

XML Schema 기반 정보 통합 시스템 설계 : DataBlender

Design of an Information Integration System based on XML Schema : DataBlender

이미영
한국전자통신연구원 데이터베이스연구팀 팀장
김명준
한국전자통신연구원 콘텐츠기술연구부 부장
이규철
충남대학교 컴퓨터공학과 교수

Mi-Young Lee
Team Leader, Contents Archiving Team, ETRI
Myung-Joon Kim
Director, Contents Technology Department, ETRI
Kyu-Chul Lee
Prof., Dept. of Computer Engineering, Chungnam National Univ.

중심어 : 정보 통합, 미디어이터, 래퍼, XML 모델

요 약

이질적이며, 분산되어 있는 정보 저장소들을 통합하여 가상 데이터베이스를 제공하는 미디어이터 기반 정보 통합 기술에서는 통합되는 정보의 의미 보존을 제공하는 공통 자료 모델과 이질 자료 모델 통합시 발생하는 충돌 해결 방법이 중요하므로 이에 대한 다양한 연구가 수행되고 있다. 본 논문에서는 XML Schema, XQuery 등 주요 XML 기술을 활용하여 통합되는 정보의 의미 손실을 최소화하고, 통합 자료 모델들간의 충돌 해결 표현이 용이하도록 설계한 DataBlender 시스템에 대하여 제안한다. 또한 DataBlender 시스템은 모바일 환경을 고려하여 이동 단말기에서도 통합 검색 서비스를 활용할 수 있도록 설계함으로써 차세대 인터넷 비즈니스 환경 구축의 기반이 된다.

Abstract

In a mediator based information integration system that integrates information distributed on various data sources as a common data model, there are many researches on a common data model providing a reservation of integrated data semantic, and on a resolution for a conflict caused by integrating various data models. This paper proposes the DataBlender system that minimizes the semantic losses and provides an easy resolution for a conflict using the XML technology like as XML Schema, XQuery. Futhermore, it provides an integrated query facility usable in the mobile environment. So it becomes a foundation for a next Internet business environment.

I. 서론

정보사회가 급속히 발전함에 따라 기업의 경쟁력 확보에 있어 가장 중요한 것은 중요한 정보에 대한 빠르고 용이한 접근이다. 기업들이 광역화됨에 따라 분산 저장 관리되고 있는 정보에 대한 접근이 요구되고 있으며, 자사에서 관리하는 정보뿐만 아니라 다른 기업에서 관리하는 정보의 활용도 요구되고 있다. 예를 들어 전자 상거래, 포탈 서비스 구축 등을 위해서는 관련 업체들과의 정보 공유, 혹은 이미 구축된 고객 정보, 상품 정보, 공급자 정보 등을 통합하여야 할 필요가 생기고 있으며, 이와 같은 정보는 하드웨어, 운영 체제, 자료 관리 시스템 그리고 네트워크 환경 등 다양한 환경에서 관리되는 경우가 많다. 또한 무선 컴퓨팅 환경의 발전에 따라 이동 단말기에서도 서비스를 지원해야 할 필요성이 가속화되고 있다.

다양한 환경에서 관리되고 있는 여러 정보들을 통합하여 새로운 서비스 구축을 용이하게 하기 위한 정보 통합 기술로는 데이터 웨어하우스 구축 기술, 멀티 데이터베이스 시스템 기술, 미디어이터 기반 통합 기술 등이 연구되어 왔다. 미디어이터 기반 정보 통합 기술에서는 XML 모델을 통합 모델로 이용하는 연구가 활발히 진행되고 있으며, 데이터 타입 등 정보 표현력이 미흡한 DTD 기반 XML 통합 모델에서 최근에는 XML Schema 활용으로 이전되고 있다. 통합 모델을 기반으로 하는 정보 통합에서는 이질 자료 모델 통합시 발생하는 충돌에 대한 분류, 이에 따른 해결 방법에 대한 연구가 필요하며, 자동으로 통합 스키마를 추론 구축하는 기술, GUI 기반의 통합 모델 구축 도구, 통합 스키마 정의 기술 등에 대한 연구가 이루어지고 있다.

