제2절 데이터웨어하우스 및 데이터마이닝

1. 데이터웨어하우스

가. 데이터웨어하우스 개요

(1) 데이터웨어하우스의 정의와 특징

데이터웨어하우스(Data Warehouse)는 1990년대 중반 이후 데이터베이스 분야에서 특히 학문계에서보다 산업계에서 그 태동이 시작되었다. 많고 다양한 형태의 오퍼레이셔날 데이터베이스(Operational Database)가 운용되고 시간이 지날수록 그 데이터베이스의 크기가 커지게 되었다. 따라서 데이터베이스 위에서 의사 결정(decision making) 등을 위해 사용되는 다양한 종류의 응용프로그램들은 그 질의(query) 수행이 원만하게 이루어지기 위하여 오퍼레이셔날 데이터베이스 위에 새로운 형태의 통합된 데이터 저장소(repository)가 필요했는데, 이것이 데이터웨어하우스가 등장한 배경이라 할 수 있다.

[그림 4-10-15] 데이터웨어하우스의 간단한 도식적 이해

데이터웨어하우스의 정의에는 여러 가지가 있지만 데이터웨어하우스 시스템 아키텍처를 초창기에 이끈 W.H. Inmon에 따르면, "데이터웨어하우스란 의사 결정 프로세스를 지원하도록 데이터를 1) 주제 지향적(Subject-Oriented)이고, 2) 통합(Integrated) 되고, 3) 시계열적(Time-Variant)이고, 4) 비휘발성(Non-Volatile)이게 모아 놓은 것(collection)"을 의미한다. 이 정의에서 사용된 4개의 의미는 다음과 같다.

(가) 주제 지향(Subject-Oriented)

데이터웨어하우스는 조직이 통상적으로 운용하는 트랜잭션 프로세싱을 위한 일반적이고 다양한 종류의 데이터의 저장소가 아니며, 의사 결정에 필요한 특정 주제(subject)의 데이터만을 가지고 그 외의 데이터는 포함하지 않는다.

(나) 통합(Integrated)

데이터웨어하우스에 저장, 관리되는 데이터는 일반적으로 다수의 서로 다른 형태의 데이터베이스로부터 통합(integrated)된 것이다.

(다) 시계열 (Time-Variant)

데이터를 이용해 의사 결정을 하는데 가장 유용한 측면중의 하나는 데이터가 시간에 따라 어떻게 변하였는지를 살피는 것이다. 따라서 대부분의 데이터웨어하우스에는 시간에 따라 변화된 데이터 정보를 저장한다.

(라) 비휘발성(Non-Volatile)

데이터웨어하우스는 오퍼레이셔날 데이터베이스와는 물리적으로 별도로 데이터를 저장한다. 오퍼레이셔날 데이터베이스에서 필요한 트랜잭션 관리, 복구 기법, 동시성 제어 기법 등은 중요시되지 않는 경우가 대다수이다. 그 대신 정기적으로 데이터를 오퍼레이셔날 데이터베이스로부터 로딩하고 로딩된 데이터를 액세스하는 기법이 중요시된다.

(2) 오퍼레이셔날 데이터베이스와의 차이점

데이터웨어하우징 시스템을 현재 상업적으로 사용되는 데이터베이스 시스템과 비교하면 데이터웨어하우징 시스템의 이해가 더 쉽다. 오퍼레이셔날 데이터베이스 시스템이 주로 조직이 필요로 하는 일상 업무(day-to-day operations)를 위한 OLTP(On-Line Transaction Processing)을 위한 시스템이라면, 데이터웨어하우징 시스템은 데이터 분석이나 의사 결정 등을 지원하는 OLAP(On-Line Analytical Processing)을 위한 시스템이다.

OLTP시스템과 OLAP시스템의 주요 차이는 다음과 같은 측면으로 요약 될 수 있다.

