실체 뷰 기법을 이용한 대규모 전자 카탈로그 시스템

A Large Electric Catalog System using Materialized Views

권일혁*, 최현화*, 김수현**, 문강식*, 이동하***, 이전영*

*포항공과대학교 컴퓨터공학과
**포항공과대학교 정보통신학과
***포항공과대학교 정보통신연구소

{ihkwon, ponoma, kimbell, ksmoon, dongha, jeon}@postech.ac.kr

대규모 전자 카탈로그 시스템은 통합 데이터베이스 기술의 대표적인 응용분야로서 최근 전자상거래 분야의 발달로 그 수요가 증대되고 있고 있다. 본 연구에서는 수 천개 이상의 공급자와 수요자를 고려한 대규모 전자 카탈로그 시스템용 통합 데이터베이스의 구조와 관리 기법들에 대해 기술한다. 통합 데이터베이스의 구축은 스키마 통합과 데이터 통합의 두 단계로 구성된다. 스키마 통합은 지역 스키마를 주어진 전역 스키마로 변환하며, 개별 지역 데이터베이스의 스키마와 표현 능력이 다른 것을 고려했다. 데이터 통합의 경우에는 실체 뷰 기법을 사용하며, 각 데이터의 식별자 처리와 함께, 지역 데이터베이스에서 갱신된 부분만을 가져오는 기법을 구현했다.

1. 서론

대규모 전자 카탈로그 시스템은 통합 데이터베이스(integrated database : IDB)기술의 대표적인 응용분야로서 최근 전자상거래(electric commerce : EC) 분야의 발달로 그 수요가 증대되고 있고 있다. 현재 운영되고 있는 전자 카탈로그 시스템에서 개개의 물품에 대한 관리는 대부분 수동적으로 이루어지며, 이에 대한 특별한 방법이 없었다. 각각의 물품 정보를 공급자가 직접 입력하거나, 배치(batch) 작업을 위한 프로그램을 통해 입력하기 때문에, 물품정보의 수정을 위해서는 물품마다 개별 처리를 할 수 밖에 없었다.

특히 대규모의 통합 데이터베이스에 대해서는 상대적으로 적은 수의 지역 데이터베이스(local database) 수를 고려해왔던 기존의 분산 데이터베이스의 연구 결과가 반영되기 어려웠었다. 최근 B2B(business-to-business)나 B2C(business-to-customer) 관련 사업이 활성화되면서, 특정 분야에 대한 물품 코드(code)의 통합과 분류 체계의 표준화가 진행중인데, 이러한 환경의 변화

는 통합 데이터베이스 구축에 대한 다른 관점의 접근이 가능하게 되었다.

본 연구에서는 수 천개 이상의 공급자와 수요자를 고려한 대규모 전자 카탈로그 시스템용 통합 데이터베이스의 구조와 관리 기법들에 대해 기술한다. 통합 데이터베이스는 데이터의 통합뿐만 아니라, 관리도 자동화한다. 예를 들어, 공급사에서 제공하는 제품이 추가되었을 경우나, 해당 물품에 대한 내용이 수정되었을 경우, 공급자 측에서 정보를 통합 데이터베이스의 내용을 수정하는 것이 아니라, 통합 데이터베이스에서 각 공급사의 지역 데이터베이스에 접근해서 데이터를 수집하고, 적당한 변환을 통해 통합 데이터베이스의 상태를 최신 정보로 유지하는 것이다.

이러한 통합 데이터베이스 구축 방법은 두 가지 전제를 가정하고 있다. 첫째는 해당 전자 카탈로그 분야에 물품 코드와 물품 분류 체계가 표준화 되어 있거나, 카탈로그 참여자가 해당 표준에 대해 동의할 것이다. 소스 데이터베이스가 이러한 표준을 따를 필요는 없지만, 표준 코드와의 관련성이 표현된 정보가 제공되어야 한다. 둘째는 전자 카탈로그에 참여하는 공급자들의 지역 데이터베이스가 JDBC(java database connectivity)와 같은 외부 접속이 가능한 DBMS를 사용할 것이다. 지역 데이터베이스를 단순한 파일로 관리하거나, 통합 데이터베이스 관리 프로그램에서 지역 데이터베이스로의 적절한 방법으로의 접근이 불가능하다면, 제안되는 자동화 기법의 적용이 불가능하다.