본 논문에서는 통합되는 정보의 의미 손실을 최소화하고, 통합 자료 모델간의 충돌 해결 표현이 용이하도록 통합 스

키마 정의를 제공하는 미디어이터 기반 데이터베이스 통합 미들웨어인 DataBlender 시스템을 제안한다. DataBlender는 자료 모델 표현력이 풍부한 XML Schema 기반 XML 모델을 지원하고, 이질 자료 모델간의 충돌 해결을 표현할 수 있는 다양한 요소를 갖고 있는 XQuery 기반 통합 스키마 정의를 제공하고, 유무선 통합 인터넷 환경에서 통합 검색 서비스를 활용할 수 있도록 이동 단말기에서도 사용 가능하다. 2장에서 기존 연구 내용과의 차별성에 대해 소개하고, 3장에서는 DataBlender에서 제안하는 주요 기능인 스키마 통합 기능, 통합 검색 기능, 통합 검색 결과 관리 기능, 유무선 통합 인터넷 환경 지원에 대해 자세히 서술하고, 4장에서는 이를 지원하기 위해 설계된 DataBlender의 시스템 구조 및 내부 모듈 구조에 대하여 기술하고 결론을 맺는다.

II. 기존 연구

정보 통합 기술은 그 동안 많은 연구가 진행되어 왔으며, 현재 많이 연구되고 있는 분야는 통합 정보의 실체화(materialize) 여부에 따라 크게 데이터 웨어하우스 기술과 미디어이터 기반 통합 기술로 구분된다.

데이터 웨어하우스 기술은 통합 정보를 실체화하기 때문에 통합되는 정보 소스들의 변경이 자주 발생하지 않고 변경 내용이 실시간으로 반영되지 않아도 되는 의사 결정 시스템, 통계 처리 응용 등에 적합하다. 미디어이터 기반 통합 기술은 가상의 통합 데이터베이스를 제공하므로 통합되는 정보 소스들의 변경이 자주 일어나고 변경된 내용이 실시간으로 통합 정보에 반영되어야 하는 환경에 적합하다.

미디어이터 기반 통합 시스템은 통합 자료 모델에 따라 관계형 자료 모델 기반, 객체지향 자료 모델 기반, OEM 모델 기반, XML 자료 모델 기반 등이 있다. 통합 대상 정보 소스가 기존의 구조화된 정보에서 반구조화된 정보까지 다양화됨에 따라 구조화된 정보를 기반으로 한 통합 모델 사용에서 반구조화 자료 모델 사용으로 변화되고 있다. 특히 최근에는 인터넷에서 XML의 사용이 활성화됨에 따라 XML 모델을 기반으로 한 연구가 활발히 진행되고 있다.

XML 기반 통합 시스템으로 MIX[1], Tukwila[2], XMF[3] 시스템 등이 있으며, 각 각의 시스템마다 응용 분야에 차이가 있다. MIX는 DTD를 기반으로 통합 스키마를 관리하며, XMAS라는 통합 스키마 정의를 자체 정의하고 있다. DTD는 다양한 자료 형태 지원 미진, 계승 등의 객체지향

개념 미지원 등 표현력이 미진한 탓에 통합되는 자료의 속성이 손실될 가능성이 많다. Tukwila 시스템은 웹 정보 소스 통합을 목표로 하고 있고 스키마 통합에 대해서는 고려하지 않고 있으며, 주로 웹 정보 소스의 동적인 변화 환경에 능동적으로 대처하는 적응형 질의 처리에 중점을 두고 있다. XMF는 자체 정의된 스키마 규격에 따라 사용자가 통합 스키마와 지역 스키마를 서술하고, 통합 스키마와 지역 스키마간의 맵핑 정보를 중재 규칙으로 직접 기술하도록 되어 있다. 즉, 통합 스키마 정의가 선언적이지 아니라 절차적이다.