나. 다차원 데이터 모델

(1) 데이터 큐브

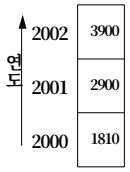
데이터웨어하우스와 OLAP 도구들은 다차원(multi-dimensional) 데이터 모델을 기반으로 하는데, 조직이 원하는 여러 차원(측면, dimension)에서 데이터 모델링이 데이터 큐브(Data Cube)를 통해 이루어진다. 데이터 큐브와 다차원 데이터 모델링을 설명하기 위해 먼저 차원(Dimension)과 사실(Fact)이란 두 용어를 설명하자.

차원(Dimension)이란 조직이 데이터 레코드를 운용하는 이유의 대상이 되는 측면을 의미하고, 사실(Fact)이란 숫자적으로 표현되는 값을 의미한다. 예를 들어 가나다 백화점이 백화점 관리에 따른 의사 결정을 위한 '가나다 판매분석 데이터웨어하우스'를 만드는데, 관심 있는 측면이 연도, 품목, 지점에 따른 총 판매 금액이라 하자. 이 경우 차원(Dimension)은 연도, 품목,

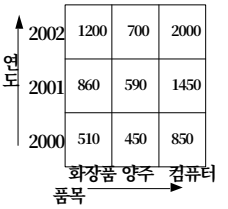
[표 4-10-16] OLTP 시스템과 OLAP 시스템의 차이점

특 징	OLTP	OLAP
주요 용도	오퍼레이셔날 트랜잭션	정보 분석
사용자	은행 창구원등의 일반 사용자, DBA등	조직 관리자, 분석가등의 지식 근로자
데이터베이스 설계 기법	ER 기반, 응용 프로그램 중심	스타/스노우플레이크 기반, 주제 중심
데이터	현재값 중심	시간에 따른 변화 중심
데이터 요약도(summarization)	요약 대신 개개 데이터 값 중심	데이터의 요약중심
작업 단위	짧고 간단한 트랜잭션 중심	대다수가 복잡한 질의
액세스 유형	읽고/쓰기 모두 필요	주로 읽기
액세스 레코드 수	수십개 정도로 많지 않음	수백만개 정도로 상대적으로 매우 많음
사용자 수	수천명 단위로 상대적으로 많음	수백명 단위로 상대적으로 적음
데이터베이스 크기	100MB에서 GB 정도 단위	100GB에서 TB 정도 단위
시스템 우선 순위	높은 성능과 높은 유용성 우선	높은 유연성과 사용자 자치성(autonomy) 우선
시스템 성능 평가척도	트랜잭션 쓰루풋(throughput)	질의 쓰루풋(throughput)과 응답 시간

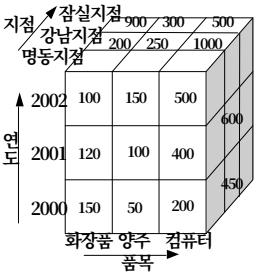
연도	총판매 금액
2002	3900
2001	2900
2000	1810



연도	품목별 총판매금액		
	화장품	양주	컴퓨터
2002	1200	700	2000
2001	860	590	1450
2000	510	450	850



연도	지점="명동지점"			지점="강남지점"			지점="잠실지점"		
	품목별 총판매금액			품목별 총판매금액			품목별 총판매금액		
	화장품	양주	컴퓨터	화장품	양주	컴퓨터	화장품	양주	컴퓨터
2002	100	150	500	200	250	1000	900	300	500
2001	120	100	400	140	200	600	600	290	450
2000	150	50	200	160	100	450	200	300	200

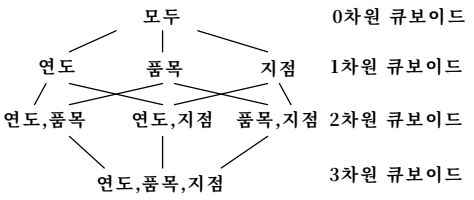


[그림 4-10-16] 다차원 모델과 데이터 큐브

지점 세 가지이고, 총 판매 금액은 사실(Fact)이 된다. 따라서 '가나다 판매분석 데이터웨어하우스'는 3차원 데이터 모델이 필요한데, 어떻게 데이터 큐브를 통해 이루어지는지 [그림 4-10-16]을 참고로 설명하기로 한다.

