한국산학기술학회논문지
Vol. 6, No. 4, pp. 309-314, 2005

웹 기반 데이터베이스로부터의 유용한 데이터 추출 기법의 설계 및 응용

황두성[1*]

Design and application of effective data extraction technique from Web databases

Doosung Hwang[1*]

요 약 본 논문에서는 생명공학 정보를 포함하는 분산 웹 데이터베이스들로부터 관련성에 기반하여 목표 데이터를 추출하는 기법들을 분석한다. 더불어 이 분석을 기본으로 단백질 데이터의 지식 확장 방법의 설계 및 구현을 제안한다. 웹 데이터베이스를 위한 데이터 추출기는 수동 추출, 반 자동 추출, 자동 추출 방법 등의 구현방법이 가능하다. 웹 데이터 추출기는 해당 웹 페이지에서 목표 데이터를 검색 및 추출하기 위하여 식별자를 이용하는 것이 일반적이다. 본 논문은 웹 데이터 추출 기법을 이용한 유기체 단백질 관련 데이터베이스 시스템의 설계와 구현을 기술한다.

Abstract This paper analyzes techniques that extract objective information from distributed web databases for bioinformatics based on relationship among information. Moreover, we discuss the design and implementation of a method for knowledge enhancement in respect of protein information. Web data extractor can be constructed by using a manual, semi-automatic, or automatic way. Data extractor generally makes use of identifiers in order to search and extract targeting information from a specified web page. This paper presents a design and implementation for the protein databases of an organism by utilizing web data extraction techniques.

Key words : data extractor, data integration, proteomics

1. 서 론

웹의 급속한 성장과 더불어 웹 데이터베이스는 여러 지역으로 분산된 다양한 분야의 지식들을 온라인에서 공유할 수 있는 지식 저장소가 되었다. 하지만, 분산된 정보들로부터 각 사용자들의 요구에 맞게 통합된 형태로 제공된다면 정보의 부가가치는 기대 이상으로 증가될 수가 있다. 분산된 환경에서 데이터 추출과 관련된 여러 연구가 진행되었으며, 일반적으로 웹 문서로부터 목적하는 정보를 여분의 자료와 식별할 수 있는 추출 규칙을 사용하여 의미 있는 데이터만을 선택할 수 있는 데이터 추출기가 많이 사용되었다. 추출기를 구성은 대체로 수동 추출, 반 자동 추출, 그리고 자동추출 방식이 사용되었다. 수동 추출 방식은 특정 웹 정보에 대한 추출기를 위한 데이터 추출 규칙을 사용자가 수동적으로 명시하여 추출기에 직접 입력하거나 추출기 내부에 구현하는 방법이다. 반 자동 추출 방식[1,2]은 필요한 데이터 정보를 사용자 인터페이스를 통해서 추출 규칙을 제공받으며, 마지막으로 자동 추출 방식[3,4,5]의 추출 규칙은 준비된 학습 데이터에 기계학습 알고리즘을 적용하여 얻어진다. 세 가지 방식 중에서 구현을 위한 선택은 대상이 되는 웹 데이터베이스의 수, 웹 데이터의 표현 방식 및 추출할 데이터의 수에 따라 결정된다. 웹 데이터 추출기의 기본적 기능은 특정 웹 데이터로부터 의미 부여가 가능한 목적 데이터 만을 분리하는 것이다. 이렇게 추출된 데이터는 주어진 목적을 위해 통합 과정을 거치게 되거나, 그 외 응용을 위해 명시적 형태로 변환되어 저장 된다.