본 논문에서는 통합 정보의 특성 손실을 최소화하기 위해 XML Schema[4] 기반 XML 모델을 지원하고, 이질 자료 모델간의 충돌 해결을 표현할 수 있는 다양한 요소를 갖고 있는 XQuery를 활용한 통합 스키마 정의를 제공하고, 유무선 통합 인터넷 환경에서 통합 검색 서비스를 활용할 수 있도록 이동 단말기에서도 사용 가능한 미디어이터 기반 정보 통합 시스템인 DataBlender를 제안한다.

III. DataBlender 기능 설계

1. 시스템 특징

통합 시스템 구축시에는 통합 대상 정보 소스의 특성 및 통합 시스템의 운영 환경을 고려하여야 한다. 수 많은 불특정 웹 정보 소스를 통합하는 환경에서는 통합 대상 정보 소스가 미리 정해지는 것보다는 수시로 변화가 발생하고, 통합 대상 정보들의 스키마도 대단히 가변적이고 미리 정해지지 않는 경우가 많다. 이와 같은 환경에서는 동적으로 정보 소스를 통합하며, 통합 스키마도 통합 대상 정보를 기반으로 자동 생성이 바람직하다. 그러나 일반적인 기업내 기업간 비즈니스 환경에서는 통합 대상 정보 소스가 미리 정해지고, 비슷한 구조를 갖는 많은 데이터가 한 곳에 집중되어 있는 경우가 많으므로 이와 같은 환경에 맞는 시스템 구축이 요구된다.

DataBlender는 일반적인 인터넷 비즈니스 환경을 대상으로 정보 통합을 제공하기 위한 시스템으로 주요 기능은 다음과 같다.

- o 통합 데이터 모델로 XML을 지원하며 통합 스키마 규격은 XML Schema를 지원한다.
- o 각 지역 스키마로부터 통합 스키마를 정의하는 통합 스키마 정의로 XQuery 기반의 XQuerySD를 제공한다.

- o 통합 정보에 대한 검색을 지원하며, 통합 검색 질의어로 XQuery를 지원한다.
- o 질의 검색 결과를 XML로 제공하며, DOM 인터페이스를 통해 질의 결과에 접근할 수 있다.
- o 응용 프로그램 인터페이스로 자바 인터페이스를 제공한다.
- o 무선 이동 단말기에서도 통합 검색 서비스를 지원한다.

2. 통합 스키마 관리

통합 스키마 정의는 지역 정보들을 통합하여 새로운 통합 정보 뷰를 정의하기 위한 것으로 통합 정보 뷰의 스키마 구조 정의는 XQuery를 기반으로 정의된 XQuerySD를 이용한다. XQuerySD는 FOR 절에 DOCUMENT() 대신 SCHEMA()라는 자체 정의한 함수를 사용한다는 점을 제외하고는 대부분의 기능이 XQuery와 유사하다. SCHEMA()는 대상이 XML 문서가 아니고 XML 스키마라는 것을 의미한다. 그리고 또 하나 차이점은 처음 시작할 수 있는 질의 표현식이 엘리먼트 생성자와 FLWR 표현식으로 한정된다는 것이다.

DataBlender에서는 XQuery 기반의 스키마 정의를 사용함으로써 사용자는 새로운 정의를 습득할 필요가 없으며, 통합이 되는 정보의 집합을 원하는 정보만으로 미리 한정할 수 있다는 장점이 있고, 또한 XQuery의 다양한 기능(element constructor, user defined function, cast function 등)을 이용하여 스키마의 구조상 충돌 및 의미상 충돌을 해결할 수 있다. 그림 1은 XQuerySD를 이용하여 통합 스키마를 정의한 예제이다.

통합 스키마 규격은 W3C에서 제안하는 XML Schema를 따른다. XML Schema는 DTD보다도 한 단계 발전된 것으로 XML 문서의 구조 표현 방법인 DTD에서 정의하고 있는 엘리먼트간의 트리 구조에 의한 XML 모델 표현뿐만 아니라 각 엘리먼트가 가질 수 있는 기본 자료의 타입 속성까지도 표현할 수 있어 XML 문서의 구조 및 속성 표현력이 한층 강화되었다. 기존 DTD 기반의 XML 통합에서는 지역 정보 소스에서 부여하던 정보의 속성을 일부 손실하는 문제가 있었으나 XML Schema은 이와 같은 정보 손실이 발생하지 않으므로 이질 자료 모델 정보 통합에 효과적이다. 또한 통합 정보 구조 정보를 XML Schema로 미리 정의함에 따라 질의시 스키마 정보 활용에 의해 검색이 용이하고 최적화가 가능하다.