[그림 4-10-16] 상단 왼쪽 테이블은 연도라는 하나의 차원에 대한 총 판매 금액을 표시하며 이에 해당하는 1차원 데이터 큐브이다. 상단 오른쪽에는 연도와 품목이라는 두 개의 차원에 대한 총 판매 금액을 테이블 형태와 이에 해당하는 2차원 데이터 큐브가 있다. 여기에 지점별로 다시 총 판매 금액을 보여주기 위해 연도, 품목, 지점의 세 차원에 대한 테이블과 이에 상응하는 3차원 데이터 큐브가 그림 하단에 나타난다. 큐브란 용어 자체는 기하학적으로 3차원을 의미해 용어상 혼동을 줄 수 있겠지만 데이터 큐브에서는 원하는 차원의 일반적인 n차원을 의미한다.

다차원 데이터 모델을 위한 데이터 큐브에서 각각의 데이터 큐브를 큐보이드(cuboid)라고도 부르며, 이러한 큐보이드 간에는 [그림 4-10-17]과 같이 래티스(lattice) 관계가 형성되게 된다.



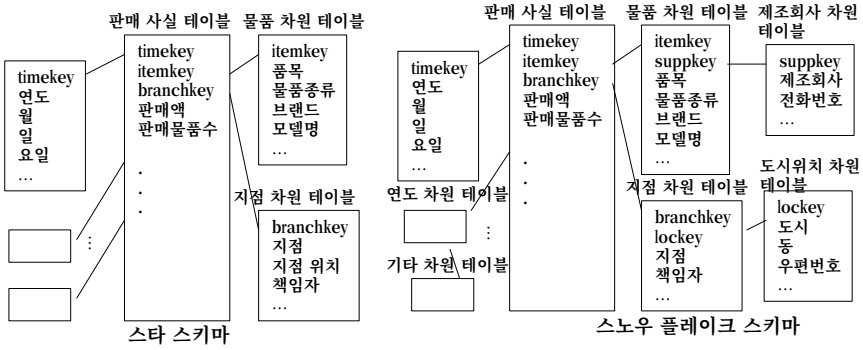
[그림 4-10-17] 데이터 큐브 래티스 구조

'가나다 판매 분석 데이터웨어하우스'를 이용하여 예를 들어 경영자가 지점별 예산 편성이나 특화 분야 선정을 하는 의사 결정을 내리기 위해 질의 패턴중 대표적인 것으로는 드릴다운과 롤업이 있다. 연도별 총 판매 금액, 다시 연도와 품목별 총 판매 금액, 그리고 연도, 품목, 지점별 총 판매 금액 정보를 계산해 내는 방식으로 점점 더 기존 차원에 또 다른 차원을 첨가해 세분화된 질의를 하는 질의 패턴을 드릴다운(Drill-Down)이라 한다. 이와 반대로 차원 수를 줄여가며 점점 요약된 형태의 정보를 얻어 나가는 질의 패턴은 롤업(Roll-Up)이라 한다. 이 밖에 슬라이싱(Slicing), 다이싱(Dicing), 피봇(Pivot) 등도 흔히 나타나는 질의 패턴이다.

(2) 스타 스키마/스노우플레이크 스키마

앞서 살펴보았듯이 데이터웨어하우스에는 다차원 모델이 적합한데 이를 논리적 (Conceptual)으로 설계할 때 제일 자주 사용되는 스키마는 스타 스키마(Star Schema)와 스노우플레이크 스키마(Snowflake Schema)이다.

