컨텐츠 통합 검색을 위한 질의어 처리 시스템 구현

김 영 균, 이 명 철, 이 미 영, 김 명 준
한국전자통신연구원 컴퓨터소프트웨어연구소 컴퓨터시스템연구부

An Implementation of a Query Processing System for an Integrated Contents Database Retrieval

Kim young-kyoon, Lee myung-cheol, Lee mi-young, Kim myung-joon
Computer System Department, Computer & Software Research Lab., ETRI
E-mail : {kimyoung, mclee, mylee, joonkim}@etri.re.kr

요 약

다양한 종류의 컨텐츠 데이터베이스를 구축하고, 이를 인터넷 서비스로 제공하는 인터넷 포탈 서비스 응용이나 전자상거래 등에서 기존에 구축되어 있는 여러 형태의 컨텐츠 데이터베이스들을 통합하여 새로운 컨텐츠 서비스를 제공하기 위한 많은 노력들이 수행되고 있다. 이는 사용자가 활용하려는 컨텐츠가 어떤 데이터베이스 또는 어느 인터넷 서비스 응용에서 제공되는지를 파악해야 하는 사용자의 부담을 줄이고, 다수의 컨텐츠 데이터베이스들을 사용자에게 단일한 뷰(view)를 제공하므로서 사용자 이용 편의성을 높일 수 있다. 본 논문에서는 다양한 컨텐츠 데이터베이스들이 인터넷에 분산되어 있을 뿐만 아니라 서로 상이한 데이터베이스 시스템들(즉, 관계형 DB/객체형 DB)에서 관리되는 환경에서 XML 자료 모델을 기반으로 통합된 하나의 가상 데이터베이스를 구축하고 검색하는 통합 검색 시스템의 핵심 요소인 질의 처리 시스템을 설계 및 구현한다.

Abstract

There have been many considerations to develop new content services that integrate a variety of contents databases being already constructed and then produce new content services which are more valuable than existing services in many applications such as Internet portal, EC, and CRM. By doing the above thing, the burden of searching databases to access interesting databases and service applications can be reduced and the database availability of users is also enhanced through a single view integrating multiple contents databases. This paper presents implementation details of the query processing system that is a core component of the database integration system, which can construct a virtual database that integrates databases being managed by multiple heterogeneous database systems using XML data model and support a query facility on the integrated database.

I. 서론

컨텐츠를 유통하는 기업들이 광역화됨에 따라 분산 저장 관리되고 있는 컨텐츠에 대한 접근이 요구되고 있으며 자신의 컨텐츠뿐만 아니라 다른 기업에서 관리하는 컨텐츠의 활용도 요구되고 있다. 즉, 다양한 종류의 컨텐츠 데이터베이스를 구축하고, 이를 인터넷 서비스로 제공하는 인터넷 포탈 서비스 응용이나 전자상거래 등에서 기존에 구축되어 있는 여러 형태의 컨텐츠 데이터베이스들을 통합하여 새로운 컨텐츠 서비스를 제공하기 위한 노력들이 이루어지고 있다. 이는 사

용자가 활용하려는 컨텐츠가 어떤 데이터베이스 또는 어느 인터넷 서비스 응용에서 제공되는지를 파악해야 하는 사용자의 부담을 줄이고, 다수의 컨텐츠 데이터베이스들을 통합하여 사용자에게 단일한 뷰(view)를 제공함으로써 사용자 이용 편의성을 높일 수 있다.

데이터베이스 통합 기술은 데이터웨어하우스 기술과 미디에이터(mediator) 기반 통합 기술로 분류할 수 있으며, 미디에이터 기반 통합 기술 연구들이 최근에 많이 이루어지고 있다. 미디에이터를 기반으로 여러 데이터베이스들을 통합하는 시스템으로 [1,2,3,4]가 있으며, 이들은 여러 데이터베이스들에 대한 가상의 단일 뷰를 제공하는 시스템이다.

본 논문에서는 통합되는 데이터베이스의 의미 손실을 최소화하고, 통합 데이터베이스 모델들간의 충돌을 해결한 DataBlender 시스템[4]의 질의 처리 시스템의 설계와 구현 내용을 기술한다.