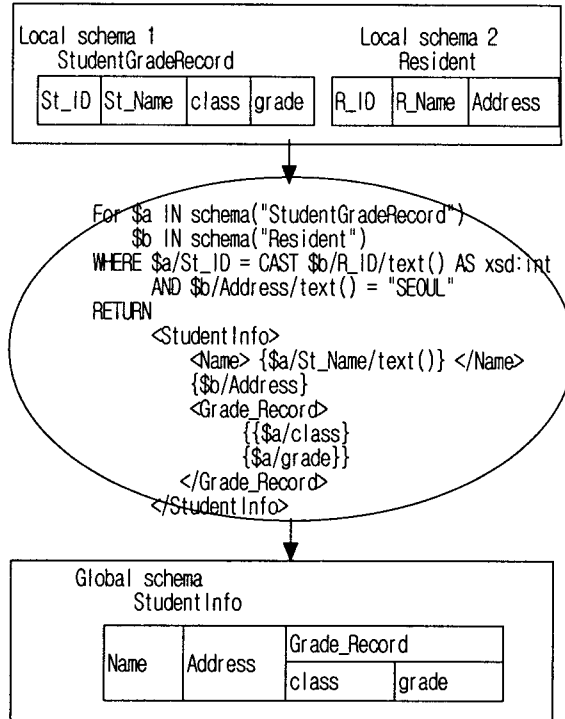


그림 1. 통합 스키마 정의 예

3. 통합 질의

통합 질의어는 W3C에서 논의되고 있는 XQuery 규격을 기반으로 한다. 표준 작업안 XQuery는 한번에 여러 질의를 동시에 전달할 수 있으며 다양한 검색 패턴을 지원하고 있으나 DataBlender 통합 질의어에서는 하나의 질의만을 지원하며, 지원하는 질의 종류는 다음과 같다.

- o path expression
- o element constructor
- o FLWR expression
- o expression involving operator :
 - numeric op : +, -, *, div, mod
 - comparison op : =, !=, ==, !=, <, <=, >, >=
 - logical op : and, or
 - sequence op : to, union, intersect, except, before, after
- o sorting expression
- o quantified expression
- o expressions relating datatype : instanceof, cast

그림 2는 XQuery를 이용한 검색 예제이다.

```
FOR $a IN schema("StudentInfo")
WHERE $a/Address/data() = "Seoul"
AND $a/Grade_Record/class/data() = "English"
RETURN
  <Seoulstudent>
    { $a/Name }
    { $a/Grade_Record/grade }
  </Seoulstudent>
```

그림 2. 통합 질의 예

4. 통합 검색 결과 관리

통합 검색 결과는 XML 문서 개념을 기반으로 DOM 인터페이스에 의해 접근할 수 있도록 한다. DOM 인터페이스에는 XML 문서를 구성하는 노드의 개수 제공과 같이 통합 질의를 완전히 처리하지 않는 한은 지원할 수 없는 인터페이스가 있다. 응용에 따라 통합 검색 결과를 모두 얻거나 일부만 접근 가능하므로 일부만 접근하는 응용에서는 통합 질의를 완전히 처리하여 결과를 제공하는 방식이 바람직하지 않다. 그러므로 DataBlender에서는 사용자의 응용 환경에 맞게 선택적으로 사용할 수 있도록 다음과 같이 3 종류 통합 질의 결과를 얻는 방법을 제공한다.