스타 스키마는 필요한 사실과 기타 속성들로 이루어진 사실 테이블과, 관심있는 차원과 그에 따른 부가적인 정보들을 각각 하나의 차원 테이

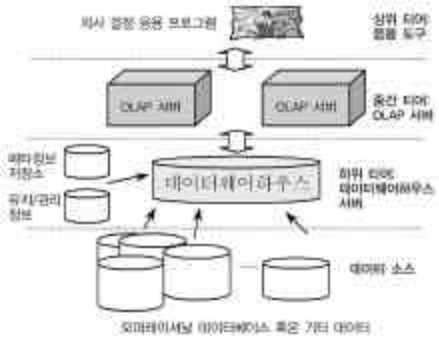


[그림 4-10-18] 스타 스키마와 스노우플레이크 스키마

블로 설계해, 사실 테이블과 차원 테이블을 외부 키(Foreign Key)로 연결할 수 있도록 한다. 그림 상으로 보면 마치 별 모양이 되도록 구성한 스키마이다. 스타 스키마의 변형인 스노우플레이크 스키마(Snowflake Schema)는 차원 테이블을 다시 정규화(Normalize)함으로서 이들 테이블들이 마치 눈꽃(Snowflake) 모양이 되도록 만든 스키마이다.

다. 데이터웨어하우스 아키텍처와 데이터웨어하우징 시스템

(1) 아키텍처와 OLAP 구현 분류



[그림 4-10-19] 데이터웨어하우징 시스템의 3-티어 아키텍처

데이터웨어하우징 시스템은 통상적으로 데이터 소스로부터 데이터를 클리닝하고 로딩하며 메타 정보 등을 관리하는 데이터웨어하우스 서버를 하위 티어에 두고, 중간 티어에 OLAP 프로세스 서버를 두며, 상위 티어에 데이터마이닝 등의 의사 결정 응용 프로그램 도구를 두는 3-티어 아키텍처를 가진다.

OLAP 서버는 크게 ROLAP과 MOLAP 그리고 HOLAP의 세 종류로 구분해 볼 수 있다. ROLAP(Relational OLAP) 서버란 관계형 데이터베이스나 확장된 관계형 데이터베이스를 사용해 다차원 모델링되는 데이터 큐브를 테이블 형태로 저장 운용하는 방식을 말한다. MOLAP(Multidimensional OLAP) 서버란 데이터 큐브를 실제로 어레이를 기반한 다차원 저장 엔진을 사용하여 저장 운용하는 방식이다. HOLAP(Hybrid OLAP) 서버는 말 그대로 ROLAP과 MOLAP을 혼용하는 방식을 의미한다.

(2) 데이터웨어하우징 시스템

현재 상업적으로 개발된 데이터웨어하우징 시스템은 많으나 그 중 몇 가지를 들면 다음과 같다. IBM사에서는 Informix eXtended Parallel Server(XPS), Readbrick Warehouse, DB2 OLAP Server 등이 있으며, Oracle사에서는 Oracle Data Warehousing이, Microsoft사에서는 MS SQL Server OLAP Services, NCR사에서는 TeraData Warehousing, Sybase사에서는 Sybase Anywhere Studio, Microstrategy사의 Microstrategy OLAP Services, Hyperion 사의 Essbase OLAP Server등을 들 수 있겠다.

2. 데이터마이닝

가. 데이터마이닝 개요

데이터웨어하우스와 비슷한 시기에 데이터베이스 분야를 중심으로 연구가 활발하게 시작된 분야가 데이터마이닝(Data Mining)이다. 90년대 중반이후에 데이터베이스나 데이터웨어하우스의 크기는 웹 로봇과 같이 데이터를 수동이 아닌 자동으로 수집하는 도구의 대중화와 오랜 기간의 시계열적 데이터 량의 증가 등으로 그 크기가 급속도로 커졌고, 이를 시스템이 기술적으로 감당할 수 있게 됨으로서 데이터마이닝이란 새로운 응용 분야가 탄생할 수 있었다.

데이터마이닝은 '대용량의 데이터에서 필요한 지식(Knowledge)을 얻고자 하는 과정'이라 간단히 정의할 수 있는데, 데이터베이스에서 분야에서는 지식 발견(KDD : Knowledge Discovery in Databases)이라고도 불리며, 그 응용 분야에 따라 비즈니스 인텔리전스, 지식 추출, 정보 분석 등의 용어로도 불린다. 여기서 발견하고자하는 지식은 뻔한 사실이 아니고(Non-Trivial), 데이터에 직접적으로 나타나지 않고 내포되었으며(Implicit), 기존에 발견되지 않은, 잠재적으로 유용한 지식을 의미한다.

