韓國컴퓨터情報學會 論文誌
第14卷 第6號, 2009. 6.

2009-14-6-4-4

하이브리드 데이터베이스 기반의 4단계 레이어 계층구조에서 메타규칙을 적용한 질의어 수행 모델에 관한 연구

오 염 덕*

A Study of Query Processing Model to applied Meta Rule in 4-Level Layer based on Hybrid Databases

Ryum-Duck Oh*

요 약

웹을 통한 생물 데이터 접근 방식은 많은 과학자들에게 대화식으로 서로 다른 형식의 생물 데이터베이스 내용을 검색할 뿐만 아니라, 한 데이터베이스에서 다른 분자생물 데이터베이스로의 연결을 위한 강력한 도구를 제공한다. 본 논문에서의 생물 개념 모델은 생물 데이터 제어를 위한 4가지 통합 레이어를 기반으로 각 생물 데이터 소스 간의 연관성에 따른 규칙 속성을 적용하고 데이터 소스 중에 관심 대상이 되는 개체를 표현하여 하이브리드 생물 데이터 모델을 구성하였다. 특정 사용자의 응용 서비스 요구가 발생하면 해당 생물 데이터베이스와 웹 서비스를 통한 데이터 소스로부터 정보를 획득한다. 본 논문에서는 통합 레이어를 기반으로 웹 데이터 소스 상에서 정보를 탐색하기 위해 메타 규칙을 적용한 질의어 처리 모형과 수행구조를 정형화하였다.

Abstract

A biological data acquisition based on web has emerged as a powerful tool for allowing scientists to interactively view entries form different databases, and to navigate from one database to another molecular-biology database links. In this paper, the biological conceptual model is constructed hybrid biological data model to represent interesting entities in the data sources to applying navigation rule property for each biological data source based on four biological data integrating layers to control biological data. When some user's requests for application service are occurred, we can get the data from database and data source via web service. In this paper, we propose a query processing model and execution structure based on integrating data layers that can search information on biological data sources.

▸ Keyword : Biological DB, Query processing, Data Integrating layers, Meta Rule, Navigation

• 제1저자 : 오염덕
• 투고일 : 2009. 06. 10, 심사일 : 2009. 06. 13, 게재확정일 : 2009. 06. 25.
* 충주대학교 컴퓨터과학과 교수

Ⅰ. 서 론

사용자는 온토로지(ontology) 기반환경의 분산 데이터로부터 내용-의존적(context-dependent)인 동적 정보 추출을 위해 서로 다른 기반 환경의 서로 다른 형식의 데이터를 접근해야 하기 때문에 이질적 분산 데이터(heterogeneous distributed data)로부터 정보 추출 및 지식 획득을 위한 기법이 요구된다[1]. 정보 소스는 물리적으로 분산되어 있고, 데이터 규모 측면에서도 대용량으로 구성되어 있다. 따라서 데이터 분석을 위해 중앙 집중화 방식으로 모든 데이터를 수집하기는 쉽지 않다. 이에 따라 대량의 데이터를 전송하지 않고 다양한 데이터 소스를 운영하기 위한 효율적인 알고리즘이 필요하다[2].

생물 데이터 소스는 전 세계의 생물학자에게 실험 데이터와 연구 내용을 포함한 많은 양의 정보를 공유할 수 있도록 제공하고 있다. 이러한 결과에 따라, 이 분야에서의 주요 관심은 유전자 기능과 주요 요소, DNA 시퀀스에 관한 연구들이 진행되고 있다. 다양한 종류의 생물 정보에서 게놈 시퀀스를 연구하고 표현하기 위해서는 새로운 모델링 관점이 필요하다[3, 4].

이미 여러 논문에서 통합 생물 데이터 소스를 위한 전형적인 데이터 모델 구축에 대한 발표와 연구가 진행되어져 왔다. 특히 Wilkinson[5]은 클래스 온토로지와 관계성을 갖는 에지 구조를 사용하여 데이터 모델을 표현하였다. 이러한 의미를 갖는 관계성은 3가지 유형으로 ISA, HASA, HAS로 구성하였다. 최근의 생물 데이터의 양은 폭발적인 증가 추세에 있다. 이러한 생물 데이터의 대부분은 국제 뉴클레어타이드 시퀀스 데이터베이스 연합(DDBJ, EMBL, GenBank)과 특정 목적의 생물 데이터베이스인 PROSITE, EC-ENZYME, GDB, Reactome, UniProt, PIR, DIP, Pfam, PDB 등에 저장되어 있다[6]. 2007년을 기준으로 전 세계적으로 968개의 생물 데이터베이스가 존재한다[7].

다양한 생물 데이터 소스에 적용할 수 있고 통합 정보 처리를 위한 도구의 필요성이 증가하고 있다. 생물학자/과학자에게 필요한 기본적인 문제는 생물 개체에 대한 특정 인스턴스(instance)를 정확하게 식별하는 데 있다. 예를 들어, 생물 개체는 특정 유전자 및 단백질 등을 다양한 형식의 상호 연결된 생물 데이터 소스 간의 탐색을 통해 이러한 개체 인스턴스의 완전한 기능적 특성을 추출할 수 있어야 한다[8].

본 논문에서는 생물 데이터의 접근 관점에 따라 개별 단위의 클래스 구조에 의한 생물 데이터 레이어를 표현하고 공통 환경의 개념적 모델을 구성하였다. 또한, 생물 데이터 소스 간에 효율적인 접근을 위해 하이브리드 데이터베이스 접근 방식에 의한 질의어 처리 방식을 정형화하였고 실행 환경에 대한 모형을 제시하였다. 본 논문의 구성은 2장에서 생물 데이터 통합을 위한 4 단계 레이어 기반 데이터 모델을 정의하고, 생물 데이터 접근을 위한 네비게이션 모형을 제안한다. 3장에서는 생물 데이터 접근 및 검색을 위한 질의 수행모형을 제시하고, 4장에서는 제안 모형의 결과를 설명하고 향후 연구방안을 다룬다.