제안된 질의 처리 시스템은 사용자 질의 언어로 W3C에 표준으로 제안되어 있는 XQuery를 채택하고 있으며, 질의 처리 결과로 XML 문서를 반환한다. 질의 처리 시스템은 효율적인 질의 처리를 위하여 질의어 구문 분석, 관계 연산자 트리 구성, 실행 계획 생성 등 일련의 과정을 통하여 각각의 단위 데이터베이스들에서 수행될 수 있는 질의어 집합을 생성한다. 이후, 각 질의어들을 분산된 데이터베이스를 관리하는 개별 시스템들에서 병렬로 수행시키고, 그 결과 문서들을 통합하여 결과 문서를 생성한다.

논문의 구성은 다음과 같다. 2장에서는 DataBlender 시스템의 핵심 요소인 미디에이터의 개략 구조를 질의 처리 시스템을 중심으로 설명하고, 3장에서는 질의 처리 시스템의 내부 구조와 모듈별 기능, 질의어 처리 과정을 소개한다. 4장에서는 개발된 질의 처리 시스템을 이용한 질의 처리 과정을 데모를 통해 보여주고, 결론을 맺는다.

II. 통합 검색 시스템의 개략 구조

DataBlender 통합 시스템은 미디에이터 서브 시스템, 랩퍼(wrapper) 서브 시스템 그리고 무선 단말기 지원을 위한 동기화 서브 시스템으로 구성된다[4]. 미디에이터는 정보 통합을 수행하는 핵심 서브 시스템으로서 통합 스키마 관리 및 통합 질의 처리 기능을 제공하고, 랩퍼 서브 시스템은 통합되는 각각의 개별 데이터베이스를 관리하는 시스템과 미디에이터를 연동시키는 기능을 수행하며 연동 시스템의 종류에 따라 RDBMS 랩퍼와 ORDBMS 랩퍼 등으로 구분된다(무선 동기화 서브 시스템은 논문의 주제와 연관성이 없어 기능 설명을 생략함). 본 논문의 질의 처리 시스템은 미디에이터 서브 시스템에서 통합 질의 처리 기능을 전담하는 블록으로 정의되며, 미디에이터 서브 시스템의 구조는 그림 1과 같다.

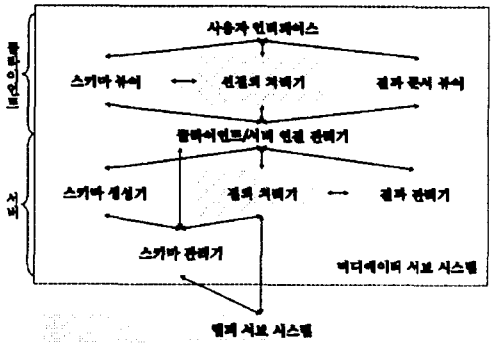

그림 1. 미디에이터 서브 시스템의 구조

미디에이터 서브 시스템은 클라이언트/서버 구조를 갖고 있으며, 질의 처리 블록 또한 수행하는 기능에 따라 클라이언트와 서버로 구분된다.

클라이언트의 선질의 처리기(query preprocessor)는 주로 사용자가 전달한 질의어를 구문 분석하고, 구문의 의미 분석을 실행하는데, 이는 통합 스키마의 정보를 이용해야 한다. 따라서 클라이언트에 있는 스키마 뷰어로부터 정보를 획득하여 구문 분석을 수행하게 된다. 또한, 서버측에 위치한 질의 처리기도 스키마 관리기와 협동 작업을 수행해야 하는데, 이는 사용자 질의어를 여러 랩퍼 서브 시스템에서 수행 가능한 지역 질의어들로 변경(transformation)할 때 지역 데이터베이스의 스키마 정보를 참조하기 때문이다.

서버의 질의 처리기는 여러 랩퍼 시스템들에서 반환된 문서들을 통합해야 하며, 이를 위해서 결과 관리기의 XML 문서 파싱 기능을 이용하여 문서 트리를 얻고, 복수개의 트리 병합을 통하여 사용자에게 반환될 최종 결과 문서를 생성할 수 있다.

