생물 정보 저장용 XML 데이터를 위한 유연한 RDB 스키마 생성 규칙

정석훈○, 박성준, 한동수
한국 정보 통신 대학교
{jsh○, psj, dshan}@icu.ac.kr

A Flexible RDB Schema Generating Rule for Biological XML Data

Suk-hoon Han○, Sung-jun Park, Dong-su Han
Information and Communications University

요 약

유전자, 단백질 등의 생물정보를 이용하는 여러 툴은 효율성의 극대화를 위하여 각각의 시스템에 맞는 데이터베이스 스키마 구성 및 필요한 정보의 선택적 저장이 필요하다. 하지만 구조 복잡성, 동일한 객체 데이터의 분산 등, 생물 정보 XML의 일반적인 특성 때문에 기존의 XML정보 저장 기법으로는 유연한 데이터베이스 스키마구성에 한계를 지닌다. 이 때문에 생물정보 XML로부터 로컬 데이터베이스를 구성하는 과정은 1:1파서를 구현하여 진행하고 있어 많은 시간과 비용이 소모된다. 본 논문에서는 생물정보 XML 의 특성과 그에 따른 유연한 RDB 스키마 구성의 제약에 대해 논하고, 이를 극복한 자유로운 RDB 스키마 구성을 위한 규칙을 소개한다. 본 규칙은 사용자가 원하는 RDB 스키마를 구성하여 생물정보 XML의 데이터를 저장하게 해주며, SQL 형태를 따르고 있어 사용자에게 익숙하다. 또한 분산된 생물정보 XML의 통합에도 유리하다.

1. 서 론

최근 생물정보학은 정보과학 기술을 이용하여 생명과학을 지원하는 하나의 학문으로서 그 역할을 견고히 하고 있다. 생물정보학자는 생물실험 정보를 모아 분석하고, 인지하기 쉬운 형태로 표현하는 툴을 제공해주거나, 기존의 생물학 실험에서 벗어나 계산, 통계 등을 기초로 하여 새로운 정보를 얻어낼 수 있는 기법을 제안하는 등, 생물학 발견에 도움을 주는 방법론을 고안해내고 있다. 각각의 생물학 연구 그룹들은 자신들의 연구분야에 해당하는 특정 데이터베이스를 구축하여 인터넷상에 공개함으로써 여러 연구자들이 이를 자유롭게 이용할 수 있게 한다. 이러한 데이터베이스는 대부분 플랫파일이나 XML형태, 혹은 그 확장 포맷으로 정보를 배포하는데, 현재 인터넷 환경에 유리하고 스스로 자료의 의미를 정의 할 수 있는 XML포맷을 장려하고 있다.

생물 정보를 이용하는 시스템은 필요에 따라 로컬 데이터베이스를 구성하고, 배포된 데이터를 저장하여 사용한다. 특히 계산 통계학적 방법을 사용하는 대단위 작업을 수행할 경우, 데이터베이스의 이용이 시스템 효율성에 큰 영향을 주므로 시스템의 성능을 극대화 시킬 수 있는 데이터베이스 스키마의 구성이 필수 적이다. 또한 필요에 따라 연구자는 여러 생물 정보 데이터 베이스를 조합하여 의미 있는 정보만을 저장하여 사용한다. 때문에 생물정보학자들이 인터넷상의 생물정보를 이용하기 위해서는 분산되어있는 데이터베이스를 자신의 목적에 맞게 통합하고 정제하는 선행과정이 필수적이라 할 수 있다.

XML기반의 인터넷 생물정보 데이터베이스를 로컬 데이터베이스에 저장할 경우 일반적으로 관계 형 데이터베이스(RDB)가 주로 사용된다. 생물정보 XML 데이터를 로컬 RDB에 저장하기 위해 현재 제안된 XML-RDB 사상 기법[1,2]을 사용한다면, 기법과 생물 정보 XML의 특성상 사용자는 XML구조로부터 생성된 특정 RDB 스키마를 사용할 수 밖에 없다. 따라서 사용자가 시스템 성능향상 등의 이유로 로컬 RDB 스키마를 변경하고자 한다면, 사용자는 원 XML파일의 데이터를 수용하면서 자신의 목적에 맞는 RDB 구조를 만든 후, 일일이 XML을 파싱하여 데이터를 저장시켜야 한다. 이와 같이 생물정보 XML로부터 로컬 RDB를 구성하는 과정은 1:1파서 구현이 필수적이기 때문에 많은 시간과 비용이 소모된다.

본 논문에서는 XML기반 생물정보를 수용하는 RDB 스키마의 유연한 정의를 위한 규칙을 소개한다. 이를 위하여 RDB 스키마 구성 시 고려돼야 할 생물정보 XML의 특성과 관련 연구 또한 살펴본다. 본 규칙은 사용자가 원하는 RDB 스키마를 구성하여 생물정보 XML 데이터를 저장하게 해주며, SQL 형태를 따르고 있어 사용자에게 익숙하다. 또한 분산된 생물정보 XML의 통합에도 유리하다.

2. 관련연구

현재까지 많은 XML연구자들에 의하여 XML 데이터를 RDB에 저장하기 위한 다양한 기법들이 소개되어 왔다.[1,2] 하지만 그 대부분은 XML과 RDB의 1:1 사상으로 XML 구조를 유지하는데 목적을 두어 자유로운 RDB 스키마 구성에 제약이 따른다. XML의 구조를 유지하는 기법은 차후 RDB로부터 XML을 재구성할 때 강점이 있지만, 시스템의 효율성을 위한 자유로운 RDB 구조 구성에는 제약이 있다. 몇몇의 연구는 데이터베이스 구성 시 유연한 구조 변경을 허락하지만[3] 여전히 특정한 XML구조에 영향을 받는다. 특히 생물 정보 XML의 경우, XML의 엘리먼트 이름 외에 에트리뷰트 이름까지 참조해야만 그 자료의 의미가 정확히 파악되는 구조가 빈번해,