Ⅱ. 4단계 레이어기반의 데이터모형

생물학적 데이터의 다양성은 생물학자를 위해 생물 시스템의 구성 요소 간에 상호 작용을 연구하는 능력을 크게 개선시키고, 그리고 이러한 상호 작용이 시스템의 기능과 행위에서 어떻게 발생하는지를 제공해준다[6]. 데이터 통합을 위한 접근 방법으로는 크게 2가지로 데이터 웨어하우징(data warehousing)과 데이터베이스 연합(database federation) 방법이 있다[9].

데이터 웨어하우징 접근 방법은 이질적(heterogeneous) 분산 데이터 소스로 부터 데이터를 획득하고 공통의 데이터 구조로 사상시켜 중앙 공간에 저장시키는 방식이다. 웨어하우스에 저장된 데이터는 개별 데이터 소스의 내용에 영향을 받기 때문에 웨어하우스의 내용을 주기적으로 변경하는 것이 필요하다. 하지만, 대다수 정보 저장 공간의 예에서 개별 데이터 소스의 변경된 내용을 찾고 검색할 수 있는 기법을 사용하지 않는 한 적용하기 힘들다. 이러한 사실은 데이터 소스의 원자성(autonomy)에 위배되기 때문이다. 데이터 통합을 위한 웨어하우징 접근 방식은 또 다른 중요한 문제점이 있다.

예를 들어, 과학적 발견을 위해 여러 관점으로 동일 데이터에 대한 분석이 필요한데, 데이터 웨어하우징 방법은 시스템 상의 모든 사용자들에게 단일 공통의 온토로지에 의존한다. 이러한 온토로지는 데이터웨어 하우징으로 설계된 부분으로, 각 사용자는 공통의 어휘와 공통의 인터페이스를 사용하여 질의가 이루어진다.

데이터베이스 연합 접근 방식은 필요한 정보에 대해 질의어를 사용하여 데이터 소스로부터 직접적으로 결과를 구한다. 그러므로 데이터 소스의 내용으로부터 얻어진 질의 결과는 최신의 데이터를 획득할 수 있다. 중요한 점은, 데이터베이스 연합 접근 방식은 사용자가 요구하는 분산된 원자 값을 갖는 정보 소스로부터의 데이터를 기반으로 그들 자신의 온토로지에 적합한 응용물(application)에 즉시 적용할 수 있다[10].

본 논문에서는 인간 유전자 응용 시스템을 위한 데이터 통합에 있기 때문에 모든 생물 데이터 소스에 대한 가상의 개념적인 동일 생물 계층 구조를 사용하고 통합된 인터페이스를 통해 접근 가능하도록 구성하여 사용자에게 다중 원자성 데이터 소스로부터 통합 데이터 적용에 융통성을 줄 수 있도록 설계하였다. 이러한 구조는 시스템 개발 시에 내부적인 메타규칙을 적하여하여 네비게이션 연결과정이 이루어진다. 따라서

본 논문에서는 모든 사용자들에게 단일 공통의 온토로지인 생물 개념적 계층구조를 적용하기 때문에 외연측면에서는 데이터 웨어하우징 기법을 사용하지만, 질의어를 사용하여 데이터 소스로부터 직접적으로 결과를 구하는 방법을 선택하였기 때문에 내부적으로는 데이터베이스 연합 방식을 사용한다. 따라서 두 가지 유형의 통합 방식의 특성을 이용하기 때문에 하이브리드 데이터베이스 통합 방식을 적용하였다. 본 논문에서 기반으로 사용하는 하이브리드 방식은 공통의 인터페이스를 통해 분산 데이터소스로 획득한 데이터를 중앙공간에 저장하여 활용하는 방식과 필요한 정보를 질의어를 사용하여 직접적으로 데이터소스로 부터결과를 구하는 방식을 혼용하는 구조로 사용하는 방식이다.

1. 의미적 통합 레이어 구조

생물 데이터 통합 시스템은 통합된 질의 인터페이스를 사용하여 다중 원자성, 분산 환경 및 이질적 데이터 소스로부터 사용자에게 정보를 효율적으로 접근 할 수 있도록 제공해야 한다[11]. 관심 데이터 소스는 그 자신이 원자성을 만족해야 하고 적합한 연산이 적용된다. 결과적으로, 데이터 소스 상에서 실행되는 연산의 범위와 대화식으로 적용되는 실제 형식은 매우 다양하다. 그러므로 데이터 소스에서 연산 제약을 만족하는 요구한 정보를 추출하기 위한 전략이 필요하다[12, 13, 14].

본 논문에서는 생물 데이터 통합을 위한 설계와 행위를 구성하기 위하여 Castillo[12]의 3 레이어(Physical layer, Onto- logical layer, Presentational layer)를 확장한 데이터 통합 4-레이어를 제안한다. 각 레이어는 생물 데이터 통합 시스템의 설계와 행위에 대한 개념적 표현을 나타낸다. 생물 데이터 통합 레이어는 네트워크(Network) 레이어, 선택(Preferential) 레이어, 온토로지(Ontology) 레이어와 추상(Abstraction) 레이어로 구성된다.

