데이터베이스의 신 응용분야

나민영
육군사관학교 전산학과 교수

데이터베이스 시스템은 산업사회가 정보화사회로 변함에따라 그 응용분야가 날로 확장되어 우리 실생활에 크게 이바지하고 있다. 처리해야 할 데이터가 점점 많아지고 컴퓨터 네트워크가 발달함에 따라데이터베이스 시스템 기술은 그 응용분야가 더욱 확장되고 있다. 본고에서는 이러한 데이터베이스 응용 분야중 최근 정보선진국에서 경쟁력 강화를 위한 새로운 응용분야로 인식되어 연구중인 응용 분야인 데이터 마이닝, 데이터 웨어하우징, CALS 통합데이터베이스 등을 간략히 살펴보기로 한다.

1. 데이터 마이닝

1.1 데이터 마이닝의 개념

데이터 마이닝(Data Mining)은 대규모 데이터베이스에 존재하는 감추어진 지식을 찾아내는 작업으로써 데이터베이스에서의 지식발견(Knowledge Discvery in Database)이라고도 하며 최근 들어 데이터베이스 분야에서 중요한 응용으로 각광을 받고 있는 분야이다.

데이터 마이닝의 목적은 의사결정에 또는 마케팅에 적용할 수 있는 중요한 정보를 캐내는 것이다.

데이터 마이닝의 기본 개념은 새로운 것이 아니라 인공지능 분야의 기계학습(Machine Learning) 이론에 그 뿌리를 두고 있다. 즉 현실 세계에서 데이터베이스가 발달하여 수많은 데이터가 쌓여가고 있으므로 이로부터 감춰진 유용한 정보를 캐내고자 하는 욕구가 데이터베이스 종사자들에게 일어나게 되어 기계학습에서 사용된 기법을 데이터베이스에 응용하기에 이르렀다.

기계학습은 규칙을 찾아내기 위한 자동화된 유도과정(Inductive process)이라 할 수 있다. 기계학습에서는 트레이닝 셋(training set)이라 불리우는 적은 양의 실험실용 데이터를 사용하여 알고리즘을 만들어 내는 작업이다. 그러나 이러한 일련의 기계학습 작업은 현실세계의 데이터베이스에는 적용하기가 곤란하다.

왜냐하면 현실 세계의 데이터베이스는 갱신이 수시로 이루어지는 등 다이내믹하고, 오류도 있을 수 있으며 데이터가 없을 수도 있고, 더우기 대량의 데이터를 보유하고 있기 때문이다. 따라서 데이터 마이닝에서는 현실세계의 대규모 데이터베이스를 트레이닝 셋으로 간주해서 이로부터 유용한 지식을 캐내는 일련의 작업인 것이다.

1.2 데이터 마이닝에서의 지식

데이터 마이닝에서 얻고자 하는 지식은 연관, 분류, 순서 등에 관한 지식들이다. 이를 간단히 설명하면 다음과 같다.

■ 연관(Association)

데이터베이스 각 항목간에 존재하는 연관규칙(Association rule)을 찾아내고자 한다.

예를 들어 수퍼마켓에서 운영하는 판매 데이터베이스를 생각해보자. 판매된 항목중 서로 같이 팔리는 연관성이 높은 항목들을 알수 있다면 상품진열, 상품주문 등 마케팅에 큰 도움이 될 것이다.

■ 분류(Classification)

데이터베이스에 있는 데이터들을 서로 중첩되지 않는 그룹으로 쪼개는 규칙을 찾아내고자 한다.

예를 들어 스키에 관련된 신상품을 판매하고자 하는 회사에서 이 상품에 관한 카달로그를 우송하여 광고하려할때 고객 데이터베이스에 있는 모든 고객에 대하여 다 보내는 것보다는 가능성 있는 고객만 분류해내서 그들에게만 보내는 것이 시간적 경제적으로 절약이 되고 그 남는 여력으로 다른 사업에 투자할 수 있게 된다.