또한, 서버의 질의 처리기는 결과 문서에 포함되는 실제 데이터를 여러 개의 지역 데이터베이스로부터 획득하기 위해서 랩퍼 서브 시스템에서 수행될 수 있는

지역 질의어들을 생성하며, 이를 랩퍼 서브 시스템들에서 수행시키기 위해 하나 이상의 랩퍼 서브 시스템들과 협력 작업을 수행하게 된다.

III. 질의 처리 시스템 설계

1. 질의 처리 단계

제안된 질의 처리기는 사용자 인터페이스를 통해 전달된 질의어를 처리하여 결과 문서를 반환하기 위해서 그림 2에서와 같이 단계별로 질의를 처리한다.

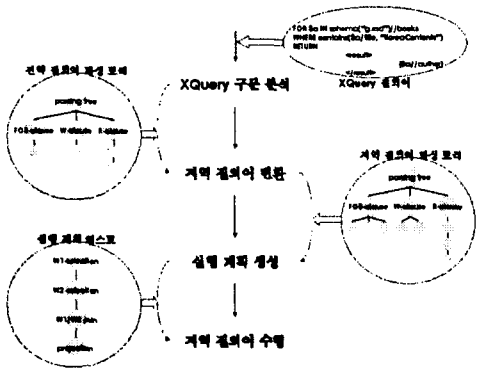

그림 2. 질의어 처리 단계

질의어 구문 분석 과정에서는 전달된 질의어가 올바른 구문인지를 분석하기 위한 파싱을 수행하고 파싱 트리를 생성하며, 이 파싱 트리를 이용하여 의미 정당성 검사를 수행한다. 의미 정당성 검사 항목들은 전역 스키마, 엘리먼트, 엘리먼트와 엘리먼트의 관계, 프레디키트의 연산자 호환성, 사용자 권한 등이다.

파싱 트리에 관리되는 질의어는 가상의 통합 스키마를 참조하는 형태이지만, 실제 데이터를 저장하고 있는 곳은 랩퍼 서브 시스템별로 연결된 지역 데이터베이스이다. 따라서, 질의를 실행시키기 위해서는 가상 통합 스키마에 대한 질의를 실제 지역 스키마를 참조하는 질의어로 변경해야 하며, 이는 질의어 변환 과정에서 처리된다. 변환의 결과는 수정된 파싱 트리가 된다.

실행 계획 생성 과정에서는 수정된 파싱 트리를 이용하여 논리적인 실행 계획을 생산하는데, 이 과정에서 파싱 트리의 노드와 구조를 관계 연산자들(relational operators)을 이용하여 관계 대수식(relational algebra)으로 변경한다[5]. 이후에 관계 대수식을 좀 더 효율적인 실행 계획으로 수정하는 기본적인 최적화(optimization)를 수행한다. 마지막으로 지역 질의어 수행 과정에서는 실행 계획으로부터 각 랩퍼 서브 시스템들에서 수행되어야 하는 지역 질의를 분류하고, 이들을 동시에 실행시킨 후, 관계 대수식 구조에 따라 결과 문서를 생성한다.

2. 질의 처리 시스템 설계

구문 분석, 질의어 변환, 실행 계획 생성, 질의어 수행 등을 처리하기 위한 질의 처리기는 그림 3에서와 같은 블록들을 갖는 구조로 설계하며, 각각의 역할은 아래와 같다.

파서는 자바 파서 생성 프로그램인 JavaCC를 사용하여 파서 클래스를 구성하며, 파싱 트리에 대한 정당성 검사는 미디에이터 서브 시스템의 스키마 뷰어를 통해 실행한다. 질의어 변환에서 정규화(normalization) 처리기는 질의 실행의 효율성을 위해 파싱 트리에 존재하는 검색 조건을 CNF(conjunctive normal form) 형태로 변경 처리한다. 실행 계획 생성기는 관계 대수식 변환기를 통해 파싱 트리를 관계 대수식 트리로 변경하고, 이후에 최적화 처리기에 의해 질의 최적화 결과 트리를 생성한다. 마지막으로 실행기는 최적화된 관계 대수식 트리를 순회하면서 각각의 지역 랩퍼 시스템들에서 수행되어야 할 지역 질의어들을 쓰레드별로 실행시키고, 그 결과를 통합하여 XML 문서를 만들어 반환하게 된다.