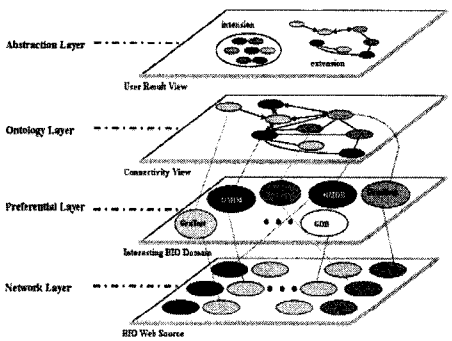

그림 1. 데이터 통합 4-레이어 구조
Fig. 1. Data integration: Structure of 4-layers

네트워크 레이어는 데이터 통합 시스템과 각 데이터 소스 간의 정보 전송 및 통신을 담당한다. 선택 레이어는 인간 유전자를 지원하는 관심 대상의 데이터 소스 그룹으로 한정하여 영역을 구성하였다. 본 연구에서 적용한 온토로지를 활용하는 방법은 이미 많은 연구들을 통해 개념들이 확립되어 왔다[15,16]. 온토로지 레이어는 정의된 온토로지를 위한 기법과 해당 사용자 질의를 질의 실행 플랜(query execution plans)으로 변환을 위한 메타 규칙을 갖는 질의어를 제공한다. 추상 레이어는 시스템에서 사용자에게 대화식으로 제공할 수 있도록 인터페이스를 제공한다. 본 논문에서는 일반적인 3단계 레이어 구조[12]에서 선택 레이어를 추가한 4-레이어 구조를 사용함으로서 분산 데이터소스의 탐색공간 범위를 줄일 수 있도록 새로운 레이어 환경을 도입하였고, 이를 기반으로 온토로지 레이어상의 대상 데이터의 범위를 설정하여 검색할 수 있도록 지원하였다.

2. 데이터모형: 추상 레이어와 선택 레이어

웹을 통해 획득할 수 있는 생물 데이터의 유형은 매우 다양하다. 또한, 생물 데이터 소스에 존재하는 데이터의 속성 자체가 매우 복잡할 뿐만 아니라, 데이터 특성 간의 연결 관계도 복잡한 유형으로 구성되어 있다. 따라서 생물 데이터 소스에서 발생하는 각종 데이터의 유형에 따라 이를 적절하게 표현하고 분류할 수 있는 적합한 형식의 데이터 모델이 요구되어진다. 이미 여러 논문에서 이를 위한 데이터 모델[3, 4, 5]이 연구되고 발표되어 왔으며, 통합 생물 데이터 소스를 위한 모델을 부분적인 구조에 한해 제시하여 왔지만, 특정 목적을 만족하는 정형화 된 모델은 거의 없는 실정이다. 따라서 본 연구에서는 관심 있는 데이터 소스의 도메인을 대상으로 생물 개체의 속성에 따라 개체를 분류하고 개체간의 적합한 의미를 갖는 관계성을 부여하여, 전체적인 개체 구조에 따른 속성을 정형화하여 개념적 생물 계층구조로 표현한 데이터모델(BioDaMo: Biological Data Model)을 구성하였다.

생물 계층 구조는 기본적으로 생물 개체, 생물 속성, 연관 관계로 구성된다. 생물 개체는 생물 데이터에 대한 공통의 의미적 속성을 공유할 수 있도록 단위 속성으로 그룹화하여 표현한다. 생물 데이터소스에서 발생하는 각 단위의 특성은 개체 형식으로, 그 구조는 다음과 같이 정의되어진다.

- 현 개체에 대한 다른 생물 개체와의 연관성
- 생물 개체에 대한 접근 및 조작을 위한 연산
- 생물 개체의 검색 및 제어를 위한 경험 규칙 적용
- 생물 데이터 소스 간의 네비게이션을 위한 경로 메타 규칙의 적용

개념적 계층구조 상에서의 관계성은 생물 개체간의 구문적, 구조적 관계성을 표현한다. 본 논문에서의 생물 개체간의 관계성은 크게 2가지 유형의 사이트-내부 관계성(intra-site relationship)과 사이트-외부 관계성(inter-site relationship)으로 구분한다. 사이트-내부 관계성은 사이트 내에서 발생되는 관심 대상의 생물 개체를 개념적 계층구조에서 서로 연관성을 갖는 생물 개체간의 연관성을 표현하기 위해 사용한다. 본 논문에서 제안된 관계성은 Consist_of, Is_A, Association_with, Supported_By, Has_A 와 같이 5가지 유형의 관계성을 사용하여 개체간의 연관성을 표현한다. 사이트-외부 관계성은 사이트 간에 접근하는 정보 형태나 속성에 따라 사이트 간의 링크 구조를 표현한 네비게이션 계층구조 상에 존재하는 링크 관계성이다. [그림 2]는 관계성을 대상으로 설정한 관심 대상이 되는 생물 데이터 소스에 대한 속성을 추상 레이어를 기반으로 개념적 계층 구조를 표현한 구조이다.

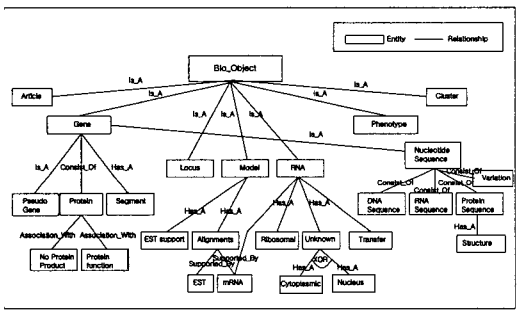

그림 2. 추상 레이어상에서 개념적 생물 계층구조
Fig. 2. Conceptual biological hierarchy in abstraction layer

[그림 2]에서 Bio_Object는 모든 생물 데이터에 대한 기본 자료형(data type)을 갖는 개체로 구성되며, 기본 개체로는 Gene, Cluster, Article, Locus, Nucleotide, Protein, Model, RNA 등으로 구성된다. 각 개체는 그 개체의 특성을 표현하기 위한 속성을 갖는다. 이를 개체의 속성이라 부르며, Gene 개체는 Gene_ID, Gene_Description 등을 가질 수 있다. 본 논문에서는 개체에 존재하는 속성이 워낙 많기 때문에 [그림2]에서는 상세하게 표현하지 않았고, 데이터 소스의 범위를 전체 생물 데이터 소스에서 인간 유전자에 관련된 데이터 소스로 한정시키고, 이를 선택 레이어 영역으로 적용된다.