추출기로부터 얻어진 데이터는 응용목적에 따라 유기적 또는 의미적 데이터 간의 관계를 이용하여 저장된다. 일반적으로 데이터간의 관계는 메타데이터를 이용하며 일관된 정보의 뷰를 제공하게 된다. 웹과 같은 분산환경

[1]단국대학교 컴퓨터과학과
*교신저자 : 황두성(dshwang@dankook.ac.kr)

에서 사용자 질의가 제공되면 시스템은 주어진 질의를 검색대상이 되는 웹 데이터베이스들에 적합하도록 분리하여 하위 질의를 발생시킨다. 생성된 하위 질의를 이용하여 시스템은 각 웹 데이터베이스 상황에 적절한 데이터 추출기를 생성-구동시켜 해당 웹 문서를 검색-분석 후 추출규칙을 이용하여 목적 데이터를 얻게 된다. 여러 웹 데이터베이스로부터 추출된 데이터는 시스템으로 보내어져 통합된 형태로 사용자에게 제공된다. 따라서 웹 데이터 추출과 통합은 응용 시스템과 웹 데이터베이스의 사이에서 용이한 데이터 변환 및 처리 메커니즘을 필요하고 있다.

본 논문의 구성은 다음과 같다. 웹 데이터를 추출하는 기법들에 관한 소개와 특징이 2절에서 소개되며 3절에서는 실험실에서 발생되는 단백질 데이터의 생물학적 의미 확장을 위한 데이터 추출 및 통합 시스템의 설계를 기술한다. 4절에서는 구축된 시스템의 Yeast 및 초파리(Drosophila)의 단백질 데이터의 지식확장에 응용을 보인다. 그리고 5절에서는 결론과 앞으로의 연구방향을 제시한다.

2. 관련연구

웹 데이터베이스로부터 데이터 추출 규칙의 생성은 제공되는 웹 데이터의 표현 방식인 HTML또는 XML 문서의 구조에 의존한다[6]. 사용자 질의로부터 검색된 웹 문서는 구조분석 과정을 거치며 처리의 단순화를 위해 정형화된 XML문서 또는 다른 트리 구조로 재 표현된다. 추출 규칙은 사용자와 대화를 통해 해당 웹 문서로부터 추출할 데이터 필드만을 대상으로 하여 생성된다. 수동 및 반 자동 추출 방식에 속하는 대부분 응용시스템이 여기에 해당된다. 대상 웹 데이터베이스의 수가 증가함에 따라 수동 추출 방식보다는 인터페이스에 의해 추출 규칙을 선택할 수 있는 반 자동 추출 방식이 많이 사용된다. 반 자동 추출 방식을 이용하는 대표적인 시스템으로는 W4F[1], LIXTO[2]가 있다.

자동 추출 시스템은 주석이 첨부된 학습 예제로부터 웹 데이터의 추출 규칙을 기계학습을 통해 학습한다. 이런 시스템들은 식별자를 이용하여 준비된 학습 패턴으로부터 추출규칙이 나타나므로 준비된 학습 데이터가 시스템의 성능에 중요한 요소가 된다. 사용될 추출규칙은 웹 데이터의 구조 및 표현에 민감하게 반응하지 않는다는 장점을 가지고 있어 가장 진보된 웹 정보 데이터 추출 시스템을 구성하고 있다. 그러나 수동 및 반 자동 추출 방식을 이용하는 시스템보다 상대적으로 복잡한 처리 과정을 거쳐야 하며, 웹 문서의 구조 변화가 빈번히 발생하는 웹 데이터베이스에 적용은 어렵다는 단점을 가지고 있다. 대표적인 시스템으로서는 STALKER[3]와 WIEN[4] 등이 있다.

프로티오믹스 관련 분야에서 이기종 데이터 서비스를 위한 응용 시스템 연구로서는 DataFoundary[7], TAMBIS[8], IBM DiscoveryLink[9] 등이 있다. 이 시스템들은 데이터 추출, 변형 및 통합에 있어 데이터 추출기-중재자(data wrapper-mediator[10,11])를 이용하며, 사용자에게 여러 데이터 소스들에 대한 일관된 처리방법을 제공한다. 메타 데이터를 사용하는 DataFoundary는 추출된 데이터를 중재자가 관리하는 미들웨어 데이터베이스에 저장하여 사용자 질의를 처리한다. TAMBIS는 논리언어 Grail를 이용하여 생물정보학 데이터베이스에 대한 지식베이스(knowledge base)를 구성할 뿐 아니라 그래픽 사용자 인터페이스를 이용한 데이터 검색 방법을 제공한다. TAMBIS의 구축된 지식베이스는 각 데이터 소스의 데이터 모델 표현, 데이터 연관성 및 상호운영성에 대한 정보를 포함한다. 선택된 웹 소스들에 대한 구조적 정보 데이터를 사용하는 DiscoveryLink의 질의 처리기는 사용자 질의를 분해, 웹 소스에 대한 적합한 질의로 변환 한 후 추출기를 통해 데이터를 검색한다.