- o singleton query : 통합 질의 결과를 하나의 XML 문서로 제공
- o iterator query : 통합 질의 결과를 여러 XML 문서의 집합으로 보고 커서 개념을 이용하여 결과 XML 문서를 하나씩 접근
- o collection query : 통합 질의 결과를 여러 XML 문서가 통합된 하나의 큰 XML 문서로 보고 XML 문서를 구성하는 각각의 서버 XML 문서에 대한 직접 접근을 지원하는 iterator query를 이용하면 XQuery의 종류에 따라 전체를 모두 수행하여 결과를 제공할 수도 있고, 요구에 따라 질의를 수행하여 결과를 제공할 수 있으므로 응용 환경에 맞는 최적의 실행 환경을 제공할 수 있다.

5. 유무선 통합 인터넷 환경 지원

유무선 통합 인터넷 환경을 지원하기 위해 DataBlender는 이동 단말기와 같이 적은 자원, 잦은 단절이 발생하는 네트워크 환경을 고려하여 무선 환경에서도 통합 검색 기능을 사용할 수 있도록 한다.

이를 지원하기 위해 통합 스키마 정보 등 메타 정보를 이동 단말기에 복제하여 활용하며, 복제된 메타 정보와 마

스터 메타 정보간의 일관성을 유지시켜 준다. 또한 비동기식 통합 검색을 지원한다. 즉, 통합 질의를 요구 후 응용 프로그램이 미디어이터와의 연결을 끊은 후 자체내 작업 수행 후 다시 접속하여 통합 결과를 얻을 수 있도록 한다.

IV. DataBlender 구조 설계

1. 시스템 구조

DataBlender는 유무선 통합 인터넷 서비스 환경을 기반으로 그림 3과 같이 크게 3개의 서버 시스템으로 구성되어 있다. DataBlender.m은 미디어이터로 정보 통합 기능을 제공하는 핵심 서버 시스템으로 통합 스키마 관리 및 통합 질의 처리 기능을 제공한다. DataBlender.w는 통합되는 각각의 정보 소스와 미디어이터 시스템을 연동하는 래퍼 서버 시스템으로 연동되는 정보 소스에 따라 RDBMS용 래퍼, ODBMS용 래퍼, 파일 시스템 래퍼를 제공한다. DataBlender.s는 이동 단말기에 복제된 메타 정보와 DataBlender 서버에서 관리되는 마스터 메타 정보의 동기화를 담당하는 싱크로나이저 서버 시스템으로 단말기에 탑재되는 Synchronizer-Sub와 서버에 탑재되는 Synchronizer-Pub로 구성된다. DataBlender.s는 이동 단말기의 특성을 고려하여 단말기내에서 처리할 수 있는 기능을 최대화하여 서버와의 연결을 가능한 지연하기 위한 구성 요소이다.

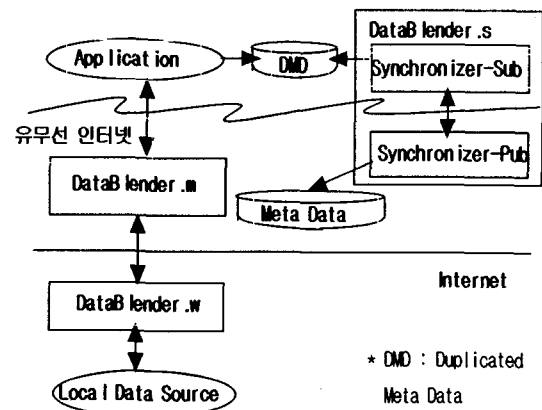


그림 3. DataBlender 시스템 구조

DataBlender.m은 클라이언트/서버 구조로 구성되며, DataBlender.w는 프록시와 서버로 구성된다. 미디어이터 클

라이언트는 사용자 응용 시스템과 통합되어 유무선 단말기에서 운영되며, 무선 단말기에 동작하는 미디어이터 클라이언트는 DataBlender 기능 중 통합 검색 기능만 사용할 수 있다. 그리고 미디어이터 서버와 프록시 랩퍼는 하나로 통합되어 같은 호스트에 동작한다. 미디어이터 서버는 데몬으로 동작하며 사용자 응용 시스템에서 연결 요구를 받으면 하나의 미디어이터 서버 쓰레드를 생성하여 서비스를 제공한다. 또한 지역 정보 소스의 서비스를 받기 위해서는 프록시 랩퍼 쓰레드를 지역 정보 소스별로 생성하여 이를 통해 실제 랩퍼 서버와 연결되어 서비스를 받는다. Synchronizer-Sub는 필요시 호출되는 유틸리티로 동작하며 Synchronizer-Pub는 데몬으로 동작하며 Synchronizer-Sub의 요구를 받으면 쓰레드를 생성하여 서비스한다.