통합 데이터베이스의 구축은 스키마 통합과 데이터 통합의 두 단계로 볼 수 있다. 스키마 통합은 지역 스키마를 주어진 전역 스키마로 변환하는 방식으로 이루어진다. 전역 스키마는 표준 코드와 표준 분류 체계를 포함하고 있다. 지역 데이터베이스의 질의 처리 능력이 각각 다르고, 통합시의 안정적인 처리를 위해, 속성 일치와 값 일치를 별개로 다루었다. 속성 일치와 값 일치에 대해서는 뒤에 다룬다.

데이터 통합의 경우에는 지역 데이터를 모아서 통합 데이터베이스에 저장하는 실체뷰(materialized view) 기법을 사용했다. 이때 각 데이터의 식별자(identifier) 처리와 함께, 지역 데이터베이스에서 갱신된 부분만을 가져오는 방법을 고려했다.

이와 함께, 시스템 성능과 사용자의 편의성 등 대규모 통합 데이터베이스와 전자 카탈로그 시스템의 구현에 따른 여러가지 문제들도 고려했다.

2. 전자 카탈로그 시스템의 요구 사항과 설계

2.1 요구사항

인터넷의 대중화와 더불어 전자 상거래의 부흥을 맞고 있는 현실 상황에서 상거래에 있어 소비자들의 제품 확인 욕구에 적절히 대응하기 위한 하나의 방편으로 나타난 전자 카탈로그는 전자 상거래를 위하여 물품 및 서비스에 대한 정보를 전자적인 형태로 저장하여 교환하기 위한 전자 문서이다. 여기에는 물품에 대한 간략한 소개, 동화상, 정지화상, 제작업체 URL, 연락처, 주문서 및 기업에 대한 기타 안내 등이 나오며, 기존의 인쇄물 형태의 카탈로그에 비해, 비교적 많은 정보를 담고 있고, 같은 종류의 물품을 여러 회사에서 생산하는 경우, 정보를 통합하여 같은 구조로 보여주고 있다.

그림 1은 전자 카탈로그 시스템의 일반적인 구성을 보여 준다.

<그림 1 전자 카탈로그 시스템 구성>

넓은 의미의 전자 카탈로그 시스템은 물품 자체에 대한 정보뿐만 아니라, 물품 관련 정보 및 거래 정보, 배송 정보, 업체 관련 정보, 소비자 보호 관련 정보들이 모두 필요하겠지만 일반적으로 거래가 이루어질 때 필요한 처리는 이마켓플레이스(e-market place) 시스템이라는 별도의 시스템을 구성하는 것이 일반적인다. 본 논문에서는 대용량 데이터의 통합을 위주로 다루기 때문에, 물품 관련 정보의 통합을 중점으로 설명한다. 표 1에 대표적인 물품 관련 정보를 보였다.

구분	내용
거래 데이터	단품 가격, 가격표시통화, 세금정보, 할인 또는 별도 요금, 가격책정 날짜, 가격책정수량, 배송조건
기본 데이터	단품코드, 제품설명, 제품그룹/제품군, 제품생산일, 특정조건, 제품규격, 제품수량, 취급조건(사항), 제품포장, 원산지,(국가), 제품상관관계, 기타
추가 데이터	위험물 관련 데이터, 제품특징, 제품관련 기술적 규정

<표 1 EAN International의 전자 카탈로그 구성요소>

2.2 설계

2.2.1 전자 카탈로그 시스템 설계

전자 카탈로그 시스템의 사용자는 구매자와 판매자의 두 부류로 나뉠 수 있는데, 이들이 보는 관점은 상당히 다르다. 일반적으로 이 두 사용자에 대해 다른 인터페이스를 제공한다.