생물 데이터 소스에서는 데이터의 특성에 따라 다양한 유형의 관계성이 존재하며, 생물 계층구조에서 사용하는 개체 간의 관계성은 5가지 유형으로 정의할 수 있다.

- Is_A 관계성은 생물 계층구조 상에서 상위개체의 속성을 하위 개체에 상속하여 사용할 수 있으며, 하위개체는 자신의 속성 정보와 상위 속성 정보를 공유한다.
- Consist_of 관계성은 생물 계층구조 상에서 상위개체의 속성을 통해 하위개체를 새롭게 생성할 수 있지만, 상속의 성질의 제공하지 않으며 변환된 속성 정보를 가질 때 사용한다.
- Association_with 관계성은 상위개체를 통해 세부화된 하위 개체를 구성할 수 있으며, 모든 하위개체를 사용하여 하나의 단위 상위개체를 표현할 수 있다. 반드시 하위개체를 가질 필요가 없다.
- Supported_By 관계성은 상위개체가 반드시 복수개의 하위개체를 가져야 하며, 하위개체는 세부화 된 구성원소를 표현한다.
- Has_A 관계성은 상위개체가 가질 수 있는 구성원소로 여러 개의 세부화 된 원소개체(element entity)를 가질 수 있으며, 개체가 가져야 하는 모든 구성원소를 표현할 필요는 없다. Association_with 관계성과 Has_A 관계성의 포함 관계는 Association_with ⊂ Has_A를 만족하며, Supported_By 관계성과 Has_A 관계성의 포함 관계는 Supported_By ⊂ Has_A를 만족한다.

[표 1]은 본 논문에서 제안한 모형(BioDaMo)과 Wilkinson [5]이 제안한 모형(BioMOBY) 간의 관계성 표현을 비교한 내용이다. 기존 생물계층구조 모형의 3개의 관계성 유형을 5개의 관계성으로 확장함으로서 생물 데이터가 가지는 특성을 의미적 속성 관계로 구성함으로서 다양한 관계표현이 가능하다.

표 1. 생물 데이터모형에서 관계성비교
Table 1. Comparison of Relationship in the Biological Data Model

	BioDaMo	BioMOBY
상속관계의 개체를 표현	Is_A	ISA
개체간의 구성요소 관계 표현	Has_A Support_By	HAS(1대n 관계) HASA(1대1 관계)
변환속성 관계표현	Consist_of	없음
한정구성 관계표현	Association_with	고정 n 값을 갖는 HAS

변환속성관계는 개체간의 변환된 속성 관계를 가질 때 사용하며, "Pseudo Gene"은 상위개체의 속성을 상속받지만, "Protein" 개체는 "Gene" 개체의 부분 속성을 새로운 유형으로 변환된 값을 갖는 속성으로 표현하기 때문에 Consist_of 관계성으로 표현한다. 한정구성관계는 반드시 고정된 속성을 갖는 관계를 표현할 때 사용한다.

3. 네비게이션 모형: 온토로지 레이어

생물 데이터 소스의 범위는 전체 생물 데이터 소스의 내용 중 사람의 질병과 관련된 데이터를 대상으로 적용한다. 각 생물 데이터 소스 정보를 가진 사이트는 해당 정보의 성격에 따라 관련 세부정보를 가진 다른 데이터 소스 방향으로 하이퍼링크(hyperlink)를 통해 연결 관계를 갖는다.

〔정의 1〕 생물 데이터 소스의 범위

> 생물 데이터 소스는 BS로 정의하며, 전체 생물 정보를 보유한 사이트에서 관심대상이 되는 정보, 영역 및 그룹을 가진 사이트로 한정한다. GenTest, MMDB, ..., GDB는 생물 데이터 소스에 존재하는 고유한 데이터베이스 이름이다. 생물 데이터 소스의 범위는 다음과 같이 정의한다. BS={GenTest, MMDB, OMIM, LocusLink, PubMed, GeneBank, UniGene, Regseq, GDB}

〔그림 3〕은 생물 데이터 소스간의 상호 연결성을 위한 네비게이션 관계를 표현하였고 온토로지 레이어 기반 환경에서 생물 데이터 소스 간의 의미적 연결 관계를 구성하였다. 각 노드(타원형 기호)는 생물 정보를 가진 데이터 소스 사이트를 의미하며, 각 데이터 소스 간에는 상호 연결 관계성을 형성하고, 정보 링크(information link) 속성에 따른 생물 데이터 소스 간의 방향성을 가진다.

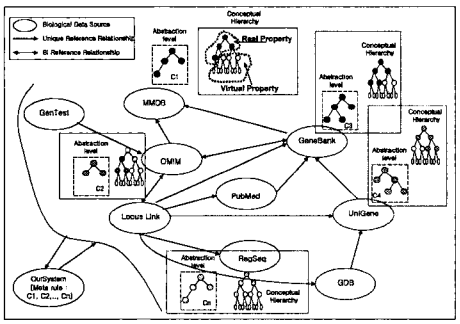

그림 3. 온토로지 레이어상에서의 네비게이션 구조
Fig.3. Navigation Structure in ontology layer

예를 들어, LocusLink와 OMIM 간에는 상호 필요 정보를 연결하는 관계로 양방향(Bi Reference Relationship) 관계성을 갖는다. OMIM 과 GenTest의 관계는 단방향(Unique Reference Relationship) 관계성을 가지며, OMIM에서 GenTest가 보유한 관심 정보를 추출할 수 있다. 단, 모든 생물 데이터 소스를 갖는 사이트는 자신의 고유한 정보를 갖고 있으며, 검색이 필요한 경우 정보링크에 의해 연관된 사이트를 참조한다. 해당 사이트에 존재하는 정보를 접근하는 방식은 〔그림 4〕와 같이 2가지 유형으로 구분할 수 있다.