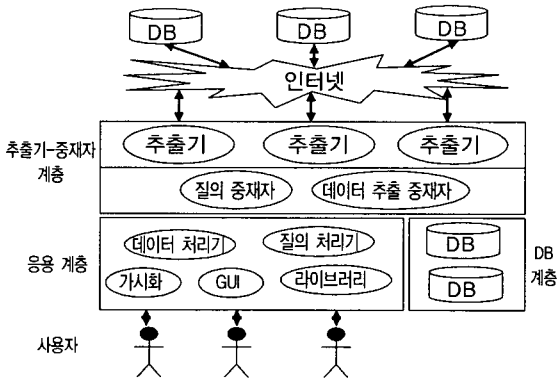

그림 1. 단백질 데이터의 지식 확장을 위한 시스템 구조

제안하고 있는 시스템은 추출규칙 생성, 데이터 통합, 확장성 등에서 살펴본 관련 시스템과 비교할 수 있다. 제안하는 시스템에서 웹 추출규칙의 생성은 상업화된 LIXTO와 같은 방식의 가시화 방식을 이용한 웹 소스의 특정 문서로부터 데이터 추출 규칙을 사용자와 인터페이스를 이용하여 제공한다. 이러한 방식은 기계학습을 이용하는 방식과 비교할 때 여러 추출 규칙을 포함시켜야 하는 경우에 효과적이다. 또한 DataFoundary나 DiscoveryLink 등과 비교할 때 관련 있는 웹 소스에 대한

메타 데이터베이스를 갖고 있지 않아도 된다. 웹 데이터 베이스들간의 교차참조 정보는 두 웹 소스간의 특정 데이터에 대한 지식을 연관시키게 되는 상호운영성을 제공하게 된다. 일반적으로 관련성이 높은 두 웹 소스로부터 추출된 데이터는 응용 프로그램에서 데이터의 연관 처리를 포함시키는 것이 방법이다. LIXTO, STALKER, WIEN 등이 이런 범주에 속한다. 제안하고 있는 시스템에서는 데이터 중재자가 상호운영성을 이용하여 데이터 통합을 하도록 설계되었다. 그리고, 시스템의 확장 가능성 면에서 제안하고 있는 시스템은 소규모의 관련성 있는 웹 데이터베이스로부터 데이터 추출 및 통합에 적합하다고 하겠다. 본 논문에서 사용된 특정 유기체의 단백질 실험 데이터는 유전자 또는 단백질의 고유 이름을 이용하여 여러 생물 정보 데이터베이스들로부터 관련 데이터의 검색 이 필요하다.

3. 시스템 설계

본 논문에서 제안하는 특정 유기체 단백질에 대한 관련 데이터 지식확장을 위한 시스템 설계는 데이터 추출기-중재자 방법을 기반으로 설계되었다. 제안하고 있는 시스템은 사용자 인터페이스를 통한 웹 데이터 추출기를 구성하므로 반 자동 추출 방식을 이용한다. 따라서 웹 데이터베이스로부터 키워드를 이용하여 검색된 문서를 분석하여 추출할 데이터를 구분한 후 추출 규칙을 결정한다. 웹 데이터 추출기는 대규모 키워드들에 대한 목적 데이터를 가져오는 데 결정된 추출규칙을 사용하게 된다.

여러 웹 문서로부터 추출된 데이터는 데이터 중재자의 입력이 된다. 데이터 중재자는 로컬 데이터베이스의 전역 스키마에 따라 추출된 데이터를 저장한다. 데이터 추출기 및 중재자의 역할은XML 데이터를 기반으로 지시된다. 데이터 통합 및 일관성 유지를 위해 데이터 중재자는 이질적 데이터 표현의 웹 데이터베이스 간의 상호운영성를 이용하여 관련 웹 데이터 추출기들을 생성시키고 운영하게 된다. 그러므로 데이터 추출 및 통합이 중재자 중심에서 설계되었고, 유기체 단백질 관련 데이터 분석에 필요한 데이터만을 데이터베이스에 통합하여 저장하도록 설계되었다.