이와 같이 분류규칙은 수많은 데이터를 필요에 맞는 그룹으로 분류해주는 것으로 카달로그 우송, 마켓지점 개설 등에 아주 유용하게 사용될 수 있다.

■ 순서(Sequence)

순서는 시간에 따른 데이터를 효율적으로 처리해서 유용한 정보를 추출하고자 하는 작업이다.

예를 들어 주식(stock) 시장에서 시간에 따른 주식값 변동을 분석해서 이에 대한 규칙을 추출해낼 수 있다면 사용자에게 큰 도움을 줄수 있을 것이다.

이 중에서 분류에 관한 예를 들어 데이터 마이닝의 이해를 돕도록 한다.

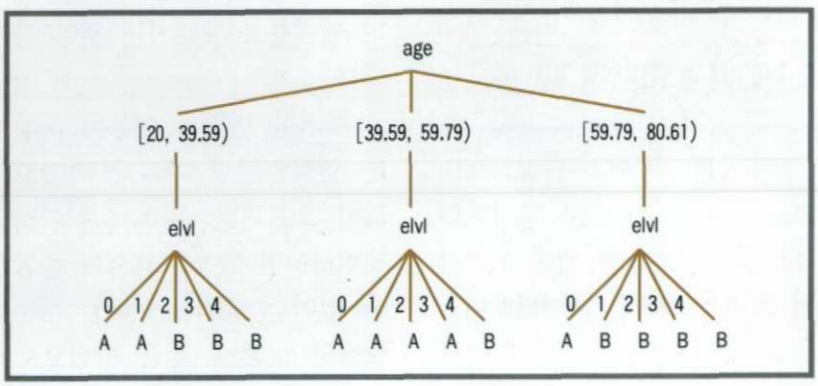
먼저 다음과 같은 세 애트리뷰트로 구성된 데이터베이스를 생각해 보자.

- age : 20-80 사이의 균등하게 분포된 값을 갖는 non-categorical 애트리뷰트이다.
- zip : 9개 zipcode 중에서 값을 갖는 categorical 애트리뷰트이다.
- elvl : 교육의 수준을 나타내는 애트리뷰트로 0-4 사이의 값을 갖는다.

이 데이터베이스에 있는 투플들을 그룹 A, 또는 B로 분류하려고 한다. 이때 먼저 어떤 애트리뷰트를 기준으로 분류하는가 즉 기준 애트리뷰트를 정해야 한다. 기준 애트리뷰트를 정하는 방법에는 information gain ratio 또는 resubstitution error rate 등이 쓰인다.

Information gain ratio를 사용해서 age가 최초의 기준 애트리뷰트로 선정되었고(애트리뷰트 선정에 관한 자세한 내용은 본 고에서는 생략함), 그 다음 기준 애트리뷰트로는 각각 elvl이 선정되었다고 하면 이때의 decision tree는 그림 1과 같이 구성되어진다.

따라서 이 트리로부터 그룹 A와 B를 구하는 다음과 같은



〈그림 1 예제 Decision Tree〉

분류 규칙을 얻을 수 있다.

● 그룹 A

((20 ≤ age 〈 39.6)
∧ (elvl ∈ [0..1])) ∨
((39.6 ≤ age 〈 59.8)
∧ (elvl ∈ [0..3])) ∨
((59.8 ≤ age 〈 80.6)
∧ (elvl=0))

● 그룹 B

((20 ≤ age 〈 39..6)
∧ (elvl ∈ [2..4])) ∨
((39.6 ≤ age 〈 59.8)
∧ (elvl=4)) ∨
((59.8 ≤ age 〈 80.6)
∧ (elvl ∈ [1..4]))

이와 같은 데이터 마이닝은 고도의 응용분야로서 많은 기업들이 연구중에 있다. 그 대표적인 예로서는 미 IBM의 QUEST 프로젝트로서 이 프로젝트에서는 데이터 마이닝 모듈을 개발해서 DBMS에 장착시킬 것을 추진하고 있다.