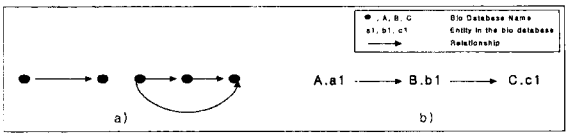

그림 4. 정보접근방식에 따른 경로 속성
Fig. 4. Pathway property on information access method

〔그림 4〕에서 a) 방식은 관심 대상 정보를 가진 데이터 소스를 기준으로 상호 연결 관계를 통해 또 다른 생물 데이터 소스를 참조하고, b) 방식은 관심 정보를 가진 개체를 기준으로 데이터 소스 간의 연결 관계에 따라 다른 사이트의 정보를 참조한다.

생물 정보가 저장되어 있는 사이트를 네비게이션하는 방법은 매우 다양하다. 이러한 생물 데이터 소스 간의 연결 방식은 의미적 연결성(semantic link)과 기능적 연결성(functional link) 방식으로 구분할 수 있다. 접근하는 정보의 의미적 연결성 구조에 따라 여러 가지 유형의 정보 링크(Information link)를 가지며, 데이터 소스 접근 방향에 따른 기능적 연결성은 양방향 참조 연결성(Bi Reference Relationship)과 단일 참조 연결성(Unique Reference Relationship)으로 구분한다. 이러한 연결성 방식을 사용하여 관계성을 구분하는 이유는 의미적 정보 속성을 갖는 네비게이션 경로 구조를 정확하게 예측할 수 있으며, 생물 데이터 소스 간의 정보를 검색을 위해 접근 방향성을 가진 메타 정보(meta information)를 사용하고 내부적인 경로 구조를 정의하여 표현한다.

4. 정보그룹과 의미적 링크구조

정보링크(information link)는 각 데이터 소스의 생물 계층 구조상에서 관심 대상이 되는 정보를 포함한 그룹의 내용과 해당 그룹에서 다른 데이터소스로의 연결하는 하이퍼링크를 결합하여 데이터 소스간의 연결 관계를 기능적, 의미적 연결 관계로 표현한다. 이러한 기법은 생물 데이터 소스 간에 메타 정보에 의해 결정된 경로를 이용하여 정보를 추출하는 방법을 제공한다.

〔정의 2〕 정보링크(IT: Information Link)

> 정보링크는 생물 데이터 소스간의 네비게이션 관계성을 표현하기 위한 의미를 갖는 정보집합 또는 정보그룹으로 구성된 자료형을 의미한다. 정보링크는 BS의 부분집합이다.
> IT={O, G, L, P, N, S, A, C, V}, 또한 정보링크와 생물 데이터 소스와의 관계는 IT ⊆ BS를 만족한다(단, O : Synonym, G : Gene, L : Locus position, P : Protein, definition, product, N: Nucleotide, S : Protein Structure, A : Article, C : Cluster, V : Sequence Variation를 의미).

각 사이트에서 연결되는 링크 관계는 정보링크의 사상(mapping) 형태에 다양한 형태가 존재하며 이러한 대상 개체에 대한 의미적 연결을 위한 각 생물 데이터 소스에 속한 멤버쉽을 갖는 관계성을 표현할 수 있다. 이러한 사상관계는 슈퍼링크(Super Link)와 비트링크(Bit Link) 유형으로 구분한다.

〔정의 3〕 슈퍼링크와 비트링크

각 BS 간에 연결되는 정보 링크는 의미적인 연결 관계에 따라 1 대 1, M 대 N의 대응관계를 갖는다. 생물 데이터 소스 간의 대응 관계가 M 대 N의 링크 관계를 가질 때, 이를 슈퍼링크라 부른다. 또한, 생물 데이터 소스 간의 대응 관계가 1 대 1의 링크 관계를 가질 때, 이를 비트링크라 부른다. 각 대응 관계는 각 생물 데이터 소스에 존재하는 개체의 멤버쉽을 전제로 한다. 따라서, IT를 정보링크라 하고 ST를 슈퍼링크, BT를 비트링크라고 하면, ST ⊆ M(사이트_1) X N(사이트_2), BT ⊆ 1(사이트_1) X 1(사이트_2)를 만족한다. 그러므로 생물 데이터 소스간의 포함관계는 BT ⊆ ST ⊆ IT 를 만족한다.

〔그림 5〕는 각 생물 데이터 소스간의 의미적, 기능적인 네비게이션 구조를 표현하였다. GenTest는 해당 사이트에 소속된 관심 데이터 항목인 4가지 정보링크(O, L, G, P)을 가지며, 이러한 정보링크를 사용하여 관련 OMIM 데이터 소스의 3가지 정보링크(G, L, O)를 통해 해당 정보그룹을 접근할 수 있다. 두개의 데이터 소스 간에는 슈퍼링크 관계가 존재한다. LocusLink와 UniGene 간에는 비트링크 관계가 존재한다.

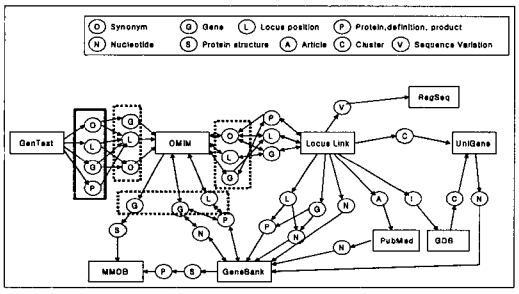

그림 5. 슈퍼링크와 비트링크의 연결구조
Fig. 5. Connection structure with super-link and bit-link

GenTest에서 OMIM으로의 정보획득은 접근방향이 단방향이기 때문에 가능하지만 반대의 정보접근은 불가능하다.

III. BioQL 질의수행 모형

특정 생물 데이터 소스에 존재하는 정보나 자신의 사이트에서 얻을 수 없는 정보는 하이퍼링크를 사용하여 다른 생물 사이트로 연결하고 필요한 정보를 해당 사이트로부터 검색하여 구한다.