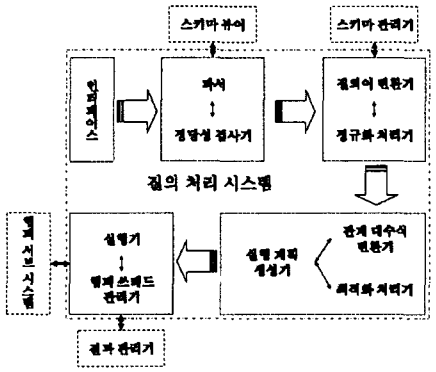

그림 3. 질의 처리 시스템의 구조

3. 지역 스키마 참조 질의어 변환

가상 스키마 참조 질의어를 통합된 지역 데이터베이스 스키마의 지역 질의어로 변환시키기 위해서는 가상 스키마와 지역 스키마의 매핑(mapping) 정보를 이용한

다. 매핑 정보의 구조는 [6]에 기술되어 있으며, 이를 기반으로 FLWR 형식의 질의어 변환 방법은 아래와 같다.

-[질의어 변환 방법]----------

(1) 가상 스키마 질의에서 변수 바인딩 정보 추출
(2) 매핑 정보에서 FOR/LET 절 정보를 새롭게 생성할 지역 질의로 복제
(3) 매핑 정보의 WHERE 절과 가상 스키마 질의의 WHERE 절을 결합(AND 연산자로 연결)
(4) 가상 스키마 질의에서 WHERE/RETURN 절에 기술된 변수를 (2)에 정의된 지역 스키마를 참조하는 변수로 변경
(5) (2)에서 복제된 FOR/LET 절 중에서 (4)에서 사용하지 않은 변수 삭제
(6) 변환된 파싱 트리 생성 및 반환

---

4. 질의어 최적화

질의 실행 계획의 효율을 높이기 위한 비용기반 최적화를 위해서는 통합에 참여한 지역 데이터베이스의 통계 정보를 파악할 수 있는 카탈로그 정보가 필요하지만, 현재 랩퍼 서브 시스템에서 이러한 정보를 제공치 못하고 있다. 따라서, 제안된 질의 처리 시스템에서 실행 계획의 최적화는 카탈로그 정보를 이용하지 않고 "Selection & Projection Push-Down" 방법을 적용한 최적화를 수행한다[6]. 제안 시스템에서 최적화 방법을 요약하면 아래와 같다.

-[최적화 방법]----------

(1) 관계 대수식 트리를 순회하면서 selection 연산자 노드들을 수집한다.
(2) projection 연산자 노드 정보를 수집한다.
(3) (1)에서 선택된 selection 연산자 노드를 그 노드의 서브 노드들을 순회하면서 가능한 단말 노드 바로 앞의 노드까지 옮긴다(단말 노드는 select 연산자를 적용할 문서를 나타내는 노드이기에 바로 상위 노드까지만 옮길 수 있음).
(3) (2)에서 선택된 projection 연산자 노드 정보를 기반으로 projection될 엘리먼트를 트리를 순회하면서 옮긴다. 이 때, 새로운 projection 연산자 노드가 필요한 경우, 이를 생성하고 대수식 트리에 추가시킨다.

---

5. 질의어 실행

가상 스키마 질의를 처리하는 실행 계획은 질의어 실행기를 통해 하나 이상의 지역 랩퍼 서브 시스템들로 전달될 지역 질의어들로 분해(decomposition)된다. 질의 처리 시스템의 실행기는 질의 처리의 효율을 높이기 위해 분해된 질의어들이 각 랩퍼 서브 시스템들에서 동시에 병렬로 처리할 수 있도록 랩퍼 쓰레드 관리기를 활용한다. 이후, 쓰레드 동기화를 통한 결과들을 수집하여 문서 집합을 구성하고, 이후 실행 계획 트리의 관계 연산자들을 처리하여 최종 질의 결과 문서를 생성한다.