2. 데이터 웨어하우징

2.1 데이터 웨어하우징의 개념

데이터 웨어하우징(data warehousing)이란 직접질의(direct query)를 위해 동질적인 또는 이질적인 데이터베이스들로부터 데이터를 미리 뽑아서 별도의 물리적인 데이터 웨어하우스에 준비하는 기법이다. 데이터 웨어하우징은 데이터를 미리 준비한다는 의미에서 데이터베이스 통합과 밀접한 관계가 있다.

여러 이질적인 데이터베이스로부터의 지금까지의 데이터 처리 방법은 질의가 들어왔을때 데이터의 통합이 시작된다는 점에서 수동적(passive)이라 할 수 있다.

이에 반해 데이터 웨어하우징 기법은 능동적(active)인 접근방법이라 할 수 있다. 왜냐하면 데이터 웨어하우징에서의 주요 개념은 질의어가 들어오기 이전에 관련된 데이터들을 뽑아 충돌을 해결한 후 통합해서 저장해 놓음으로써 질의어가 들어오면 바로 그 결과를 제공해 주고자 하는 방법이다.

따라서, 데이터 웨어하우징 방법의 핵심 개념은 질의어에 앞서 관련 정보를 추출하고 걸러내고 통합하는 것이다. 질의어가 도착했을때 이 질의어는 실행을 위하여 번역되거나 원래의 테이블로 보내지는 것이 아니다.

이러한 변환과 이동은 복잡한 연산일뿐만 아니라 시간도 많이 소요된다. 따라서 데이터 웨어하우징 방법은 데이터 통합에 있어 능동적(active) 또는 열심인(eager) 접근 방식으로 간주된다.

2.2 데이터 웨어하우징 시스템 구조

그림 2는 데이터 웨어하우징 시스템의 간략화된 구조를 보여준다. 이 그림의 맨 하부는 정보소스 즉 이질 데이터베이스를 보여준다. 이 소스는 스키마변환 및 통합을 통하여 웨어하우스로 모여지게 된다. 변환기(translator)는 소스 데이터를 웨어하우징 시스템에서 사용하는 모델과 포맷으로 바꾸는 역할을 담당한다.

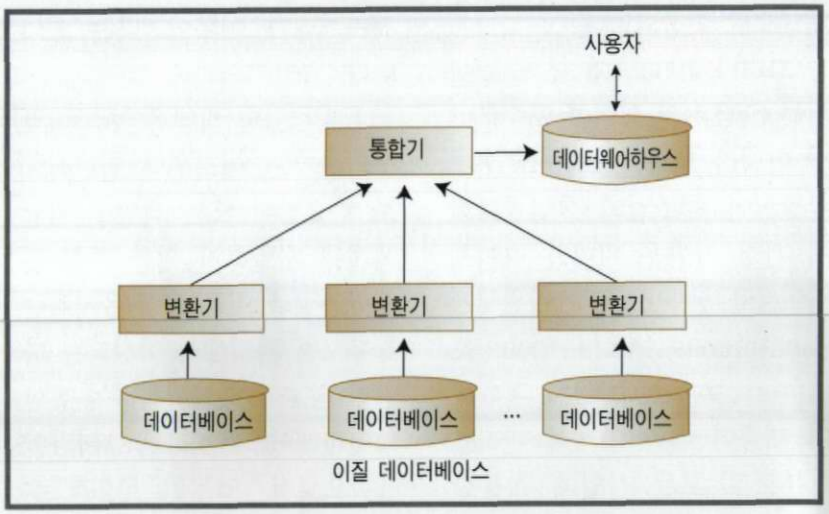
통합기(integrator)는 웨어하우스에 데이터를 설치하는 역할을 담당하는데 이는 데이터를 여과하고 요약하고 여러 소스로부터 취합해서 통합하는 일을 말한다. 이질 데이터를 기존재하는 스키마 속으로 통합시키는 것은 여러 단계를 요구하는 어려운 문제다.

