# 시맨틱 웹 기반의 바이오 온톨로지 시스템의 설계

정희준°, 유명환*, 이강찬**, 김성한**, 민재홍**, 정인정*
*고려대학교 전산학과 {joonny96°, myong, chung}@korea.ac.kr
** 한국전자통신연구원 표준연구센터 {chan, sh-kim, jhmin}@etri.re.kr

## An Architecture for Bio-Ontology System Based Semantic Web

Hee-Joon Chung°, Myong-Hwan Yoo* Kang-Chan Lee**,
Sung-Han Kim**, Jae-Hong Min**, In-jeong Chung*
* Dept. of Computer Science, Korea University
** Protocol Engineering Center, Electronics and Telecommunications Research Institute

**요 약**

대부분의 생물학에서는 많은 지식을 도출 할 수 있는 공리(axiom)에 의한 응용 프로그램보다는 기존의 지식을 적용하고 있으며, 바이오인포매틱스 데이터베이스에 저장된 복잡한 생물학 자료는 추가, 변경이 자주 발생한다. 이러한 바이오인포매틱스의 데이터베이스와 응용 프로그램에서 지식을 표현하는 방법으로써 온톨로지의 사용이 제시되고 있다. 온톨로지는 사람과 컴퓨터간의 공유되는 지식을 개념화하고, 이를 명세화 하는 것으로 정의된다. 즉, 온톨로지는 도메인 내의 지식을 개념화한 구체적인 형식이며, 개념화와 개념화간의 관계를 표현할 수 있다. 또한, W3C에서 제안한 시맨틱 웹은 온톨로지는 중요 기술로 사용하고 있으며, 온톨로지를 통한 추론과 컴퓨터가 이해 가능한 형식을 제공하여 상호운영성등을 향상시킨다. 본 논문에서는 기존의 바이오 온톨로지들에 대해서 알아보고, 바이오 온톨로지 시스템의 설계와 시스템의 각 구성 요소에 대해서 제시한다. 마지막으로 이러한 시스템을 구축할 때에 고려되어야 하는 이슈들에 대해서 설명한다.

## 1. 서 론

바이오인포매틱스에서 온톨로지의 중요성은 최근 몇 년 동안 부각되어지고 있다[2]. 생물학 작업의 대부분이 기존의 지식을 기반으로 하여 알려지지 않은 물질에 대해서 분석과 연구를 한다. 그러나, 이러한 생물학 작업은 생물 데이터의 폭발적인 증가로 인하여 지식 기반의 데이터 처리가 쉽지가 않은 작업이며, 바이오인포매틱스 내의 복잡한 데이터베이스는 사용자가 새로운 지식의 추가와 그에 따르는 데이터의 정제를 하는 작업을 자주 수행하고 있다[1]. 이러한 환경에서 온톨로지는 바이오인포매틱스 지식의 표현하는 방법으로써 제시되고 있다. 온톨로지의 정의는 사람과 응용프로그램 간의 공유되는 지식을 개념화(conceptualization)하고, 이를 명세화(specification)하는 것이다[8]. 바이오인포매틱스 커뮤니티에서 온톨로지는 커뮤니티 지식의 공유와 표현을 할 수 있는 방법을 제공하고, 이는 공통적인 표현형식으로써 정의되어져 있으며, 이 기종 데이터베이스에서 지능적인 질의를 지원한다[2]. 또한, 온톨로지와 메타데이타는 자원을 조직하고 그 내용을 기술한다. 이러한 온톨로지의 정의와 설명은 자원의 상호운영성과 통합을 위한 필요조건이다[3].

또한, 현재의 웹이 가지고 있는 문제점인 비효율적인 정보와 자료의 검색을 해결하기 위해서 1990년대 말에 W3C(World Wide Web Consortium)에서 시맨틱 웹을 제시하였다. 시맨틱 웹은 현재의 웹 환경의 문제점을 해결하고 자동화된 웹 서비스를 제공하며 컴퓨터의 지능적인 정보처리가 가능토록 웹 문서내에 지식표현을 위한 온톨로지를 삽입하고 지식간의 관계를 설정하며 추론규칙을 포함시킨다. 이를 통해서 이기종간의 상호운용성을 보장하고 사용자가 원하는 웹서비스의 발견, 자동적인 웹 서비스의 실행과 동시에 웹 서비스들의 통합과 상호작용을 하여 원하는 정보를 검색하고 추론이 가능토록 한다[5]. 그리고, 시맨틱 웹 언어로써 제안된 웹 온톨로지 언어인 RDF(S), OIL, DAML, SHOE[4], [7]와 같은 언어는 웹 문서내에서 온톨로지를 표현함으로써 웹 문서에 나타난 정보의 의미와 정보들 간의 관계를 설정해 주어 시맨틱 웹의 기능을 수행할 수 있도록 한다.