1. BioQL 질의모형

웹기반 환경에서 생물 데이터를 검색하기 위한 BioQL 질의어는 웹을 접근하는 측면에서 기존 웹 기반 질의어와 유사한 특성을 지닌다. 본 논문의 BioQL 질의모형은 웹 접근을 위한 기능과 생물데이터의 4-레이어를 기반으로 하는 네비게이션을 통한 검색기능을 갖도록 구성하였다. 또한, 질의언어에서 데이터 소스간의 다양한 관계성을 표현할 수 있도록 설계하였다. BioQL 언어는 생물 개념 계층구조와 경로지향 네비게이션 환경에서 질의를 만족하는 생물 데이터소스를 검사하여 질의 수행을 위한 최적의 접근경로를 설정한다. 생물 데이터에 대한 통합스키마는 데이터베이스의 의미적 속성을 명시적 형식으로 표현하며, 구조적 속성, 연산적 속성, 데이터베이스의 모든 객체와 관련된 규칙을 포함한다.

생물 개념 계층구조는 생물 데이터소스의 의미적인 내용을 정의하지만 실질적으로 사용자나 DBA에게 생물 데이터소스 간의 내부적 관계성(inter-relationship)을 완전하게 전달하지는 못한다. 따라서 질의결과에 대한 보다 다양한 정보를 제공하기 위하여 생물 정보를 가진 개체와 인스턴스를 표현하기 위하여 두 가지 유형의 결과형식을 갖도록 설계하였다. 질의 결과의 외연(스키마 구조를 나타냄)을 표현하는 N-Diagram (Navigation Diagram)과 내포를 표현하는 I-Table(Instance Table)로 구성되는 질의 결과형식을 제공한다.

N-Diagram은 생물 데이터 소스간의 네비게이션 관계성을 시각적 그래픽 형식으로 표현하고 노드와 링크형식의 네트워크 구조를 가진다. 또한, I-Table(Instance Table)은 개체에 속한 인스턴스와 네비게이션을 통해 추출된 정보를 표현한다. 본 논문에서의 BioQL 질의어 형식은 〔그림 6〕과 같다.

```
RangeOfContext  〔range_variables〕
PathDB  〔association_relationships〕
      〔Link  functional_path〕
Where  search_condition
Retrieve  result_properties ;
```

그림 6. BioQL 질의어 형식
Fig. 6. Query form on the BioQL query language

RangeOfContext 절에서는 생물 데이터소스의 범위변수를 설정한다. 범위변수를 설정하는 이유는 검색 대상이 되는 바이오 데이터베이스의 규모를 줄이기 위한 목적이다. 범위변수 설정을 위한 range_variables에 대한 기본형식은 〔형식1〕과 같다.

〔형식1〕 RangeOfContext 절의 range_variables 설정

```
range_variables ::=
 (Bio_Source_Name RangeVariable[,range_variables])*
```

RangeOfContext 절의 바이오 소스명에 대한 범위변수는 복수 개를 설정할 수 있다. 범위 변수를 생략하는 경우에는 LocusLink 바이오 데이터소스를 기본 바이오 데이터소스로 설정하고 LocusLink 로부터 탐색을 시작하여 질의어에서 탐색조건을 만족하는 다른 데이터소스로 네비게이션이 진행된다.

PathDB 절의 association_relationships 관계성은 범위변수로 설정된 바이오 데이터 소스를 참조하기위한 연결관계성을 표현하며, 관계성의 의미를 나타내기 위해 *, +, | 기호를 사용한다. 기본 형식은 〔형식2〕와 같다.

〔형식2〕 association_relationships의 설정

```
association_relationships ::=
    Src_Group { toggle_op association_relationships}*
Src_Group ::=
    RangeVariable {navi_op  Range_Variable}*
     | Src_Group
navi_op ::= '*' | '+'
toggle_op ::= '|'
```

〔형식2〕에 따라 관계식을 구성하면 다음과 같은 연산식을 작성할 수 있다.

〔관계식1〕 PathDB DataSrc1 * DataSrc2*DataSrc3

〔관계식2〕 PathDB DataSrc1 + DataSrc3

〔관계식3〕 PathDB (DataSrc1 * DataSrc2 * DataSrc3)
| (DataSrc4 + DataSrc5)

"*" 기호는 바이오 데이터소스 간의 직접적인 연결 관계를 표현하며, 〔관계식2〕에서 "+" 기호는 범위변수로 설정된 DataSrc1 바이오 데이터소스에서 시작하여 DataSrc3 까지의 간접적인 연결 관계를 설정하여 중간의 연결통로는 질의처리시에 메타규칙(meta rule)을 사용하여 연결되는 바이오 데이터소스를 설정되어 질의어 구조가 확장되어진다. 〔관계식3〕에서 "|" 기호는 좌측 관계식에 존재하는 바이오 데이터소스의 영역과 우측 관계식에 존재하는 바이오 데이터소스의 영역 중에서 가장 최적의 검색 영역으로 데이터를 추출하고자 할 때 사용한다.

Link 절은 연결 관계를 갖는 데이터 소스 간에 적용할 필요한 메타함수(meta function)를 표현한다. 메타함수는 메타규칙을 운용하기 위해 독립적으로 정의하여 사용하거나 별도의 내장함수를 활용한다. 규칙 관리자(Rule Manager)에 의해 메타규칙을 포함한 질의어로 확장된다. Where 절은 사용자가 원하는 바이오 데이터를 검색하기 위한 탐색조건을 기술한다. Retrieve 절은 검색 목표가 되는 바이오 데이터의 속성을 표현한다.

2. BioQL 질의수행 절차

BioQL 질의어를 사용하여 생물 데이터소스의 정보를 검색하기 위해서는 해당 질의 처리 작업을 수행하기 위한 절차가 필요하다. 질의어 처리 모듈은 웹을 제어하는 기능과 통합 환경에 적합하도록 질의어를 제어하는 기능이 요구된다. 질의어에서 사용되는 각 데이터소스 자체가 웹 환경을 기반으로 하는 대상 도메인이 되기 때문이다.