구매자는 물품들의 리스트를 카테고리 별로 볼 수 있게 했으며, 각 카테고리별로 하위 카테고리를 둔다. 보고 싶은 특정 물품 리스트를 보여주기 위한 적절한 search 기능을 전체 물품에 대해서 수행하거나, 현제 보고 있는 카테고리의 하위에 대해서만 수행하게 했다.

판매자의 관점에서는 자신이 공급하는 물품 중심의 카테고리 표시를 기본으로 했으며, 자동화된 지역 데이터베이스에서의 정보 수집 방법에 대한 인터페이스가 있어야 한다. 자동으로 수집되고 관리되는 통합 데이터베이스에는 여러가지 이유로 인해, 해당 표준 카테고리가 없는 물품이 수집되었거나, 통합된 상태에서 정보를 수정할 필요가 있을 수 있다. 이러한 관리에 필요한 인터페이스도 같이 제공된다.

2.2.2 전자 카탈로그 시스템을 위한 통합 데이터베이스 시스템 설계

전자 카탈로그 시스템을 위한 통합 데이터베이스의 구성에 있어서 필요한 구성 요소들은 다음과 같다.

(1) 전자 카탈로그를 위한 전역 데이터베이스

각각의 로컬 데이터 소스(source)들로부터 가져온 데이터들이 변환을 거쳐 하나의 통합된 데이터베이스로 재구성되기 위한 스키마가 필요하다. 물품 코드와 물품 분류는 표준화 과정을 거쳐

서 얻어지고, 그외의 스키마는 응용 환경에 따라 재구성되어 질 것이다.

(2) 소스 데이터베이스

전역 스키마에 참여할 지역 데이터베이스이다. 각 로컬의 데이터를 재구성하여 통합에 필요한 정보들인 소스 데이터의 위치, 접속 정보 등이 관리되어야 한다.

(3) 소스 데이터베이스에서 전역 데이터베이스로의 통합 관리

소스 데이터베이스에서 가져온 데이터를 어떻게 전역 스키마에 맞게 재구성 혹은 변환 할 것인지에 대한 전반적인 처리를 하는 부분이다. 주어진 전역 데이터베이스의 스키마에 소스 데이터베이스의 스키마를 변환하는 정보가 관리되어야 한다. 이때 필요한 정보는 소스 데이터베이스 관리자가 제공하게 되고, 이를 쉽게 하기 위한 인터페이스가 제공되어야 한다.

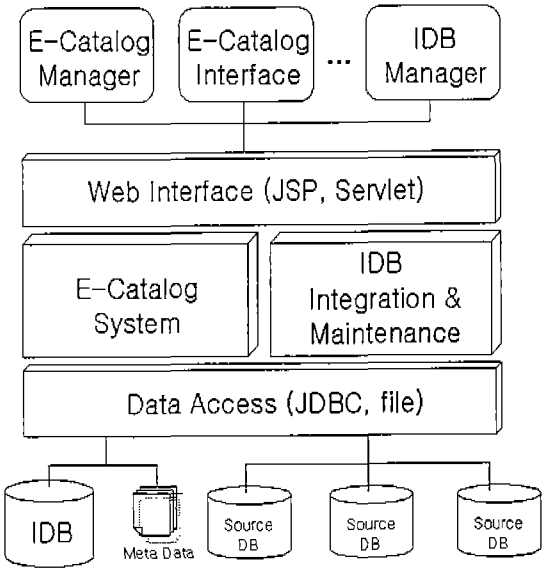
(4) 소스 데이터베이스 접속 관리

소스 데이터베이스의 등록, 자동화된 데이터 수집의 스케줄 관리, 데이터 변환 통계 처리 등이 필요하다.

위의 구성 요소들을 수렴한 전자 카탈로그 시스템을 위한 통합 데이터베이스의 전반적인 구성도가 <그림 2>이다.