본 논문에서는 바이오인포매틱스에서의 바이오 온톨로지를 가지고 시맨틱 웹 서비스를 제공하기 위한 시스템의 설계와 시스템의 각 구성 요소를 제시한다. 또한, 각 구성요소들에 대한 이슈들에 대해서 설명한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 현재의 바이오 온톨로지의 현황과 시맨틱 웹에 대해서 알아보고 3장에서는 지능적인 서비스가 가능한 바이오 온톨로지 시스템과 각 구성요소들의 기능과 관련 이슈들에 대해서 제시한다. 마지막으로, 결론과 향후과제를 언급한다.

## 2. 바이오 온톨로지와 시맨틱 웹

### 2.1 바이오 온톨로지

바이오인포맥틱스 분야에서 온톨로지의 사용은 최근에 들어서 사용이 되었으며, 현재 바이오 온톨로지로 개발된 수도 몇 가지 정도이며, 전체적인 생물학에 대한 온톨로지를 구성이 아닌 부분적으로 바이오인포매틱스나 분자 생물학 분야의 온톨로지가 개발되었다.[2] 온톨로지가 생물학 분야에서 사용이 이유는 것은 종(species)의 분류가 잘 되어져 있으며, 또한 풍부한 분류(taxonomy)가 있기 때문이다. 이러한 분류법에 의해서 계층적인 포함 관계를 표현할 수 있다. 즉, 분류법에 의한 관계의 설정이나 포함 관계 등이 명확하게 분류되어져 있는 것이다. 이 절에서는 현재 개발되어진 바이오 온톨로지를 소개한다.

현재 개발된 바이오 온톨로지는 다음과 같다[2].

- RiboWeb ontology
  http://smi-web.stanford.edu/project/helix/riboweb.html
- EcoCyc ontology
  http://ecocyc.PangeaSstems.com/ecocyc/ecocyc.html
- Schulze-Kremer ontology for molecular biology (MBO)
  http://igd.rz-berlin.mpg.de/~www/oe/mbo.html
- The Gene ontology
  http://genome-www.stanford.edu/GO/
- TAMBIS ontology
  http://img.cs.man.ac.uk/tambis

이상의 바이오 온톨로지들이 개발되었으며, 이러한 온톨로지를 표현하기 위하여 사용하는 지식표현 언어로는 프레임(frame)과 기술로직(description logic)을 사용하고 있다 또한, 이 바이오 온톨로지는 2가지의 중요한 특징을 가지고 있다. 먼

저 온톨로지는 커뮤니티 내의 데이터베이스와 응용프로그램에 입력되는 지식을 제공을 하고, 두 번째로 온톨로지는 도메인에 적합하도록 만들어져서 매우 상이하다는 것이다.

### 2.2 시맨틱 웹

1990년대 후반에 W3C에서는 차세대 웹으로써 시맨틱 웹을 제안하였다. 시맨틱 웹을 간단히 정의하자면, 시맨틱 웹은 웹 데이터에 의미(semantics)을 부여하여 컴퓨터 기계가 이해할 수 있는 언어로 만들어 컴퓨터에 의해 처리될 수 있도록 고안된 차세대 인터넷이다. 따라서 시맨틱 데이터는 다양한 이기종 분산처리 환경에서의 데이터 원천에서 결합이 가능할 수 있도록 구성되어 있으며, 데이터의 개념화와 관계를 토대로 이루어짐으로 온톨로지 기반으로 되어 있다[4],[5]. 온톨로지의 정의는 사람과 응용프로그램간의 공유하는 지식을 개념화하고, 이를 명세화하는 것으로 정의된다.[8]

시맨틱 웹을 사용함으로써 다음과 같은 이점이 있다. 먼저 웹 데이터로부터 구조화 및 정형화 된 정보를 효과적으로 추출할 수 있으며, 둘째로는 서로 다른 데이터 소스들은 온톨로지를 사용하여 상이한 데이터들의 통합 처리가 가능하게 된다. 셋째는 폭발적으로 증가하는 웹 데이터 증가를 효과적으로 관리할 수 있고, 넷째로 이러한 새민틱 웹을 위한 지능형 에이전트의 도움을 받아 지금까지의 인터넷 서비스와는 근본적으로 다른 서비스 변화를 가져 올 수 있다[5]. 즉, 현재의 웹에서는 사용자가 검색엔진을 사용하여 키워드로 조사하는 정도이지만, 지능형 에이전트가 온톨로지 기반의 시맨틱 웹을 사용한다면 다양한 데이터 소스들의 의미를 파악 처리하고, 이용자의 검색기록을 추론한 후 자신이 다른 웹을 찾아다니면서 필요한 정보들을 자발적으로 제공하여 사용자로 하여금 가장 효과적인 의사결정을 할 수 있도록 도와준다.