2. 모듈 구조

DataBlender를 구성하는 각 서버 시스템의 모듈 구조는 그림 4와 같다.

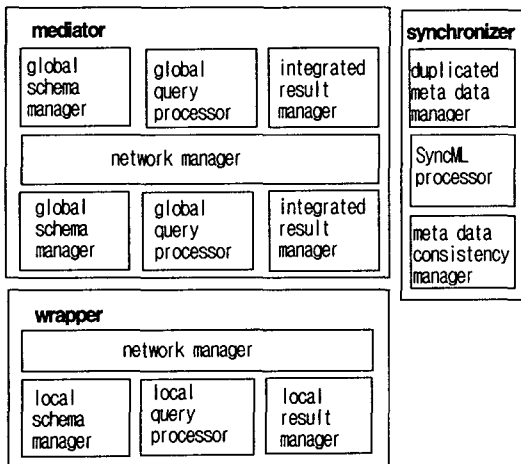


그림 4. DataBlender 모듈 구조

2.1. 미디어이터

연결 관리자(network manager)는 클라이언트/서버 구조를 지원하며 미디어이터 서버의 쓰레드를 제어한다. 통합 스키마 관리(global schema manager)는 스키마 정의어 XQuerySD를 처리하여 통합 스키마 구조 정보와 통합 스키마와 지역 스키마간의 맵핑 정보를 관리한다. 또한 통합 스키마 정보, 지역 정보 소스에 대한 정보, 사용자 정보 등을 저장 관리한다. 통합 질의 처리기(global query processor)는

XQuery를 처리하는 모듈로 XQuery에 대한 파싱, 정당성 검사 및 지역 스키마를 이용한 질의로 변환하고 이를 실행하기 위한 실행 계획을 작성하여 각 각의 지역 정보 소스에 질의를 분배 및 결과를 얻어 최종 질의 처리 결과를 구성한다. 통합 결과 관리자(integrated result manager)는 DOM 인터페이스를 지원하며 통합 질의에 의해 생성된 결과 문서들을 저장, 관리하며 커서 개념에 따라 요구되는 결과 문서를 제공하는 역할을 수행한다.

2.2. 랩퍼

연결 관리자(network manager)는 미디어이터에 랩퍼 인터페이스를 제공하고 랩퍼 서버의 해당 서비스 요청을 전달한다. 지역 스키마 관리자(local schema manager)는 지역 정보 소스의 스키마 정보를 XML 뷰로 제공한다. 지역 질의 처리기(local query processor)는 미디어이터에서 전달한 XQuery를 지역 정보 소스에서 제공하는 질의 기능으로 변환하여 처리하고, 지역 결과 관리자(local result manager)는 결과를 얻어 XML 문서로 구성하여 제공한다. 지역 스키마 관리기 중 지역 정보 소스에 따라 스키마 정보를 XML로 재구성하는 부분과 지역 질의 처리기 중 XQuery를 지역 정보 소스의 질의 기능으로 변환하는 부분만이 연동되는 랩퍼 타입에 따라 재구성이 되는 부분으로 모듈화하여 쉽게 plug-and-play가 가능하다.

2.3. 싱크로나이저

메타 정보 요구 메타 정보 전달 등을 SyncML로 변환하여 실제 Synchronizer-Sub와 Synchronizer-Pub간 전달을 담당하는 SyncML 처리기(SyncML processor)와 전달된 메타 정보를 단말기 운영 환경에 맞게 저장 관리하는 복제 메타 정보 관리기(duplicated meta data manager), 단말기로 복제된 메타 정보들에 대한 정보를 유지 관리하며 미디어이터의 통합 스키마 관리기에 의해 메타 정보의 변경이 발생하면 이를 해당 단말기에 변경된 메타 정보를 전달하는 메타 정보 일치성 관리기(meta data consistency manager)로 구성된다.