먼저 데이터를 웨어하우스에 의해 사용된 개념스키마로 표현한 후 소스와 웨어하우스 데이터 간에 존재할 수도 있는 불일치를 해결하면서 이미 존재하는 데이터와 통합되어진다. 새로운 소스가 웨어하우징 시스템에 첨가되거나 소스에 있는 연관 데이터가 변경되면 새로운 또는 수정된 데이터가 통합기로 전달된다.

통합기는 또한 변경에 대한 보고를 받아 그 변화를 웨어하우스내로 반영 통합시킨다. 새

로운 데이터를 웨어하우스로 적절히 통합하기 위해서 통합기는 다른 소스로부터 추가적인 정보를 필요로 할 수도 있다.

데이터 웨어하우스 그 자체는 특별 목적의 DBMS를 사용할 수 있다. 그림에는 웨어하우스가 하나이고 중앙집중형이지만 이 웨어하우스는 물론 분산 데이터베이스 시스템으로 구현할 수도 있다.



〈그림 2 데이터 웨어하우징 시스템〉

이러한 데이터 웨어하우징 기법의 장점을 요약하면 다음과 같다.

- 질의어의 수행이 데이터의 변환이나 이동을 수반하지 않으므로 복잡한 질의어도 쉽고 효과적으로 수행될 수 있다.
- 단말 사용자는 단일 모델과 단일 질의어를 사용할 수 있다.
- 시스템 설계가 훨씬 쉬어진다. 예를 들어 기존의 다른 접근 방법에서는 매우 어려운 문제인 이질 데이터베이스간에 질의어 최적화 같은 작업이 필요없게 된다.
- 웨어하우스에 있는 정보는 단말 사용자의 통제아래 있게 된다. 즉 필요한한 안전하고 신뢰성있게 저장되어질 수 있다.

이러한 데이터 웨어하우징 방법의 잠재적 단점은 데이터가 원래의 소스로부터 실제로 물리적으로 복사되어 유지되므로 추가적인 저장공간을 소모한다는 점이다. 그러나 이 점은 최근 저장 매체의 가격이 계속 하락하고 있으므로 그리 심각한 문제는 아니다.

오히려 더 심각한 문제는 데이터의 복사는 원래의 데이터와 불일치를 초래하거나 시대에 뒤떨어진 데이터를 보유하게 될 수도 있다는 점이다. 이러한 이유 때문에 데이터 웨어하우징 기법은 현재의 데이터가 덜 요구되는 분야에서 그 진가를 발휘할 수 있다.

3. CALS 통합데이터베이스

3.1 CALS의 역사와 개념

최근 들어 전산 및 경영 관련 분야에서 CALS에 대한 관심이 높아져 가고 있다. 국방 군수 산업을 중심으로 시작된 CALS는 이제 군수분야뿐만 아니라 각종 제조업 중심의 전 산업에 걸친 산업 정보화로 발전해 가고 있다.

CALS의 역사를 살펴보면 CALS는 1985년 미국방부에서 처음 시작되었다.

이때의 CALS의 의미는 컴퓨터를 이용한 군수지원체계(Computer Aided Logistics Support)를 말하는 것으로 이는 무기체계의 보급 조달과 이의 정비 유지를 위해 디지탈 정보의 통합과 정보의 공유를 통한 신속한 자료처리 환경을 구축하는 전략을 의미하였다. 이후 CALS의 개념은 컴퓨터를 이용한 무기체계 획득 및 군수지원 전산화(Computer-aided Acquisition & Logistics Support)로 바꾸어 무기체계의 군수지원 뿐만 아니라 획득과정을 포함하는 총체적 군수지원

개념으로 확장되었다.

그러나 1990년대 초에 들어서는 CALS의 의미가 군수분야만 아니라 모든 제조업으로까지 확장되어 제조업의 산업정보화에 가장 가까운 의미인 라이프 사이클을 통한 지속적 지원체계(Continuous Acquisition & Life-cycle Support)로 확대되어 제품에 대한 총체적 관리를 기본으로 모든 산업에 적용할 수 있는 개념으로 발전하였다.

