# IHWA(Information Harvest Warehouse) 시스템의 정보수집 서비스

오동익, 정종석
순천향대학교 공과대학 정보기술공학부

# Gathering Services of the Information Harvest Warehouse(IHWA) System

Dong-Ik Oh, Jong-Suk Jung
Division of Information Technology Engineering, Soonchunhyang University

**요 약**

IHWA는 기존의 객체 컴포넌트를 통합함으로써 새로운 응용프로그램을 작성할 수 있는 CBSE(Component Based Software Engineering)기법을 바탕으로 개발되어진 웹 기반 정보 저장/검색 시스템이다. 이 시스템은 1997년에 그 1차 버전이 개발되어 사용되어 왔으며, 현재는 보다 나은 견고함과 일반성의 제공, 그리고 전자상거래 영역에 대한 준비를 위해 시스템에 대한 중요한 수정작업이 진행 중에 있다. 본 논문에서는 새로운 IHWA시스템의 설계에 적용된 원리 및 내용을 설명하고, 특히 정보 수집 서비스를 제공하기 위해 필수적으로 요구되는 정보 수집기들의 구조 및 구현에 대해 설명하고자 한다.

## 1. 서론

IHWA 시스템[1]은 다양한 웹 지향 정보검색 시스템 구축을 위한 정형화된 하나의 모델을 제공하고자 개발된 통합형 웹 기반 정보 저장/검색 시스템이다. IHWA는 CBSE (Component Based Software Engineering) 기법에 바탕을 두고 설계되었으며, 다양한 소프트웨어 컴포넌트들을 통합하고 새로운 컴포넌트를 개발함으로써 구현되었다. 그림 1에서 보는 것처럼 IHWA시스템의 구조는 웹 프로그래밍에서 사용될 수 있는 EJB(Enterprise JavaBeans)와 JCC(Java Commerce Client) 객체 컴포넌트를 기반으로 구현되어져 있다.

IHWA시스템에는 Query Agent와 Rebuild Agent라 불리는 2개의 중요한 서버 측 컴포넌트가 있다. Query Agent는 클라이언트의 요구에 의해 데이터베이스에 대한 쿼리를 수행하고, 정보를 CORBA 객체 형식으로 클라이언트들과 다른 시스템에 제공하는 역할을 담당한다. Rebuild Agent는 CORBA/IIOP[4]를 통해 인터넷상에 존재하는 여러 IHWA 사이트간의 정보교환 작업을 수행하는 역할을 담당한다. IHWA는 Rebuild Agent를 통해 다른 IHWA사이트로부터 정보를 획득하고, 획득한 정보를 로컬 데이터베이스에 저장하며, 이러한 정보는 Query Agent를 통해 사용자들에게 제공되게 된다.

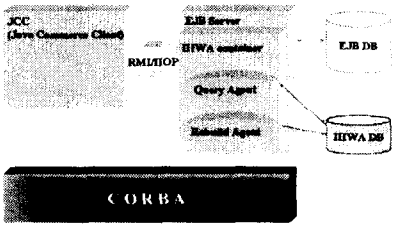


그림 1. 기존 IHWA시스템의 구조

최근의 CBSE 기술 발전에 힘입어[6] 본 연구팀은 기존의 IHWA시스템에 대한 수정 및 보완작업이 필

본 연구는 정보통신부의 대학 S/W 연구센터 지원사업에 의해 수행된 것임.

요하다는 판단을 하게 되었는데, 이는 새로운 디자인 기술들을 활용함으로써 보다 나은 시스템의 구성 및 시스템의 정확성을 높일 수 있는 가능성들이 파악되었기 때문이다. 이러한 기술들에 대한 연구 및 분석을 통하여 우리는 IHWA 시스템에 대한 재구축 작업을 수행하게 되었고 그림 2에서 보는 것과 같은 형태의 새로운 시스템 아키텍처를 구성하게 되었다[2,3].