3. 통합 데이터 베이스의 관리

통합 데이터베이스를 유지하고 관리하는 기법으로 우리는 실체 뷰 유지 기법을 선택했다. 대규모의 전자 카탈로그를 통합하기 때문에, 검색 질의를 처리하기 위해, 다양한 속도 차이가 있는 소스 시스템을 직접 접근하기에는 무리가 따른다.



<그림 2 IDB System의 전체 구성도>

그리고, 구현된 시스템은 실체 뷰를 직접 통합 데이터베이스에 구현하지 않고, 소스 데이터베이스의 이미지(image)를 먼저 생성한 뒤에 통합하는 과정을 거친다. 통합 데이터베이스를 직접 update 하게 되는 경우, 운영중인 전자 카탈로그의 성능에 영향을 줄 수 있을 뿐만 아니라, 관리상의 어려움이 있기 때문이다. 특히 객체 확인 과정에서 운영중인 시스템에 과부하를 줄 수 있다. 결국, 소스 시스템 또는 네트웍과 통합 시스템 양쪽에 과부하를 막기 위해 두 단계의 변환 과정을 거친다.

3.1 객체 확인

통합 데이터베이스 구성에서 객체 확인 과정이 필요한 이유는 지역 데이터베이스로부터 가져온 데이터가 이미 통합 데이터베이스에 존재하는 지를 확인해야 하기 때문이다. 개발된 시스템은 통합 과정만을 자동화한 것 뿐 아니라, 수정 등의 관리 과정도 자동화 했기 때문에, 소스에서 가져온 하나의 물품 정보에 대해, 통합 데이터베이스의 내용을 갱신(update)할 것인지, 해당 물품을 신규로 삽입할 것인지 결정할 필요가 있

다.

객체 확인을 위한 전역 객체 식별자는 로컬 데이터 소스의 식별자와 지역 데이터베이스내의 객체 식별자를 연결(concatenate)하고, 이를 다시 4 바이트(byte)의 식별 코드(identifying code)로 변환해서 사용한다. 4 바이트의 자체 식별 코드로 변환하는 이유는 최종적으로 얻어지는 전역 식별자가 너무 길어지는 것을 방지하기 위해서이다.

실제 전역 데이터베이스로 삽입이 일어나는 경우는 이렇게 생성된 전역 식별자가 새로운 식별자인 경우이고, 갱신하는 경우는 해당 전역 식별자의 정보에 소스에서 가져온 최종 수정시간(last update) 값을 비교해서 갱신을 수행하게 된다.

3.2 물품 분류 체계 관리

위에서 기술한 객체 확인 과정에서는 표준화된 물품 코드나, 물품 분류 체계에 대한 고려가 없다. 소스 데이터베이스를 운영하는 공급자들이 자신의 시스템에서 표준 물품 코드나, 표준 분류 체계를 운영한다는 보장이 없기 때문이다.

그러므로 소스 데이터베이스의 물품 코드는 객체 확인 과정에서 활용되지만, 소스 데이터베이스의 분류 체계는 통합 데이터베이스에서는 전혀 의미를 가지지 않는다. 통합 데이터베이스에는 표준화된 물품 코드와 표준화된 분류 체계를 따르고 있고, 물품 구매자에게는 이 분류 체계로 인터페이스가 제공된다.

소스 데이터베이스에서 가져온 물품을 표준화된 분류로 체계화할 때 필요한 정보는 소스 데이터베이스의 특정 물품에 해당하는 표준 물품 코드이다. 일반적인 경우, 이 물품 코드는 전역 데이터베이스 내의 최종 분류에 해당한다.