이때부터 CALS 개념은 국방 영역을 벗어나 일반 상업적 응용으로 확대되기 시작했다. 이어 정보통신 기술의 급속한 발전에 힘입어 1994년부터는 광속 교역(Commerce At Light Speed)란 개념으로 발전하였다. 이 개념은 각국의 인터넷 사용의 확산과 초고속 정보망 기반 구축이 실용화 단계에 다달음으로서 광속과 같이 빠른 속도의 전자거래가 이루어진다는 의미이다.

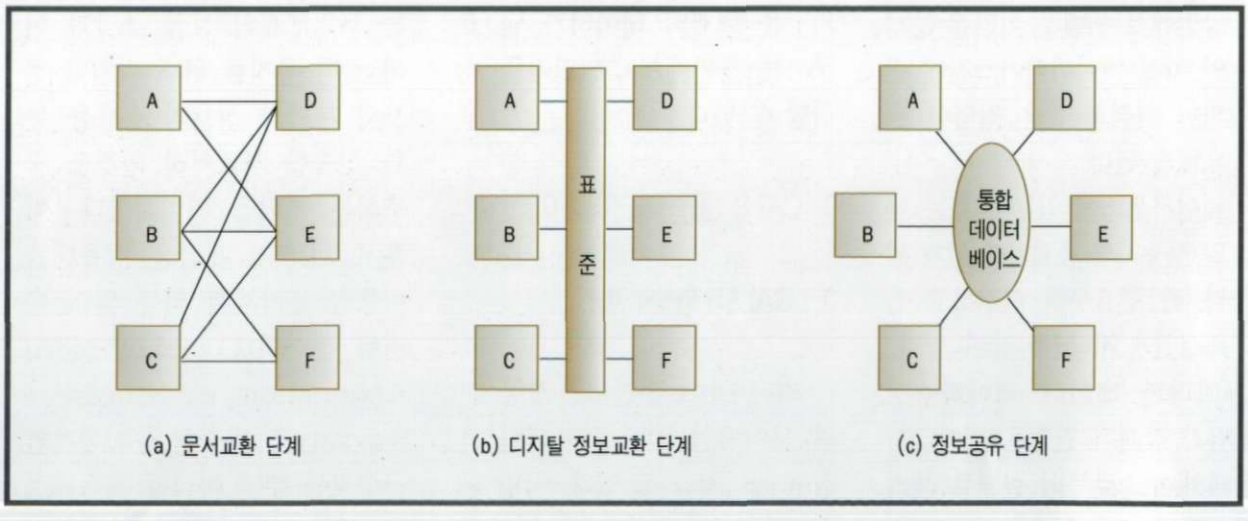
CALS의 개념을 한마디로 정의하기는 어려우나 일반적으로 주요장비 또는 다양한 지원체계를 개발하기 위한 설계 및 제작과정과 이를 운영 유지하는 물류지원 과정에서 작성되는 다양한 정보를 디지탈화하여 통합 데이터베이스(IDB)에 저장 활용함으로서, 업무의 효율적 수행과 신속한 정보 공유 및 전달 체계를 통한 비용절감 효과를 추구하는 전략이라 정의할 수 있다.

따라서 CALS는 기존의 시스템 수명주기 전반에 걸친 문서 중심의 업무처리 방식에서 탈피하여 디지탈 형식으로 상호 정보 공유 및 교환이 이루어질 수 있도록 하기 위한 체계 통합 및 자동화 전략이라고 할 수 있다. 그림 3은 이러한 CALS의 단계별 구축 개념을 보여준다.

즉 CALS는 어떤 시스템이나 사업이 아닌 고도의 경영 철학이자 전략이라고 할 수 있으며, 이는 시스템의 획득 및 운용 지원 과정에서 디지탈 기술정보를 이용하게 하는 자동화된 환경을 제공함으로써 업무의 과학적 효율적 수행, 그리고 정확하고 신속한 정보 공유 및 유통 체계를 통해 시스템 획득 및 운용 비용의 절감, 획득 및 운용지원 시간의 단축, 그리고 종합 품질 경영 능력을 향상시키자는 전략이다. 이러한 CALS의 개념은 최근 미국, 일본, 유럽 등 주요 선진국은 물론 환태평양 지역까지 그 연구가 확산되고 있는 실정이다.