〔그림 7〕은 사용자 질의에 대해 생물 정보를 찾기 위한 질의처리 절차를 구성하였다. 특정 사용자가 필요로 하는 생물 정보를 찾을 경우 먼저 사용자 인터페이스를 통해 사용자 질의어 입력하면, 시스템에서는 질의 관리자(Query Manager)에 의해 BioQL 형식의 질의어를 분석한다.

질의어가 BioQL 형식을 만족하면 BioQL는 규칙관리자(Rule Manager)에 의해 적용된 네비게이션 구조를 갖는 메타규칙(meta rule)을 포함한 형식으로 질의어가 확장된다. 또한, 추론규칙은 질의어 내에서 필요한 경험적인 지식(heuristic knowledge)을 함수를 사용하여 적용한다. 질의관리자가 최적화된 질의어를 생성하면 규칙관리자에 의해 생성된 변환된 질의접근 플랜에 대한 질의실행 플랜을 생성한다. 시스템 매퍼(System Mapper)는 각각의 서로 다른 모듈들을 통합 운영하고 모니터링하는 데 사용한다. 이러한 통합 운영모듈은 웹추출기(Web Extractor)에게 질의실행 플랜에서 탐색조건을 만족하는 생물 데이터 소스를 요구한다.

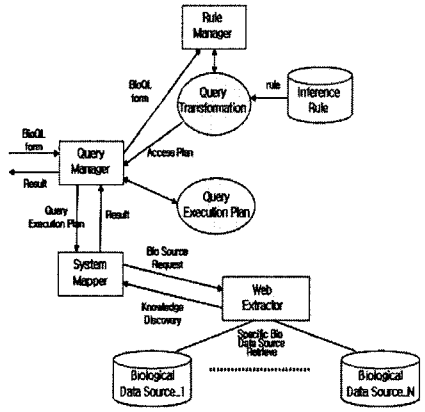

그림 7. 질의수행구조
Fig. 7. Query execution procedure

웹추출기를 통해 온토로지 기반환경에서 특정 데이터소스에 대한 네비게이션 과정이 진행된다. 질의 수행 결과는 외연(extension)과 내포(intension)로 구성된다. 외연은 시각적 형식의 데이터 소스간의 의미적 네비게이션 관계를 표현하고 내포는 네비게이션을 통한 생물 정보와 테이블 형식의 개체에 속한 인스턴스로 표현된다. 다음의 질의어 "Query"을 사용하여 질의 수행절차를 구성하였다.

〔Query〕 GenTest 바이오 데이터소스에서 Locus가 "IMJE_B"이고, "BRCA2" 암 중에서 "Brca2-Dss1-Ssdna"를 만족하는 유방암의 3차원 단백질 구조를 검색하라.

〔Query Expression〕

```
RangeOfContext   GenTest G,  OMIM O, MMDB M
PathDB  G*O*M
Where G(LocusID="IMJE_B"),
      M.Structure(Description="Brca2-Dss1-Ssdna")
Retrieve M.View
```

주어진 〔Query Expression〕는 규칙 관리자에 의해 메타 규칙을 적용하여 다음과 같은 형식으로 변환된다.

〔BioQL 변환〕

```
PathDB   GenTest(Locus)*OMIM(Locus)*
   MMDB(Gene.Nucleotide.ProteinSequence.Structure)
   Association_Path(GenTest*OMIM*MMDB)
Where    GenTest.Locus(LocusID="12190"),
         LocusLink.Gene.Nucleotide.ProteinSequence.
         Structure(Description="Brca2-Dss1-Ssdna")
Retrieve
   LocusLink.Gene.Nucleotide.ProteinSequence.
   Structure.View ;
```

Association_Path(GenTest*OMIM*MMDB) 함수는 데이터소스간의 물리적 연결관계를 설정하기 위한 네이게이션 설정에 사용된다. 질의수행에 따라 결과는 외연을 표현하는 N-Diagram와 내포를 표현하는 I-Table로 구성된다.

N-Diagram은 연관 바이오 데이터소스와 탐색 대상으로 사용한 세부구조(structure)를 표현한다. 〔Query〕에 대한 N-Diagram 결과는 〔그림8〕과 같이 네비게이션 링크구조와 생물 개념 계층구조로 구성된다.

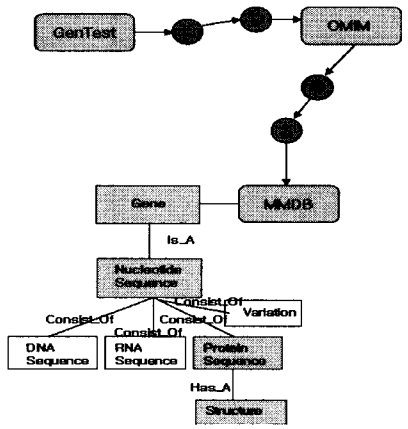

그림 8. 질의수행결과: N-Diagram
Fig. 8. The result of query execution: N-diagram

〔그림 9〕의 I-Table은 질의에 대한 Protein 결과 Structure View에 대한 정보를 획득한다.