데이터마이닝은 잠재적으로 폭 넓은 응용 분야를 가지고 있는데 고객 구매 성향 분석, 타겟 마케팅, 크로스-마켓팅 등의 마켓팅 분석 및 관리 분야, 위험 분석 및 관리 분야, 사기 상행위 발견 및 예측 분야, 텍스트 나 웹 등에서의 정보 발견, DNA 데이터 분석 및 의료정보학, 기타 스포츠, 과학 등을 망라하는데 데이터마이닝 기술이 발전하면서 그 응용의 폭이 점점 더 커지고 있는 추세이다.

나. 지식발견과정 : 데이터마이닝

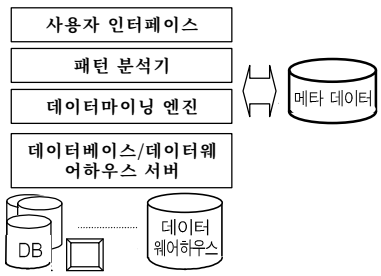
데이터마이닝은 일반적으로 다음 [그림 4-10-20]과 같은 과정을 밟는다.

[그림 4-10-20] 데이터마이닝 과정

- 데이터 클리닝(cleaning)과 통합 : 데이터 소스는 서로 다른 관계형 데이터베이스로에서부터 파일 혹은 이메일 자료 등 다양할 수 있으며 그 데이터 형식도 다양할 수 있다. 이 과정에서는 에러를 보정하고 포맷을 통일시키고 데이터의 일관성을 유지하며 스키마를 통합하는 등의 작업을 수행한다.
- 데이터 선별과 변환 : 데이터웨어하우스와 같은 통합된 데이터 저장소로부터 분석작업에 필요한 데이터를 선별하고 데이터마이닝을 수행할 수 있는 형태로 데이터를 변환한다.
- 데이터마이닝 : 데이터로부터 다양한 형태의 마이닝 기법을 적용시켜 패턴을 추출해 낸다.
- 패턴 평가와 표현 : 추출된 패턴이 얻고자 하는 지식에 필요한 것인지를 척도에 맞추어 평가해보고 궁극적으로 사람이 이해할 수 있는 방법으로 지식을 표현한다.

다. 데이터마이닝 시스템 아키텍처와 필요 기능

[그림 4-10-20]에서 제시된 데이터마이닝 과정을 처리하기 위한 데이터마이닝 시스템의 아키텍처는 [그림 4-10-21]과 같이 데이터베이스/데이터웨어하우스 서버, 데이터마이닝 엔진, 패턴 분석기, 사용자 인터페이스 등으로 구성될 수 있다.



[그림 4-10-21] 데이터마이닝 시스템 아키텍처

데이터마이닝 시스템에서 필요한 기능적 요소를 요약해 보면 다음과 같다.

(1) 개념 설명 : 특징화(Characterization)와 비교(Discrimination)

예를 들어 백화점 판매 분석 데이터웨어하우스에서 고객들의 분류에 따른 물품별 판매 패턴을 분석하여 그로부터 백화점 경영에 대한 지식을 얻는데 데이터마이닝 기법을 도입한다 하자. 그러면 SRAM, DRAM 등은 메모리로 분류되고, 다시 이는 하드디스크, PC, 모니터 등을 포함하는 컴퓨터라는 물품으로 분류되는 등의 물품 분류가 필요하다. 또 고객별 성향을 20대 30대 등의 나이별 분류에서 씀씀이가 큰 고객, 신세대 고객 등의 분류도 필요할 수 있다. 이와 같이 데이터를 클래스화하고 개념화하기 위해서는 비슷한 부류의 데이터가 가지고 있는 성질 등을 일반적인 용어를 사용하여 요약(Summarize)하는 특징화(Characterization)와, 비교되는 부류의 데이터와 대조를 하여 클래스화하는 비교(Discrimination) 기법 등이 필요하다.