3.2 CALS에서의 IDB

통합데이터베이스(Integra-



〈그림 3 CALS 구축단계〉

ted Data Bases: IDB)는 그림 3에서 보는 바와 같이 CALS 시스템의 핵심 요소로서 기존의 시스템과 제품 개발자의 데이터가 통합된 형태의 데이터베이스이다. 여기서 통합이란 표현은 논리적 통합으로 이는 즉 다양한 형태와 성질의 정보를 어디에서나 투명하게 실시간에 액세스할 수 있다는 의미이지 물리적으로 하나의 컴퓨터나 한 장소에 모아 둔다는 것은 아니다.

CALS의 목표 달성은 궁극적으로 처리 및 액세스가 가능한 데이터를 한번 생산하여 여러번 사용할 수 있는 통합데이터베이스에 달려 있다. 다시 말하면 CALS 전략의 1단계 목표는 디지탈의 흐름이고 최종 목표는 정부와 업체가 통합된 데이터베이스를 구축하고 이를 공유하면서 제품의 전 수명 주기에 활용한다는 것이다.

IDB에 포함되는 정보의 형태는 참고도서 데이터베이스, 사전/법령 데이터베이스, 제품 기술정보, 사건 기록 및 추적체계 데이터베이스, 기술도면 데이터베이스 및 다양한 통계 데이터 등이 있으며 이들 각각의 구성 요소들은 서로 다른 데이터베이스의 데이터를 받아 결합하여 새로운 데이터를 구성할 수 있다.

물론 이때 각각의 데이터베이스들은 서로 독립적이어야 한다.

이러한 IDB 개념이 회사에 적용되면 IPDB(Intrgrated Product Data base)라 불리고 군수 분야에 적용되면 IWSDB (Integrated Weapon System Data base)로 불린다. 최근들어 IPDB는 IPDE (Integrated Product Data Environment)로 발전해가고 있고 IWSDB는 IDSDB(Integrated Defense System Data Base)로 발전해 가고 있다.

IDB가 구축되면 구축된 IDB는 획득, 설계, 생산, 정비, 후속지원, 물자획득, 교육 등의 단계별로 효과적으로 활용되어 통합된 정확한 정보를 제공할 수 있게 된다.

3.3 CALS IDB 시스템 구조

CALS IDB 시스템을 이루는 각각의 데이터베이스는 기존의 문자나 수치뿐만 아니라 그림, 이미지, 그래픽, 텍스트, 소리 등 다매체를 제공하게 되므로 데이터베이스는 다양한 형태의 정보를 보유하게 된다. 이 각각의 정보베이스는 독자적으로 구축되고 독립적으로 운용되는 특성을 갖는다.

CALS IDB 시스템 구조는 먼저 다음과 같은 이질 데이터베이스 통합 기법에 근거하여 접근해 볼 수 있다.

■ Tightly-coupled 방법

이질 데이터베이스 통합을 이루는 한가지 방법은 tightly-coupled 방법이라 하여 전역스키마(통합스키마, global schema)를 사용하는 것이다. 전역스키마를 사용하는 시스템에서는 스키마 변환 및 스키마 통합을 통해 전역스키마를 구성한 다음 이 전역스키마에 해당되는 데이터 조작언어를 이용하여 질의어를 표현한다.

이러한 연구의 대표적인 예로서는 CCA의 MULTIBASE, UNISYS의 Mermaid 등이 있다. 그러나 이 방법은 다양한 형태의 많은 스키마를 통합하는 일이 쉽지 않고 더우기 계속 변화하는 스키마를 통합에 반영하는 것이 어렵다.