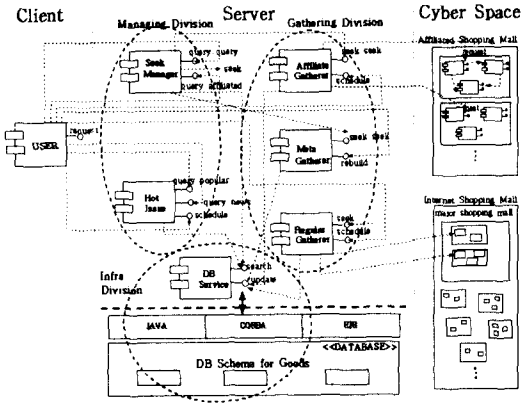


그림 2. 새로운 IHWA시스템의 구조

새로운 IHWA시스템에서는 여러 개의 새로운 서버측 컴포넌트들이 추가되었고, 그들 상호간의 인터페이스는 컴포넌트의 재 활용성을 최대한 고려하여 디자인되었다. 원래의 IHWA시스템에서와 마찬가지로 새로운 IHWA시스템의 컴포넌트 구조는 JCC기반의 클라이언트 컴포넌트와 EJB기반의 서버 컴포넌트들을 사용한다. 그리고 이 두 가지 컴포넌트들 사이의 통신은 CORBA/IIOP채널을 통하여 이루어지게 된다.

새로운 IHWA시스템에서는 6개의 서버측 컴포넌트들이 Managing, Gathering, Infrastructure의 3가지 Division으로 분리되어 제공되고 있다. Managing Division은 시스템에서의 모든 동작을 컨트롤하는 역할을 담당하는데, 외부로부터의 검색요구, 데이터베이스 갱신, 정보수집과 같은 다양한 시스템 동작을 제공하고 관리한다. Gathering Division은 3가지의 정보수집기로 구성되어 있고 이들은 분산된 환경에서의 정보를 수집하는 역할을 담당한다. 마지막으로 Infrastructure Division은 모든 시스템자원과 관련된 데이터 및 EJB와 CORBA객체에 대한 정보들을 조작하고 저장하는 역할을 담당한다.

## 2. IHWA의 정보수집 서비스

본 논문의 주요 내용은 IHWA시스템의 Gathering Division에서 제공하는 정보수집기의 개발에 관한 것이다. 이 Division은 Affiliated Gatherer와 Regular Gatherer 및 Meta Gatherer의 세 가지 수집기로 구성되는데, 이들을 통해 IHWA 시스템은 분산된 환경에서 정보 수집 활동을 할 수 있게 된다.

### 2.1 Affiliated-Gatherer

Affiliated-Gatherer를 통해 각각의 IHWA 사이트들은 상품 카탈로그에 대한 정보를 안전하게 교환할 수 있다. 이러한 정보수집 동작은 일정한 시간 간격을 두고 실행되어지며, 일단 하나의 IHWA 사이트의 상품정보가 갱신되면 일정시간 후에는 서로 협약된 모든 IHWA사이트들이 이 정보를 수집하여 동일한 상품 카탈로그를 유지하게 된다. 따라서 사용자는 어떠한 IHWA사이트를 방문하더라도 전체 IHWA사이트들의 상품정보를 검색할 수 있게 된다.

모든 IHWA사이트들이 서로의 상품 카탈로그 정보가 공유될 수 있도록 하는 또 하나의 방법은 상품 카탈로그 정보가 갱신되었거나 또는 새로운 정보를 획득한 IHWA사이트로 하여금 그 정보를 다른 사이트에게 전달하도록 하는 방법이 있을 수 있다. 이렇게 함으로써 각각의 IHWA사이트들은 갱신된 정보를 리얼타임으로 활용할 수 있게 된다.

위와 같은 Affiliated-Gatherer의 모든 정보교환은 CORBA/IIOP통신을 통해서 3-tier 계층으로 수행되도록 구현되었다. 그림 3은 이러한 Affiliated-Gatherer의 역할을 설명하고 있다.

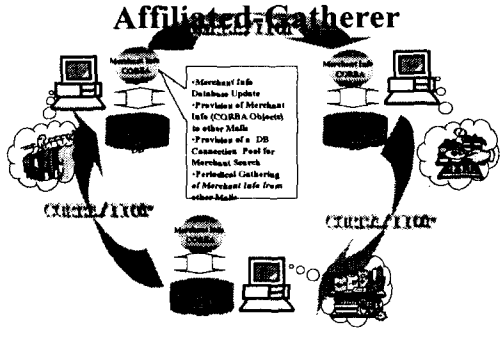


그림 3. Affiliated Gatherer

그림 4는 Affiliated-Gatherer에 대한 시스템 구성을 나타내고 있다. 그림에서 보는 것과 같이 각 IHWA시

스템의 CORBA객체는 일정한 시간간격을 두고 서로의 데이터베이스에 저장된 상품정보를 교환한다. 각각의 IHWA 사이트는 서로의 인증과 통신상의 보안 및 전송 데이터의 무결성을 위하여 SSL (Secure Socket Layer)프로토콜에 바탕을 둔 CORBA 객체 통신을 수행한다.