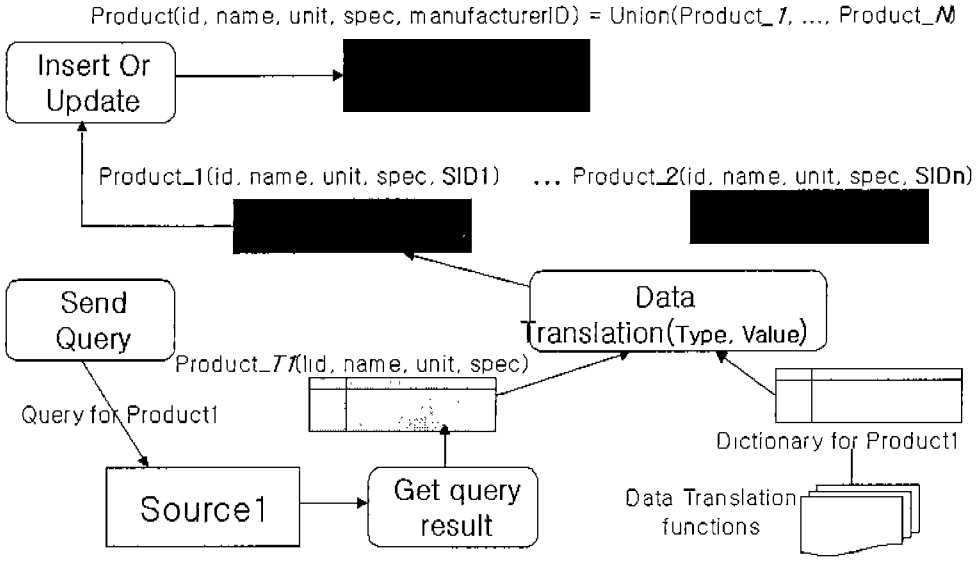
소스 데이터베이스에는 해당 물품에 해당하는 표준 물품 코드를 반드시 제공해야 하고, 전역 데이터베이스에서 표준 물품 코드는 최하위 분류의 역할을 하기 때문에 하나의 표준 물품을 제공하는 여러 업체들의 물품 정보가 최하위 분류에 동시에 나타나게 된다.

3.3 실체 뷰 관리

실제적인 데이터 통합에 있어서 본 시스템은 2-스키마 구조(2-schema architecture)를 사용하였다. 하나는 전역 스키마로서 실체화를 위한 전자 카탈로그의 스키마이고, 또 다른 하나는 로컬 스키마로서 로컬 데이터의 뷰 집합으로 로컬 데이터베이스의 이미지(image) 데이터베이스의 스키마 집합이다. 로컬 데이터베이스의 이미지에 해당하는 스키마를 로컬 뷰(local view)라고 정의한다.

또 일반적인 전자카탈로그의 특성상 통합되어야 할 테이블은 상품 테이블 하나이거나, 소수에 그칠 것이므로, 이후의 설명에서는 통합 데이터베이스를 통합 테이블로 지칭한다.

통합 테이블 구축의 절차를 보면 먼저 데이터를 가지고 올 소스를 등록하게 된다. 그리고 등록된 소스로부터 데이터를 어떻게 가져 올 지에 대한 정보를 담은 로컬 뷰를 등록하게 된다. 그리고 가져온 데이터에 대한 변환 정보를 가지는 데이터 사전(data dictionary)을 등록하게 된다. 로컬 뷰는 정해진 스케줄에 따라, 소스로부터 데이터를 수집해서 정보가 채워지게 되며, 후에 로컬 뷰의 결과를 데이터 사전을 참조해서 전역 테이블에 삽입하거나 수정 작업을 함으로써



<그림 3 데이터 통합 과정>

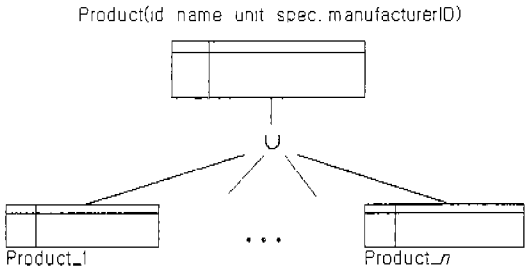
통합 과정을 마치게 된다.