[그림 1]은 웹 데이터를 이용한 단백질 데이터의 지식 확장 구조를 나타낸다. 이 시스템은 관계형 데이터 베이스를 사용하고 있으며 응용, 추출기-중재자 모듈의 계층화 구조로 구성되었다.

3.1 추출기-중재자 계층

데이터 추출기는 웹 데이터 소스의 HTML 혹은 XML 웹 문서로부터 목적 데이터를 추출하여 중재자에 제공한다. 추출기는 제공된 맵 파일의 검색어를 가지고 해당 웹 소스의 HTML 혹은 XML형태의 문서를 검색-분석하고, 목적 필드의 값을 추출하여 중재자에게 제공한다. 추출기의 맵 파일은 중재자로부터 제공되며 웹 문서 검색 경로, 검색어 그리고 추출할 데이터 속성을 포함하게 된다. 추출기와 중재자 행위를 정의하는 XML 맵 파일은 데이터 추출, 변형 및 저장을 위한 세부 명세를 정의 한다. 이 파일로부터 중재자는 새로운 웹 데이터 추출기의 생성, 데이터 매핑 및 연관 그리고 데이터 저장에 대한 역할을 수행한다. 현재, HTML과 XML 웹 문서를 처리할 수 있는 데이터 추출기가 개발되어 운영되고 있다. HTML 데이터 추출기는 웹 문서를 분석 후 트리 구조로 변환하여 이용된다. 표현이 자유로운 HTML 문서의 XML과 같은 트리 구조 문서의 변환에는 제약이 있어 50개의 추출할 데이터의 속성들까지만 제공되는 제한을 가지고 있다.

데이터 중재자는 그 역할에 따라 분리하여 설계되었다. 질의 중재자(query mediator)는 로컬 데이터베이스와 사용자 질의간에 서비스를 처리하도록 설계되었다. 응용계층의 질의 처리기는 질의 중재자의 도움을 받아 사용자 질의의 결과를 로컬 데이터베이스 또는 웹 데이터베이스로부터 도출한다.

추출 중재자(extraction mediator)는 전역적 스키마 통합을 위해 설계되었으며, 또한 스키마 엔티티, 데이터 관계성(relationship)과 데이터 무결성을 제공한다. 추출 중재자는 웹 데이터 추출기의 출력 결과인 XML 데이터의 검사, 로컬 데이터베이스에 저장을 위한 SQL 문의 생성 및 수행을 한다. 각 웹 소스에 대한 데이터 추출기는 하나씩 생성되며 추출된 데이터는 하나의 추출 중재자에게 제공된다. 추출 중재자는 여러 웹 데이터베이스에서 데이터 추출이 필요하게 되면 여러 개의 웹 추출기로부터 데이터 입력을 받는다. 추출 중재자는 웹 추출기로부터 가져온 목적 데이터를 단백질 데이터의 분석에 이용될 수 있도록 데이터베이스에 저장한다. 데이터 추출 중재자는 추출된 데이터와 데이터베이스의 필드간 매핑, 데이터 통합 및 저장에 필요한 일들을 수행한다. 그러므로 로컬 데이터베이스의 데이터 무결성에 대한 역할은 데이터 추출 중재자가 수행한다.

현재 5종류의 추출기 프로그램(FlyBase[12], GadFly [12], GeneOntology[14], GenBank[15], Swiss-Prot[16], CYGD[17])이 개발되어 시범 운영되고 있다.

3.2 데이터베이스 계층

데이터 추출 중재자의 결과는 로컬 데이터베이스에 저장되어 사용자의 질의를 처리하는데 제공된다. 표현 방식이 다른 웹 데이터들로부터의 일관성 있는 데이터를 제공하기 위해 전역 스키마를 고려하였다. 전역스키마 설계는 서로 다른 형태의 웹 소스에 존재하는 데이터 연관성을 이용하여 구성하였다. 데이터 추출 중재자는 웹 데이터 추출 시 생성되는 메타 데이터를 이용하여 추출된 데이터와 관계형 테이블의 세부 데이터에 대한 매핑 정보를 관리한다.