■ Loosely-coupled 방법

이질적인 데이터베이스들을 통합하는 또 하나의 다른 방법은 loosely-coupled 방법이라 하여 전역스키마를 사용하지 않는 대신 강력한 데이터베이스 조작 언어를 이용하는 것이다. 이 언어를 Litwin은 멀티데이터베이스 언어(multidatabase language)라 불렀다.

대표적인 예로서는 휴스톤 대학의 Omnibase, DIRECT 등이 있다. 이 방법은 위치 투명이 제공되지는 않으나 각 정보베이스의 독자성을 최대한 보장

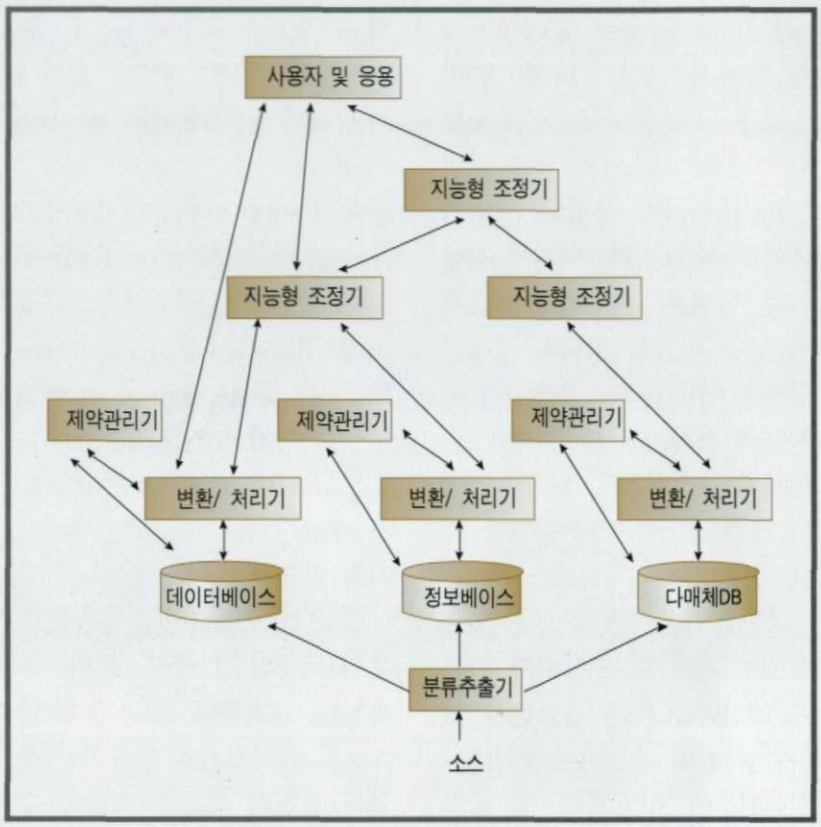
해줄 수 있으며 시스템의 확장 및 변경에 능동적 대처가 가능하다.

■ 데이터 웨어하우징 방법

웨어하우징 방법의 핵심 개념은 질의어에 앞서 관련 정보를 추출하고 걸러내고 모아 놓는 것이다. 질의어가 도착했을 때 이 질의어는 실행을 위하여 번역되거나 원래의 테이블로 보내지는 것이 아니다. 이러한 변환과 이동은 복잡한 연산일뿐만 아니라 시간도 많이 소요된다. 그러나 자주 변하는 정보베이스에 있어서는 그 유지 관리가 곤란하다.

과거의 연구를 보면 tightly-coupled 방법에 가까운 JCALS의 GDMS(Global Data Management System)과 IWSDB를 이용한 방법, 데이터 웨어하우징 방법에 가까운 RGS社의 Smart Agents를 이용한 방법 등이 있다.

그림 4는 loosely-coupled 방법에 기초한 IDB 시스템 구조를 보여준다. 각 구성 요소를 간략히 설명하면 다음과 같다.