개념의 효율적인 요약, 특징화 등을 위하여, 데이터 큐브 기반 일반화 기법, 속성중심 추론(Attributed-Oriented Induction) 기법, 표준편차(Standard Deviation)나 중앙집중 경향(Central Tendency)등의 통계적 요약기법, 일반화 기반 추론(Generalization-Based Induction) 기법, 속성 연관성(Attribute Relevance) 파악 기법 등이 사용된다.

(2) 연관성 법칙(Association Rule)

연관성 법칙(Association Rule)이란 어떤 속성들이 가지는 값이 자주 나타나는 조건을 보여주는 것인데, 형식적으로는 $A1 \wedge A2 \wedge ... \wedge An \Rightarrow B1 \wedge B2 \wedge ... \wedge Bm$ 같이 논리적 폼으로 쓰여질 수 있다. 여기서 프레디킷(Predicate) Ai와 Bj에는 각각 속성과 그 값이 나타내며 $\wedge$는 논리곱을 의미한다. 예를 들어 나이(X,"35..45")$\wedge$성별(X,"남자")$\wedge$자녀여부(X,"예")$\Rightarrow$구매(X,"컴퓨터") [지지도=40%, 신뢰도=75%]라는 연관성 법칙은 "나이가 35에서 45세 사이에 있고 자녀가 있는 남자는 컴퓨터를 구매한다"를 의미한다. 연관성 법칙은 지지도(support)와 신뢰도(confidence)가 같이 수반될 때 연관성 법칙으로서의 의미가 제대로 파악될 수 있다.

지지도는 연관성 법칙에 나타나는 각 프레디킷을 모두 만족시키는 데이터가 전체 데이터에서 나타나는 확률을 의미하고, 신뢰도란 주어진 조건 안에서 법칙이 성립하는 확률을 의미한다: 지지도$(A \Rightarrow B) = P(A \cup B)$, 신뢰도$(A \Rightarrow B) = P(A|B)$. (P는 확률을 의미함)

예를 들어 위에서 지지도와 신뢰도의 확률 계산이 트랜잭션 숫자에 기준한 통계로 이루어 졌다면, 지지도=40%가 의미하는 바는 전체 백화점 판매 데이터베이스에서 나타나는 트랜잭션 숫자 중에서 나이가 35에서 45세 사이에 있고 자녀가 있는 남자가 컴퓨터를 구매한 트랜잭션의 숫자의 비율이 0.4임을 의미한다. 또 신뢰도=75%가 의미하는 바는 나이가 35에서 45세 사이에 있고 자녀가 있는 남자가 구매한 트랜잭션 가운데 컴퓨터를 구매한 트랜잭션의 비율이 0.75라는 의미이다.

일반적으로 높은 지지도와 높은 신뢰도를 가진 연관성 법칙일수록 좋은 법칙이라 할 수 있겠는데, 최소지지도와 최소 신뢰도를 시스템이 정하고 이를 넘는 지지도와 신뢰도를 가진 연관성 법칙을 스트롱(Strong) 연관성 법칙이라 부른다. 대용량 데이터베이스에 존재하는 연관성(특히 스트롱 연관성)을 어떻게 효율적으로 찾을 수 있는지, 얼마만큼 많은 연관성 법칙들을 찾을 수 있는지, 혹은 필요한 연관성 법칙이 무엇인지를 알아내는 등이 주요 이슈가 된다.