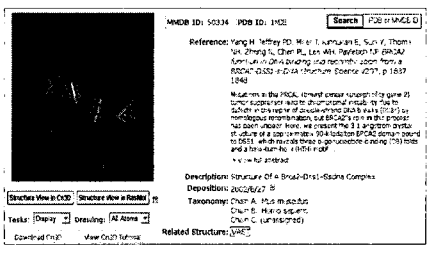

그림 9. 질의수행결과: I-Table
Fig. 9.The result of query execution: I-Table

IV. 결 론

본 논문에서는 바이오 데이터에 대한 여러 유형의 의미 있는 관계성을 표현하기 위해 하이브리드 형식의 바이오 데이터

통합 시스템에서 바이오 데이터의 설계와 행위를 포함하는 4개의 통합 바이오 레이어 구조를 제안하였다. 각 레이어는 바이오 데이터 통합 구조의 행위와 데이터 속성에 따른 설계부분을 포함한다. 네트워크 레이어는 데이터 통합 시스템과 데이터 소스간의 통신을 담당한다. 선택 레이어는 인간 유전자를 지원하는 관심 대상의 데이터 소스 그룹으로 구성된다.

본 논문에서는 온토로지 레이어상에서 바이오 데이터 소스간의 연결 관계와 구조를 표현하고 목표로 설정한 데이터소스를 네비게이션 하도록 설계하였다. 의미있는 바이오 정보를 탐색하기위한 통합 질의어 처리는 선택 레이어를 기반으로 전체 바이오 데이터베이스 대신에 선택된 데이터 소스만을 탐색하도록 하여 탐색공간에 대한 범위를 줄였다. 본 논문에서 제안된 질의 수행절차에 대한 검증을 위해 질의수행에 따른 다양한 변환규칙을 정형화하여 질의 수행 결과에 대한 완성도를 높이기 위한 연구를 추가적으로 진행할 예정이다.

Acknowledgement

이 논문은 충주대학교 대학구조개혁지원사업비(교육인적자원부 지원)의 지원을 받아 수행한 연구임.

참고문헌

[1] Levy A., "Logic-based techniques in data integration ", Logic Based Edited by Jack Minker, Kluwer Publishers, 2000.

[2] Caragea, D., Silvescu, A., and Honavar, "Towards a Theoretical Framework for Analysis and Synthesis of Agents That Learn from Distributed Dynamic Data Sources", In: Emerging Neural Architectures Based on Neuroscience, Berlin: Springer-Verlag, pp. 547 - 559, 2001.

[3] Peter Mork, Ron Shaker, Alon Halevy, Peter Tarczy - Hornoch, "{PQL}: A declarative query language over dynamic biological data", Proceedings of the American Medical Informatics Association (AMIA) Anual Symposium, American Medical Informatics Association, San Antonio, Texas, pp. 533-537, 2002.

[4] Peter Mork, Alon Halevy, Peter Tarczy-Hornoch, "A Model for Data Integration Systems of Biomedical Data Applied to Online Genetic Databases", Symposium of the American Medical Informatics Association, pp. 473-477, 2001.

[5] M. Wilkinson, D. Gessler, A. Farmer, L. stein, "Bio MOBY2003 - Hooked on MOBY!", Virtual Conference on Genomics&Bioinformatics program, , 2003.

[6] Z. Wang, X. Gao, C. He, J.A. Miller, J.C. Kissinger, M. Heiges, C. Aurrecoechea, E.T. Kraemer and C. Pennington., "A Comparison of Federated Databases with Web Services for the Integration of Bioinformatics Data", In Press, Conference Proceedings: The 2007 World Congress in Computer Science, Computer Engineering, & Applied Computing, pp. 334-338, 2007.

[7] Michael Y. Galperin, "The Molecular Biology Data- base Collection: 2007 update": NUCLEIC ACIDS RES., 25, pp. D3-D4, 2007.

[8] Vladimir Zadorozhny, Louiqa Raschid, "Query Optimi -zation to Meet Performance Targets for Wide Area Applications", ICDCS, pp. 271-279, 2002.

[9] Hass L. M., "Discovery Link: A System for integr -ated access to life sciences data sources", IBM System Journal, Vol. 40, No 2, pp. 489-511, 2001.

[10] Reinoso J., Silvescu, A., Caragea, D., Pathak, J., and Honavar, V. "A Federated Query-Centric Approach to Information Extraction and Integration from Heterogeneous, Distributed and Autonomous Data Sources", In: Proceedings of the 2003 IEEE International Conference on Information Reuse and Integration(IRI 2003), October 27-29, Las Vegas, NV, USA. pp. 183-191, 2003.

[11] Davidson, S.B., et al., "K2/Kleisli and GUS: Experiments in integrated access to genomic data sources": IBM. Systems J. Vol. 40, pp. 512 - 531, 2001.

[12] Reinoso-Castillo J., "Ontology-driven information extraction and integration from heterogeneous distributed autonomous data sources: A federated query centric approach", Masters Thesis, Iowa StateUniversity, 2002.

[13] Oh R. D, "NAMO : A Navigation Model on the Abstraction layer for Heterogeneous Biological Databases", The 4th Asia Pacific International Symposium on Information Technology, , Goldcoast, Australia, Jan., Vol. 1, pp. 26-27, 2005.
[14] Oh R. D, "Semantic Navigaion Model and Query Processing in Multi-Bio Layer based on Federated Biological Databases", The 2nd Asia Pacific International Conference on Information Science and Technology(APICIST2007), Hanoi, Vietnam,, Dec. 13-14, pp. 1-7, 2007.
[15] 신진섭, 안우영, 오일용, "생체정보측정을 통한 진단시스템 개발", 한국컴퓨터정보학회논문지, 제 13권, 제 1호, 219-226쪽, 2008년 1월
[16] 이승근, 김영민, "계층적 상황 온톨로지 관리를 이용한 상황 인식 서비스 미들웨어 설계", 한국컴퓨터정보학회 논문지, 제 11권, 제 1호, 185-194쪽, 2006년 3월

저 자 소 개

오염덕(Ryum-Duck Oh)

1993년 2월 홍익대학교 대학원 전자계산학과 이학박사
2001년 ~ 2002년 충주대학교 전산정보원장
1990년 ~ 현재 충주대학교 컴퓨터과학과 교수
2009년 ~ 현재 충주대학교 공학교육혁신센터장
관심분야: USN 데이터베이스, Data Mining, 바이오데이터베이스