V. 결론

XML 기반 데이터베이스 통합 미들웨어인 DataBlender를 제안하고, DataBlender의 기능, 구조 및 세부 모듈 설계 내용에 대해 기술하였다. DataBlender는 통합되는 정보의 의

미 손실을 최소화하고, 통합 자료 모델들간의 충돌 해결 표현이 용이하며 통합 스키마 정의시 통합 대상 정보의 범위 제한이 가능하다는 장점이 있다. 또한 XML Schema, XQuery, DOM 등 XML 관련 주요 표준안을 수용하고 이동 단말기에서도 통합 검색 서비스를 활용할 수 있도록 설계함에 따라 유무선 통합 인터넷 환경에서 새로운 서비스 환경 구축의 기반이 된다.

DataBlender는 현재 본 설계 내용을 바탕으로 자바로 구현 중이다. DataBlender는 인터넷 비즈니스 환경에서 통합 검색 서비스를 지원하기 위하여 만들어진 기반 시스템으로 차후 바이오 정보 처리 분야에서 활용할 수 있도록 바이오 정보 소스와의 효과적인 연동을 위해 이를 보완 확장할 예정이다.

참 고 문 헌

- [1] Chitanya Baru, 외 5인, "XML-Based Information Mediation with MIX," In Proc. of ACM-SIGMOD'99, Philadelphia, PA., 1999.
- [2] Zachary G. Ives, Alon Y. Levy, Daniel S. Weld, "Adaptive Query Processing for Internet Applications," IEEE Data Engineering Bulletin, Vol.23, No.2, June 2000.
- [3] 이경하, 조정수, 이강찬, 이규철, "XMF: XML 기반 분산 이질 정보 자원의 통합 프레임워크, KDBC2000 학술발표논문집, pp.262~270, 2000.
- [4] W3C Recommendation, "XML Schema Part 0:Primer," <http://www.w3.org/TR/xmlschema-0>, 2001.
- [5] W3C Working Draft, "XQuery 1.0: An XML Query Language," <http://www.w3.org/TR/xquery>, 2001.
- [6] W3C Recommendation, "DOM Level 2 Core," <http://www.w3.org/TR/DOM>, 2000.

이미영(Mi-Young Lee)

중신회원



1981년 2월 : 서울대학교 식품영양학과 졸업

1983년 2월 : 서울대학교 계산통계학과 졸업(이학석사)

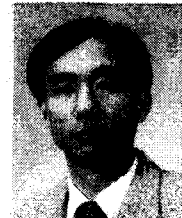
2001년 2월 : 충남대학교 컴퓨터공학과 박사과정 수료

1988년 3월 ~ 현재 : 한국전자통신연구원 책임연구원

<관심분야> : 데이터베이스, 콘텐츠 관리

김명준(Myung-Joon Kim)

중신회원



1978년 2월 : 서울대학교 계산통계학과 졸업

1980년 2월 : 한국과학기술원 전산학과 졸업(이학석사)

1986년 5월 : 프랑스 Nancy제1대학교 응용수학및전산학과(이학박사)

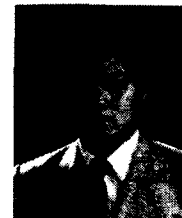
1986년 7월 ~ 현재 : 한국전자통신연구원 책임연구원

컨텐츠기술연구부/부장

<관심분야> : 데이터베이스, 분산처리, 콘텐츠 관리 및 보호

이규철(Kyu-Chul Lee)

정회원



1984년 2월 : 서울대학교 컴퓨터공학과 졸업

1986년 2월 : 서울대학교 컴퓨터공학과 졸업(공학석사)

1990년 8월 : 서울대학교 컴퓨터공학과 졸업(공학박사)

1989년 2월 ~ 현재 : 충남대학교 컴퓨터공학과 교수

<관심분야> : 데이터베이스, XML, ebXML