IV. 구현 결과

제안된 질의 처리 시스템은 순수 자바 언어로 구현되었으며, JDK 1.3.1을 운영 환경으로 갖는다. 구현 결과는 그림 4와 그림 5에서 질의 처리 시스템의 동작 화면을 통해 확인할 수 있다.

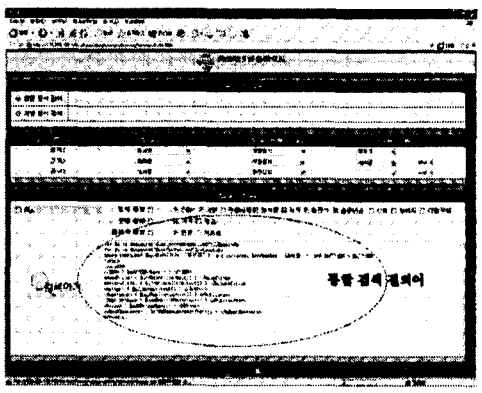

그림 4. 통합 검색 질의 화면

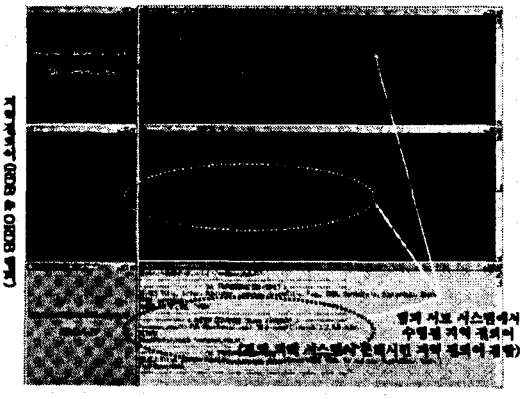

그림 5. 랩퍼 서브 시스템에서 지역 질의어 처리 화면

먼저, 그림 4는 가상 통합 스키마에 대한 통합 질의어를 표현하는 것이며, 그림 5는 실제 정보가 저장되어 있는 랩퍼 서브 시스템들에서 실행되는 지역 질의어를 보여주고 있다. 이 실험에서는 3개의 지역 데이터베이스를 통합한 통합 검색을 시행한 경우이다.

V. 결론

본 논문에서는 다양한 컨텐츠 데이터베이스들이 인터넷에 분산되어 있고, 이들이 서로 다른 데이터베이스 시스템들에서 관리되는 환경에서 통합 데이터베이스 검색 서비스를 지원하는 DataBlender 시스템의 핵심 요소인 질의 처리 시스템의 구조와 상세 설계 내용 그리고 구현 결과를 기술하였다.

향후에는 앞서 3장에서도 언급했듯이 통합 검색 성능을 향상시키기 위한 질의 최적화 방법에 대한 개선과 비용기반 최적화 알고리즘을 적용시킨 연구가 진행되어야 한다.

참 고 문 헌

[1] Chitanya Baru, et al, "XML-Based Information Mediation with MIX," Proceedings of ACM SIGMOD, Philadelphia, PA, 1999.

[2] Xachary G. Lves, et al, "Adaptive Query Processing for Internet Applications," IEEE Data Engineering Bulletin, Vol. 23, No. 2, 2000.

[3] 이경하 외 4인, "XMF: XML기반 분산 이질 정보 자원의 통합 프레임워크", KDBC2000 학술발표논문집, pp.262-270, 2000.

[4] 이미영 외 2인, "XML Schema기반 정보 통합 시스템 설계: DataBlender", 한국컨텐츠학회논문지, 제2권, 제2호, pp.36-41, 2001.

[5] Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer Widom, Database System Implementation, Prentice Hall Internaltion Inc., pp.329-422, 2000.

[6] 김병섭, 이미영, "통합 스키마 정의를 위한 XQuerySD", KDBC2002 학술발표논문집, pp.159-163, 2002.