〈그림 4 CALS IDB 시스템 구조〉

● 데이터/ 정보베이스(Data/ Information Base): 데이터베이스를 근간으로 하는 각종 정형 또는 비정형 정보의 저장소이다. 특히 최근 들어서는 인터넷의 각종 정보를 온라인 데이터베이스로 하여 서비스할 수 있는 수준으로 발전될 것을 요구하고 있다.

● 변환기/ 처리기(Translator/ Processor): 데이터베이스의 스키마를 관리하고 사용자 요구에 대한 검색등의 처리를 담당한다. 예를들어, 하부의 데이터가 정형의 관계형 데이터베이스로 구축되어 있다면 이때의 처리기는 RDBMS엔진으로 볼 수 있다.

● 분류추출기(Classifier/ Extractor): 비구조화된 정보 소스들을 여러 매체별로 분류하고 이들로부터 제약 관리기나 변환기에서 사용될 규칙들을 추출해 낸다.

● 제약관리기(Constraint Manager): 저장된 여러 이질 정보들이 시맨틱상으로 일관성 있게 유지 관리될 수 있도록 해준다. 이러한 제약은 통상 규칙(rule)의 형태로 유지된다.

● 지능형 조정기(Intelligent Controller): 지능형 조정기는 사용자나 응용 프로그램이 원하는 정보를 얻을 수 있도록 해주는 IDB 시스템 구조의 핵심 부품이다. 지능형 조정기는 지식 베이스로서 각

지능형 조정기가 관할하는 정보베이스에 대한 구조에 대한 정보를 가지고 있다가 사용자의 요구시 변환기를 통하여 정보베이스들을 보여주고 더 나아가 즉시 해당 정보 베이스들로부터 정보를 가져와 이를 필터링하고 종합 및 융합하여 사용자에게 제공하는 역할을 담당한다.

이와 같은 구조의 IDB는 기존의 데이터베이스 통합 시스템과 비교해 볼때 다음의 장점을 갖는다.

- 시스템의 확장 즉 정보베이스의 확장에 매우 능동적으로 대처해 나갈 수 있다. 왜냐하면 해당되는 지능형 조정기만 준비되면 시스템 확장이 이루어지기 때문이다.
- 지능형 조정기는 계층구조를 가질 수 있으므로 앞으로 지식베이스 기법을 잘 응용하면 매우 효율적인 사용자 인터페이스를 제공할 수 있다.

이와같이 CALS IDB를 구축하기 위해서는 공통 모델의 선정 및 변환 기술, 언어 번역 기술, 정보의 편집 및 저장 기술, 정보 필터링 기술, 제약조건 명세기술, 분산 트랜잭션 기술, 지식 기반 사용자 인터페이스 기술, 그리고 다매체 정보융합 기술 등 다양하고 복합적인 최신 기술들이 요구된다.

4. 결론

데이터베이스는 정보산업의 발전에 따라 그 중요성이 점점 증가되고 있는 분야로써 최근 들어서는 그 응용분야가 날로 확장되어가고 있다.

본 고에서는 데이터베이스의 새로운 응용분야로 인식되어 그에 대한 관심이 고조되고 있는 응용분야인 데이터 마이닝, 데이터 웨어하우징, CALS IDB 등에 관하여 그 개념을 위주로 간단히 살펴보았다. 이러한 기술들은 개별적으로도 적용될 수 있지만 서로 연관이 될 수도 있다. 예를 들면 데이터 웨어하우징을 구축한 후 데이터 마이닝을 실시하면 더 좋은 효과를 얻을 것이고, 또한 데이터 웨어하우징을 CALS에 응용하면 좋은 효과를 얻을 수 있을 것으로 기대된다.

이러한 새로운 응용분야는 인터넷의 발달과 초고속 정보통신망의 활성화와 더불어 우리 산업 전반에 걸쳐 영향을 미치게 될 것이므로 이에 관한 연구 및 개발은 국가 경쟁력 강화를 위해서 또한 정보선진국과의 경쟁을 위해서 필수적일 것으로 생각된다. DC