(3) 클래스화(Classification) 혹은 클러스터링(Clustering)

클래스화란 서로 비슷한 특징을 보이는 데이터 혹은 객체들로 분류하는 모델을 찾는 과정인데, 흔히 훈련 데이터(Training Data)로부터 적당한 클래스화를 얻어본 후 실제 데이터에 구해진 클래스에 따른 클래스화 및 클래스에 속한 객체에 대한 특징 예측(Predict)을 해나간다. 클

러스터링도 객체를 분류한다는 측면에서는 비슷하지만, 훈련 데이터를 통하거나 사전에 미리 클래스화해서 실제 데이터를 그 클래스화에 적용하는 방식이 아니고, 비슷한 특징을 가진 객체끼리 클러스터간 유사성은 커지면서 동일 클러스터내의 객체간 유사성은 높이도록 분류해 나간다는 것이 그 차이점이다. 흔히 기계학습(Machine Learning)분야에서는 이러한 클래스화와 클러스터링의 차이를 감독학습(Supervised Learning)과 비감독학습(Unsupervised Learning)으로 말하기도 한다.

클래스화 기법은 크게, 의사 결정 트리 추론(Decision Tree Induction) 기법, 인스턴스 기반(Instance Based) 기법, 베이지안 네트워크(Bayesian Network) 기법, 신경망(Neural-Net) 기반 기법, k-근접 이웃(k-Nearest Neighbor) 기법, 유전자(Genetic) 기법 등으로 분류될 수 있다. 이에 비해 클러스터링 기법은 파티셔닝(Partitioning) 기법, 계층화(Hierarchical) 기법, 밀집-기반(Density-Based) 기법, 그리드-기반(Grid-Based) 기법, 모델-기반 기법 등이 있다.

(4) 비쥬얼화(Visualization)

분석된 패턴이 이해할 수 있는 지식이 되기 위해서는 비쥬얼화가 중요하다. 비쥬얼화에는 데이터를 3차원 큐브, 분포 차트 등의 다양한 형태로 보여주는 데이터 비쥬얼화 뿐만 아니라, 데이터마이닝으로 얻어진 연관성 법칙이나 클러스터 등을 플로팅하는 등의 데이터마이닝 결과 비쥬얼화, 데이터마이닝 과정 자체에 대한 비쥬얼화 등이 포함된다.

라. 데이터마이닝 시스템

현재 상업적으로 개발된 데이터마이닝 시스템 중 몇 가지로서는, IBM사의 Intelligent Miner, SAS Institute 사의 Enterprise Miner, Silicon Graphic 사에서 개발한 MineSet, Oracle사의 Data Mining Suite 등을 들 수 있겠다.

마. 향후 전망

현재까지 데이터마이닝 연구 및 개발은 주로 수평적 시스템을 구성하기 위한 기본 연구에 치중하였으며, 그 응용 영역도 확대해 왔다. 앞으로의 연구는 특정 응용 분야에 한정한 수직적(Vertical) 데이터마이닝으로 그 활용 영역을 넓혀갈 것으로 보인다. 예를 들어 웹마이닝(Web Mining), 바이오정보학 마이닝(Bioinformatics Mining)은 좋은 예라 할 수 있다. 또 수직적 데이터마이닝 못지 않게 더 인텔리전트하고, 효율적이고, 또 대용량 데이터베이스에서도 적용가능한(Scalable) 데이터마이닝 시스템을 만드려는 연구 개발도 당분간 지속되리라 본다. 관련분야의 세계적인 학술대회인 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 이나, 데이터베이스 전반을 다루는 ACM SIGMOD International Conference on Management of Data, 그리고 International Conference on Very Large Data Bases 등에서의 데이터웨어하우스나 데이터마이닝의 2001, 2002년도 최근 연구는 이를 반영한다 할 수 있다.

데이터웨어하우스와 데이터마이닝은 그 수요에 있어서는 별개의 제품으로서가 아닌, 데이터마트와 같은 전자상거래 시스템(e-Commerce), 고객관리시스템(CRM, Customer Relationship Management), 공급사슬관리(SCM, Supply Chain Management), 기업애플리케이션통합(EAI, Enterprise Application Integration), 비즈니스 인텔리전스(Business Intelligence) 등의 비즈니스 시스템 통합형태로 시장을 형성해 왔으며, 앞으로도 그러한 수요가 지속되리라 예견된다.