통합 과정에서 고려되어야 할 사항으로는 데이터 소스의 접근 횟수에 대한 문제와 데이터의 삽입과 수정에 대한 문제 그리고 데이터 사전의 설계 즉 변환 함수를 어떻게 기술 할 것이며, 필요한 변환 함수를 어떻게 적절히 호출 할 것 인가의 문제들이 있다.

소스 데이터 접근 횟수에 대하여 기본적으로 본 시스템에서는 한시간 단위로 모든 트랜젝션이 수행된다. 각각의 소스 데이터에 따라서 고유의 스케줄링 관련 정보를 가지고 있고, 그 스케줄링 정보는 일 단위, 주 단위, 월 단위의 골격을 가지고 있다.

예를 들자면, 매일 몇 시, 혹은 매주 무슨 요일 몇 시, 매달 몇째 주 무슨 요일 몇 시, 혹은 매달 몇 일 몇 시 등의 정보를 스케줄러가 항시 체크 하여 일정 시간이 되면 소스데이터를 가지고 와서 통합 데이터에 변환하여 넣는 식의 일을 하게 된다.

<그림 3>는 로컬 소스로부터 가져온 데이터가 로컬 소스의 이미지인 임시 테이블에 저장되고 난 후 전역 테이블로 통합되는 과정을 간략하게 보여 주고 있고, <그림 4>에 전역 스키마인 Product가 각각의 소스 테이블들인 Product_1, Product_2... 등의 합집합(union)으로 구성되는 모습을 도식적으로 보여주고 있다. 실제로는 소스 데이터베이스 Source1으로 정해진 질의를 수행해서 임시 테이블인 Product_T1을 얻고, 이를 통합하여 Product를 얻는 것이다. 임시 테이블과 통합된 테이블은 전자 카탈로그 서버에 저장된다.



<그림 4 테이블 통합>

임시 테이블은 실제 로컬 테이블과 스키마가 일치하지 않는다. 로컬 테이블에 적절한

질의를 수행해서 얻어진 임시 테이블에 저장된 로컬 데이터들로부터 어떻게 전역 스키마에 맞는 데이터로 변환 되는지에 대해서 살펴 보도록 한다. 로컬 소스에서 지원하는 질의 처리기가 강력한 경우, 별도의 변환 없이, 질의만으로도 전역 테이블과 완전히 동일한 스키마를 얻을 수 있고, 이런 경우에는 별도의 데이터 변환과 데이터 사전이 필요없다. 그렇지 않은 경우, 각각의 속성에 대해 적당한 변환 과정이 필요하다. 모든 경우에, 로컬 소스에 행해지는 질의는 통합 테이블을 구성하기 위한 충분한 정보를 가져오는 것이 주 목적이 된다. 즉 질의를 수행하는 단계에서는 그 속성에 해당하는 정보가 있는지의 여부만 보장되면 된다. 이 과정을 속성 일치라고 하는데, 실제 가져온 값이 전역 데이터베이스 내의 값과는 의미가 다른 것에 대한 처리는 후에 데이터 사전을 통해 처리하는 것이다.

첫번째로 전역 스키마의 속성과 같은 경우 당연히 데이터의 변환 과정이 필요 없게 된다. 로컬 스키마의 속성값이 그대로 전역 스키마의 값이 된다.

이 경우에 처리되는 식은 다음과 같다.

Product_3.name = Product_T3.name

두 번째는 로컬 데이터의 값이 전역 데이터의 특정 값으로 변환되는 경우이다. 즉 특정 도메인 1 로부터 특정 도메인 2로의 변환이다. 예를 들면 로컬 데이터 소스에서 국가정보를 문자열(Korea, America, Japan, ...)으로 지정했는데, 전역 스키마에는 각각의 값을 특정 상수로 쓴다면 <표 2>와 같은 값의 변환이 있어야 한다. 이러한 값의 변환을 수행하는 것을 값 일치 처리(value mapping)이라고 한다.