## 3. 바이오 온톨로지 시스템 설계

### 3.1 시스템 개요

지금까지 바이오 온톨로지의 연구는 온톨로지의 생성에 초점이 맞추어져 있다.[2] 또한, 온톨로지의 생성 및 관리를 위한 지식표현의 언어와 시맨틱 웹 언어로써 개발된 OIL, DAML, RDF(S)와 같은 언어로써 바이오 온톨로지에 적용을 하는 연구가 진행 중이다. 그러나, 아직 바이오 온톨로지를 활용하는 시스템의 개발과 연구는 미흡한 실정이다. 본 논문에서는 바이오 온톨로지를 활용하여 지능적인 서비스를 제공하는 시스템을 제안하며, 시스템의 전반적인 흐름은 그림1과 같다.

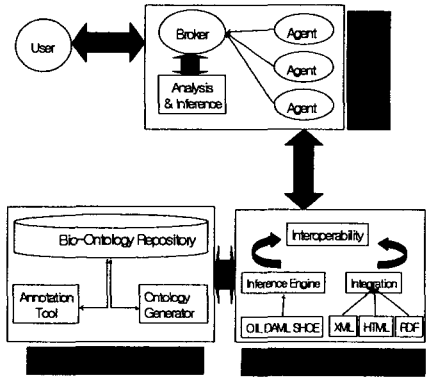


그림 1 바이오 온톨로지 시스템

그림 1에서와 같이 바이오 온톨로지를 활용하여 지능적인 서비스를 제공하기 위해서는 다음과 같은 세 가지 구성요소가 필요하다.

① 에이전트 컴포넌트
② 시맨틱 웹 컴포넌트
③ 바이오 온톨로지 컴포넌트

먼저 에이전트 컴포넌트에서는 검색과 분석 및 추론의 기능을 담당하는 에이전트의 생성과 에이전트가 수합한 정보를 수합하여 이 결과를 알려주는 구성 요소이다. 두 번째로, 시맨틱 웹 컴포넌트는 HTML, XML, RDF(S)의 언어들로 작성되어진 웹 페이지들과 OIL, DAML, SHOE와 같은 언어로 온토로지를 사용한 온톨로지 기반의 페이지들간의 상호운용성을 제공하는 컴포넌트이다. 마지막으로, 바이오 온톨로지 컴포넌트는 바이오 온톨로지를 저장하기 위한 바이오 온톨로지 저장소와 온톨로지 생성기 그리고, 온톨로지의 주석(annotation) 도구로써 온톨로지의 관리와 저장 및 생성을 담당하는 구성요소이다.

### 3.2 시스템 구성 요소

본 시스템의 구성요소인 에이전트 컴포넌트, 시맨틱 웹 컴포넌트, 바이오 온톨로지 컴포넌트와 컴포넌트의 구성 요소들에 대해서 설명한다.

#### 3.2.1에이전트 컴포넌트

에이전트 컴포넌트는 사용자가 어떠한 검색이나 분석을 하기 위하여 요청을 할 때에 정보를 검색하고 추론하여 반환하는 역할을 한다. 에이전트 컴포넌트에서 브로커(broker)는 크게 2가지의 역할을 담당한다. 먼저, 브로커는 에이전트를 생성하고 에이전트를 전파하는 역할을 담당한다. 즉, 브로커는 사용자로부터 요청을 받으면 이에 적합한 에이전트를 생성하여 해당하는 페이지들에 전파를 한다. 두 번째로는 에이전트가 수집해온 정보들과 사용자의 과거 검색기록으로부터 추론을 하여 사용자에 원하는 정보와 관련된 정보를 되돌려 주는 역할을 한다.

#### 3.2.2 시맨틱 웹 컴포넌트

시맨틱 웹 컴포넌트의 중요성은 바로 기존의 웹 언어인 HTML, XML 및 RDF(S)로 작성된 정보들과 온톨로지 기반으로 작성된 웹페이지간의 상호운용성을 제공하는 것이다. 현재의 웹페이지들의 대부분이 HTML과 XML로 작성이 되어져 있으며, 이는 온톨로지를 적용한 웹페이지간의 연동이 필수적으로 지원을 해야 한다. 본 시스템의 통합(integraion)모듈은 이러한 역할을 담당한다.

