc. Int'l Conf. on Data Mining and Knowledge Discovery(KDD 96)*, Portland, Oregon, 1996.

[11] J. Ham et.Al., "GEOMiner : A System Prototype for Spatial Data Mining", Proc. SIGMOD, 1997.

[12] "WebMiner : A Resource and Knowledge Discovery System for the Internet", http://db.cs.sfu.ca/WebMiner/

[13] M.-S. Chen, J. S. Park and P. S. Yu, "Data Mining for Path Traversal Patterns in a Web Environment", *Proc. the 16th ICDCS*, pp. 385-392, 1996.

[14] C. Hall ed., "MineSet2.0 for Data Mining and Multidimensional Data Analysis", http://www.cgi.com/Products/software / MineSet / DMStrategies / index.html.

[15] R. Kohavi et. al., "Data Mining Using MCL^{++} : A Machine Learning Library in C^{++}", Proc. Tools with AI, pp. 234-245, 1996.

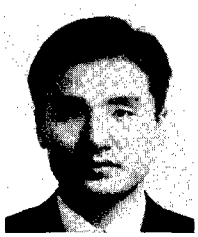
[16] R. Agrawal, M. Mehta, J. Shafer, R. Srikant, A. Arning, and T. Bollinger, "The Quest Data Mining System", *Proc. Int'l Conf. on Data Mining and Knowledge Discovery(KDD96)*, Portland, Oregon, 1996[10].

[17] G. Piatetsky-Shapiro and C. Matheus, "Knowledge Discovery Workbench for Exploring Business Databases", Int'l Journal of Intelligent Systems, Vol. 7, No. 7, pp. 675-686, Sep. 1992.

[18] R. Brachman and F. Halper, "Knowledge Representation Support for Data Archaeology", *Proc. CIKM*, pp. 457-464, 1992.

김 정 자

1985 전남대학교 자연과학대학 계산통계학과(이학사)
1988 전남대학교 자연과학대학 계산통계학과(이학석사)
1997~현재 전남대학교 자연과학대학 계산통계학과 박사과정 재학중
관심분야: 데이타 마이닝, OLAP, 데이타 웨어하우스
E-mail: jjkim@dbcore.chonnam.ac.kr



이 도 헌

1990 한국과학기술원 과학기술대학 전산학과(공학사)
1992 한국과학기술원 전산학과(공학석사)
1995 한국과학기술원 전산학과(공학박사)
1995 인공지능연구센터 연구원
1996~현재 전남대학교 전산학과 조교수
관심분야 : 데이타 마이닝, 데이타 웨어하우스, 워크플로우 관리, 데이타베이스
E-mail : dhlee.dbcore.chonnam.ac.kr