Product_2.attr	Product_T2.attr
1	Korea
2	America
3	Japan

<표 2 value mapping table>

세 번째는 문자열 연산을 들 수가 있다. 예를 들자면 로컬 데이터 소스에서 특정 문자열들의 연결로서 전역 데이터를 만드는 경우이다.

Product_4.id = Product_T4.id + SID4

마지막으로 로컬 데이터의 값에 수치적인 연산이 있을 수가 있다. 예를 들면 환율의 차이에 의한 가격의 변환 등에서 처럼 가격에 특정 상수를 곱하는 연산을 볼 수 있다.

*Product_5.price = Product_T5.price * 1326.55*

그리고 데이터 사전에서의 함수 호출은 <표 3>와 같은 형태로 불려지게 된다.

4. 구현 환경과 결론

본 연구에서는 수 천개 이상의 공급자와 수요자를 고려한 대규모 전자 카탈로그 시스템용 통합 데이터베이스의 구조와 관리 기법들에 대해 기술하였다. 그리고 통합 데이

Product_N.id =
trans_id(Product_TN.id, SID_N)
Product_N.name =
trans_name(Product_TN.name)
Product_N.unit =
trans_unit(Product_TN.unit)
Product_N.spec =
trans_spec(Product_TN.spec)
Product_N.manufacturerID = SID_N

<그림 5 데이터 사전 함수 호출>

터베이스의 구축은 스키마 통합과 데이터 통합의 두 단계로 나누었다. 이 두 단계에서 속성 일치와 값 일치를 별개로 보았으며, 지역 데이터를 모아서 통합 데이터베이스에 저장하는 실체 뷰 기법을 사용했다.

본 논문에서 제안한 시스템은 IDB platform으로 Oracle 8i을 통합 데이터베이스로 활용해 구현했으며 로컬 데이터베이스들은 mySQL, MS-SQL, Oracle 8i에 설치하여 수행하였다. 그리고 모든 데이터베이스 연결 인터페이스로는 JDBC를 사용하였다. 마지막으로 전자 카탈로그시스템이 기본적으로 구동하는 환경은 Oracle 9i AS(application server)이다. 개발이 완료된 본 시스템은 수개의 로컬 시스템에서 실험했고, 열개 이상의 다수의 로컬 데이터베이스에 실험하지는 못했지만, 상용 서비스 준비중이다.

대규모의 통합 데이터를 관리하기 위해서는 기존의 분산 데이터베이스 연구에서 수행되었던 방법들이 적절하지 않다. 특히 전역 스키마의 관리 부담이 큰 문제 였다. 본 연구와 같이, 지역 데이터베이스 관리자의 참여를 통해 안정적인 통합 데이터베이스를 구축하는 것이 현실적으로 가장 적절한 방법이라고 생각한다.

구축된 시스템은 현재 Agent 기반 시스템으로도 볼 수 있으나, 지역 DB를 수집하는 과정에 좀더 지능적인 서비스가 통합되어, 분산 에이전트 시스템으로 발전시킬 수 있을 것이다.

참고문헌

[1] R. Hull and G.Zhou. A framework for supporting data integration using the materialized and virtual approaches. In *Proc. ACM SIGMOD*, 1996.

[2] Patrick Martin and Wendy Powley. Database Integration using Multidatabase Views.

[3] Steve Olson. Distributed Query Processing Using Sybase Adaptive Server Enterprise and OmniConnect 11.9.2

[4] D. Quass and J. Widom, "On-line warehouse view Maintenance for Batch Updatess," *Proc. of the ACM SIGMOD*, pp. 393-404, May, 1997.

[5] A. P. Sheth and J. A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing surbeys*, 22(3):183-236, 1990.

[6] Jeffrey D. Ullman. Information Integration Using Logical Views. In *Proc. Int. Conf. On Database Theory*, 1997.

[7] 한국 전산원. 전자 카탈로그 관련 기술 및 사업의 현황 분석과 개선방안.

[8] 한국정보통신 기술협회. 전자 상거래에서 물품 정보교환을 위한 전자 카탈로그 공통 표준.