또한, 온톨로지를 웹에서 표현하기 위한 언어인 OIL, DAML과 SHOE와 같은 언어들은 온톨로지를 웹에서 표현을 할 수 있고, 추론을 가능하게 한다. 이를 위하여 추론 엔진(inference engine)에서 OIL, DAML, SHOE등의 언어가 온톨로지 컴포넌트로부터의 정보를 받아서 추론의 기능을 지원하게 된다. 마지막으로 상호운용성을 위하여 상호운용(interoperability) 모듈에서 기존의 웹 페이지와 온톨로지 기반의 언어로부터 나온 정보를 연동이 가능하게하는 역할을 담당한다.

#### 3.2.3 바이오 온톨로지 컴포넌트

바이오 온톨로지 컴포넌트는 바이오 온톨로지 저장소, 온톨로지 생성기, 주석(annotation) 도구로써 구성을 한다. 바이오 온톨로지 저장소는 생성된 온톨로지를 보관하고, 시맨틱 웹 컴포넌트에서 요청한 온톨로지를 제공하는 역할을 수행한다. 온톨로지 생성기는 온톨로지의 자동 생성을 해주는 역할을 수행한다. 이는 온톨로지의 추가적인 생성과 분할 등의 역할을 하게 된다. 마지막으로 주석 도구는 온톨로지에 주석을 처리할 수 있도록 하는 역할을 한다. 온톨로지에 주석을 명시하는 것은

온톨로지의 관리 및 관리자 및 사용자들에 정보를 제공하는 역할을 수행한다.

### 3.3 시스템의 장점

차세대 인터넷으로써 주목을 받고 있는 시맨틱 웹과 바이오인포매틱스의 바이오 온톨로지의 결합은 지능적이고 효율적인 기능을 수행하기 위해서 필요하다. 본 시스템은 시맨틱 웹의 기능을 사용함으로써, 다음과 같은 장점을 가진다.

① 온톨로지 사용으로 인한, 지식의 공유 및 재사용
② 검색 및 분석의 효율성의 증대
③ 추론을 통한 예측 및 복잡한 분석 기능 제공
④ 자동화된 기능 수행

온톨로지는 지식의 공유와 재사용을 가능하게 하는 기술이며, 현재 개발된 온톨로지의 공유와 통합을 할 수 있다. 또한, 검색 및 분석의 효율성의 증대는 시맨틱 웹 서비스의 기능을 사용함으로써, 검색에 따른 결과와 연관된 결과도 같이 제공함으로써, 보다 품질 높은 검색의 기능을 제공한다. 마지막으로, 추론을 통한 예측 및 복잡한 분석기능은 온톨로지의 추론 엔진을 사용함으로써, 바이오 온톨로지에 명시된 사항들에 대한 예측과 분석을 가능하게 한다. 예를 들어, 온톨로지에 명시된 데이터에서 새로이 발견된 데이터를 입력을 하면 추론을 하여 기능에 대한 예측을 가능하게 하는 것이다. 에이전트 컴포넌트를 사용함으로써 사용자의 계속적인 간섭이 없이 사용자의 요구를 에이전트가 분석하여 그에 적합한 결과를 제공한다.

즉, 본 시스템은 차세대 인터넷 환경에 맞도록 설계를 함으로써, 기존의 인터넷과의 상호운영성과 지능적인 웹 서비스 및 복잡한 분석이나 검색의 정확성을 높이는 성능을 갖는다.

### 3.4 기술적인 이슈

본 절에서는 이 시스템을 구현할 때에 필요한 기술적인 이슈들에 대해서 논하고자 한다.

#### 3.4.1 속도

본 시스템에서는 사용자의 요청에 의한 최적의 응답을 하기 위해서 시맨틱 웹 컴포넌트와 에이전트 컴포넌트를 적용을 하였다. 그러나, 온톨로지의 지속적인 개발과 온톨로지의 변경이 일어나는 동적인 환경에서는 최적의 응답을 짧은 시간에 해결하는 것은 어려운 일이다.

이를 해결하기 위해서는 메타 검색 엔진과의 결합으로써 속도 문제를 해결할 수 있다. 메타 검색 엔진은 사용자가 요청을 하였을 때, 관련성이 없는 데이터나 웹 페이지를 줄일 수 있으며, 시간적인 낭비를 줄일 수 있다.