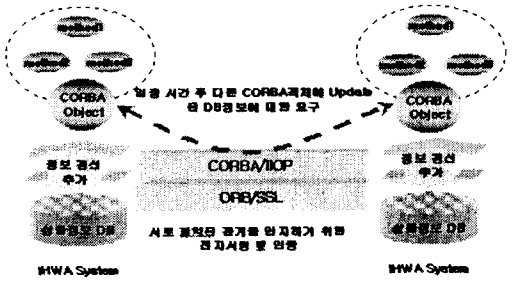


그림 4. Affiliated-Gatherer의
시스템 디자인

그림 5는 각 IHWA사이트의 서버 측 CORBA객체에 대한 구현을 설명하고 있다. 그림에서 보는 것처럼 각 IHWA사이트의 CORBA객체는 시스템에 적절한 수의 JDBC 연결 객체를 미리 생성 한 후 이를 통해 데이터베이스에 연결된다. 이렇게 함으로써 클라이언트의 요구가 있을 시마다 하부 데이터베이스에 대한 연결을 설정해야 하는 오버헤드를 줄일 수 있고, 이를 통해 클라이언트의 요구에 빠르게 응답 할 수 있다. 서버는 각 클라이언트의 요구를 수용하기 위하여 서비스를 하고 있지 않은 연결 객체에 대한 정보를 유지하고 이를 적절히 할당함으로써 시스템의 로드 밸런싱을 유지할 수 있다.

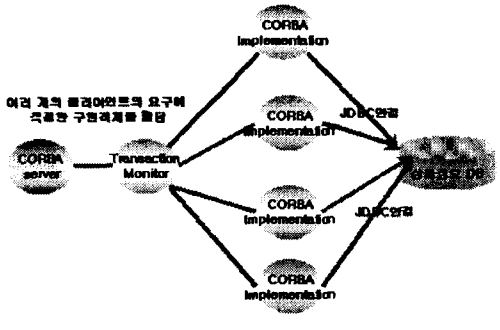


그림 5. IHWA사이트의 CORBA 객체

## 2.2 Regular-Gatherer

모든 인터넷 몰들이 상품 카탈로그 및 보유한 정보를 검색하고 저장하는 방법으로서 IHWA를 채택한다고 가정할 수는 없다. 따라서 IHWA시스템이 웹 상에서 광범위한 검색 능력을 갖게 하기 위해서는 협약되지 않은 인터넷 몰에서도 상품정보를 수집할 수 있어야 할 것이다. 이러한 문제에 대한 해결방법으로 개발된 것이 Regular-Gatherer다.

Regular-Gatherer는 서로 협약되어지지 않은 인터넷 쇼핑몰로부터 상품 카탈로그를 수집하고 수집된 카탈로그에서 상품정보를 추출하여 데이터베이스에 저장하는 기능을 수행한다. 문제는 각각의 몰에서 제공하는 상품 카탈로그가 아주 다양한 방식으로 제공될 수 있다는 것인데, 이러한 문제를 해결하기 위해서 Regular-Gatherer는 표준화 된 상품 정보만을 수집하는 역할을 담당한다. 상품정보는 XML [5] 데이터로 표현되도록 하여 표준화의 문제를 해결하였다.

Regular-Gatherer는 각각의 상품 종류마다 DTD를 구성하게 하고, XML로 표현된 상품 카탈로그 정보를 수집한 후, 문서의 Validity를 확인하고, 문서에서 상품정보를 추출한 후 이를 하부 데이터베이스에 저장하게 된다. 분산된 인터넷 환경에서의 XML문서의 수집은 FTP프로토콜을 통한 Java 에이전트에 의해 수행되어진다. 그림 6은 이와 같은 Regular-Gatherer의 기능을 보여주고 있다.

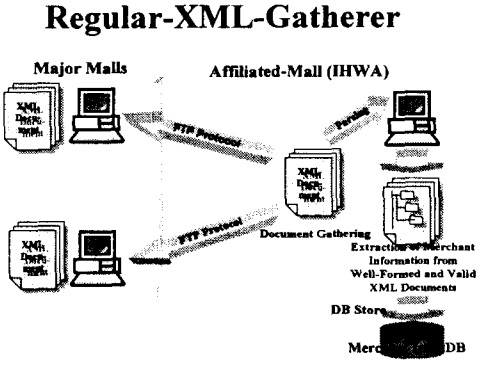


그림 6. Regular Gatherer

Regular-Gatherer는 크게 XML문서 수집부분, XML문서 파싱부분, 추출된 정보를 데이터베이스에 저장하는 부분의 세 부분으로 나눠져 구성된다. 먼저 XML문서 수집부분에서는 FTP프로토콜을 사용하여 XML문서들이 저장되어 있는 디렉토리의 내용을 읽어온다. 읽어 온 문서는 HTML 형식이며 수집기는

이 문서의 태그들을 파싱 함으로써 사이트에서 제공하는 XML파일의 이름들을 추출 할 수 있고, 따라서 그 파일들을 수집 할 수 있다. 일단 수집된 XML 문서는 이미 정해진 DTD를 이용하여 그것의 Validity가 검증되고, 파서에 의하여 상품정보가 추출되며, 추출된 정보는 Java 객체형태로 구성되어 하부 데이터베이스에 저장되게 된다.

### 2.3 Meta-Gatherer

Meta-Gatherer는 이미 언급된 두 개의 다른 수집기들을 통해서 얻을 수 없는 상품정보를 수집하는 역할을 담당한다. 이를 위해 Meta-Gatherer는 웹 상에서 표준화 되어있지 않은 정보, 즉, 일반적인 웹 문서(HTML)나 DTD가 정의되어 있지 않은 XML문서들을 수집한 후, 이를 통해 관련된 상품정보를 추출한다. 하지만 정형화되지 않은 이와 같은 문서들에서 온전한 상품정보를 추출한다는 것은 거의 불가능하다. 따라서 Meta-Gatherer는 완벽한 정보를 수집하기보다는, 불완전할 수도 있으나 이러한 정보에 대한 접근을 필요로 하는 검색자들을 위한 서비스를 제공하는 수단으로 사용되게 될 것이다.

## 3. 결론

다양한 웹 정보 시스템 구축을 위한 정형화된 모델을 제공하고자 하는 목표로 개발이 시작된 IHWA는 이제 그 목적을 좀 더 효율적으로 달성하기 위한 대대적인 재구성의 단계에 있다. 새로 구성되는 IHWA는 웹 응용 개발 컴포넌트 객체인 EJB와 JCC를 기반으로 설계되었고, Managing, Gathering, Infrastructure의 세 Division간의 정교한 인터페이스를 제공함으로써 컴포넌트의 재 활용성을 높일 수 있도록 설계되었다.

이 중 Gathering Division에서는 분산환경에서의 정보교환을 위해서 세 가지의 Gatherer를 제공하는데, 이들을 통하여 IHWA는 웹에서 효과적이고 안정적으로 상품정보를 유지하고 검색 할 수 있는 기반을 제공한다.

세 가지 정보 수집기를 통해 얻어진 상품 카탈로그 정보들은 Affiliated-Gatherer를 통해 분산된 IHWA사이트간에 서로 교환되어지며, 따라서 모든 협약된 IHWA사이트들은 동일한 상품 카탈로그를 유지하게 된다. 사용자는 어떤 IHWA사이트에 접속하더라도 동일한 상품정보를 검색하고 이에 대한 자료를 얻을 수 있다.

본 연구팀에서는 현재 Meta-Gatherer가 수집하는 문서에서의 효과적인 정보 추출 방법에 대해 연구하고 있으며, 다양한 상품정보 형식에 맞게 수집기가 자동으로 생성 될 수 있는 "자동 Gatherer 생성기"에 대한 연구도 앞으로 진행 할 계획을 가지고 있다.

## [참고문헌]

[1] Incheon Paik, W Lee, "Design of Scalable User Oriented Internet Information Search System Using Distributed Object", Proceedings of TOOLS 24, Sep. 1997.

[2] Incheon Paik, Tongwon Han, "Design and Implementation of electronic commerce Search Engine Component" , Proceedings of AoM/IAoM 17, August. 1999.

[3] Incheon Paik, Tongwon Han, "A Novel Component Architecture for Electronic Commerce Search System" , To appear in the Proceedings of SNPD 00 Conference.

[4] Robert Orfali, Dan Harkey, *Client/Server Programming with Java and Corba*, Wiley Computer Publishing, 1998.

[5] St. Laurent and Cerami, *Building XML Applications*, McGraw-Hill., 1999.

[6] Sherif Yacoub, "A Model for Classifying Component Interfaces", 1999 International Workshop on Component-Based Software Engineering, Software Engineering Institute, Carnegie Mellon