#### 3.4.2 온톨로지 자동 생성 기술

현재 바이오 온톨로지의 생성은 생물학의 전문가에 의해서 수작업으로써 작성이 되고 있다. 이러한 온톨로지의 생성을 자동화함으로써, 새로운 온톨로지 생성에 소요되는 시간을 줄이는 방안이 필요하다. 이러한 방법은 휴리스틱(heuristic)한 방법을 적용하여 온톨로지를 개념화하고 이를 명세화하기 위하여 주석을 처리할 수 있는 자동화된 온톨로지 생성기 기술의 연구가 이루어져야 한다.

#### 3.4.3 온톨로지 버저닝(versioning) 기술

온톨로지 저장소에서는 기존의 온톨로지와 새로 작성된 온톨로지를 저장하는 역할을 담당한다. 그러나, 새로 작성되고 변경이 되는 온톨로지에 대한 명시가 되어야 하고, 변경 시에 기존의 온톨로지 또한 함께 보관을 하여야 한다. 이에 온톨로지 버저닝 기술에 대한 연구가 필요하다. 온톨로지 버저닝 기술이란, 새로운 온톨로지의 생성이나 다양한 다른 온톨로지의 관리를 하면서 발생되는 변경사항을 관리하는 기술을 말한다[9].

## 5. 결론과 향후 과제

본 논문에서는 바이오인포매틱스 분야에서 바이오 온톨로지를 활용하고, 지능적인 서비스를 위해 시맨틱 웹 기술을 적용한 시스템을 제안하였다. 바이오 온톨로지는 현재 해외에서 연구기관과 대학에서 생물학의 각 분야에 대한 온톨로지를 개발하였다. 또한, 시맨틱 웹은 그 개념이 소개가 최근에 이루어졌으나, 서비스를 위한 연구가 활발하게 진행되고 있다. 시맨틱 웹에서는 웹 문서에 온톨로지를 삽입하여 이상적인 기능과 서비스를 제공한다. 즉, 바이오 온톨로지와 시맨틱 웹의 결합은 보다 효율적이고 지능적인 처리 및 서비스를 위하여 필요한 실정이다. 이에 본 논문에서는 바이오 온톨로지와 시맨틱 웹의 기술을 결합한 시스템을 제시하였다.

이 시스템을 위해서는 온톨로지의 저장 기술과 온톨로지 버저닝 및 온톨로지 자동 생성을 위한 기술을 연구하여야 하며, 또한 효과적인 서비스를 제공하기 위한 에이전트의 연구가 병행되어야 한다. 또한, 재사용성 및 확장성을 오랫동안 진행해 온 소프트웨어 공학측면에서의 연구도 필요하다.

## 6. 참고 문헌

[1] Particia G. Baker, Carole A. Goble, Sean Bechhofer, Norman W. Paton, Robert Stevens, Andy Brass, "An ontology for bioinformatics application", Bioinformatics, Volume:15, no:6, page(s):510-520, 1999

[2] Robert Stevens, Carole A. Goble and Sean Bechhofer, "Onotlogy-based knowledge representation bioinformatics", 2001, Briefings Bioinformatics

[3] Robert Stevens, Carole Goble, Ian Horrocks, Sean Bechhofer, "Building a bioinformatics ontology using OIL", IEEE Transaction On Information Technology In Biomedicine, Volume:6, No:2, page(s):135-141, June 2002

[4] Gomez-Perez. A, Corcho.O, "Ontology languages for the semantic web", IEEE Intelligent Systems, Volume:17 Issue:1, Jan-Feb, page(s):54-60, 2002

[5] McIlraith.S.A, Son T.C, Honglei Zeng, "Semantic web services", IEEE Intelligent Systems, Volume:16 Issue:2, page(s):46-53, March-April, 2001

[6] J.Hendler, "Agents and the Semantic web", IEEE Intelligent Systems, Volume:16, Issue: 6, page(s):30-37, March-April 2001

[7]Lassila. O, van Harmelen.F, Horrocks I, Hendler J, McGuiness D.L, "The semantic web and its languages", IEEE Intelligent Systems, Volume:15, Issue:6, page(s):67-73, Nov-Dec 2000

[8]T.R.Gruber, "Towards Principles for the Design of Ontologies Used for Knowledge Sharing", International Workshop on Formal Ontology, Padova, Italy, 1993

[9]Michel Klein, Dieter Fensel, "Ontology versioning on the Semantic web", First Semantic web working symposium, stanford, CA, USA, August, 2001