3.3. 응용 계층

사용자의 질의에 대한 다양한 결과 및 분석 방법을 제공한다. 사용자 질의 처리기는 일반적인 텍스트 형태의 질의 결과를 제공한다. 사용자 질의는 그래픽 사용자 인터페이스를 이용하여 입력된다. 질의는 로컬 데이터베이스로부터 우선 처리되나 만약 질의 결과가 없으면 추출기-중재자 계층의 질의 중재자에 질의에 대한 결과를 요구한다. 가시화 툴은 단백질 상호작용 데이터의 2차원 그래프의 표현을 제공한다. 가시화된 단백질-단백질 상호작용으로부터 선택된 단백질에 대한 추출된 속성 데이터를 제공하는 기능을 가지고 있다.

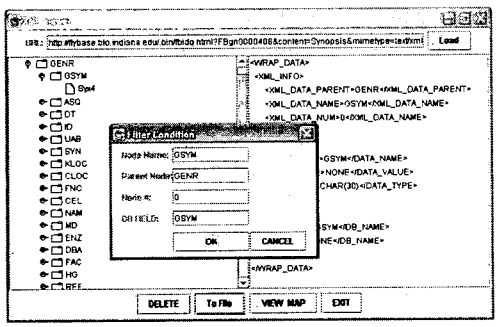

그림 2. 웹 데이터 추출을 위한 XML 맵 파일의 생성의 예

[그림 2]는 FlyBase의 Cyt-c-d의 검색된 웹 문서로부터 synonym 데이터를 위한 추출기의 맵 파일의 생성 예이다. URL 필드에 대상되는 단백질의 웹 문서의 검색 경로 정보를 입력하고 Load 버튼을 선택하면 왼쪽 화면에 검색된 특정 단백질의 FlyBase웹 문서를 트리 구조 문서로 변형시켜 보여준다. 보여진 웹 문서로부터 추출할 데이터 값 Syx4을 선택하면 Filter Condition 팝업 창에서 선택 필드에 대한 구조 정보가 나타난다. 이 예에서 synonym에 대한 FlyBase의 필드 속성의 태그는 GSYM이며, 부모 태그는 GENR이다. Node #는 검색된 웹 문서로부터 선택된 데이터 필드의 수이다. DB FIELD는 로컬 데이터베이스에서 사용할 필드의 이름이다. FlyBase 웹 데이터베이스로부터 여러 단백질 데이터의 synonym 값을 추출하기 위한 XML 맵 파일이 오른쪽 화면에 구성되었다.

4. 응용 사례

설계된 시스템은 단백질-단백질 상호작용이 알려진 두 단백질에 대한 웹 관련 데이터를 추출하는데 응용되었다. 선택된 유기체는 초파리와 Yeast이다. 두 유기체의 단백질 관련 추출된 데이터를 가지고 단백질 데이터베이스를 구축하였으며 단백질 데이터 분석에 사용하고 있다. 대부분의 생물정보학 데이터베이스들은 각기 설계 및 개발되어 제공되는 정보는 상이하지만, 각 정보는 생물학적으로 서로 상호운영성을 가지고 있다. 예를 들어 같은 유전자에 대해 GenBank는 그 유전자와 관련된 유기체이름, FlyBase접근번호 및 관련 데이터가 검색된다. 여기서 얻어진 FlyBase접근번호를 가지고 FlyBase 데이터베이스로부터 실질적인 유전자 이름, phenotype, 그리고 Swiss-Prot접근번호가 얻어진다. Swiss-Prot 접근번호는 그 유전자에 대해 밝혀진 단백질 속성 데이터들을 Swiss-Prot데이터베이스의 검색에서 이용된다. 이와 같이 여러 개의 개별적인 데이터베이스들로부터 얻어진 데이터들의 유기적인 결합은 데이터들 간에 상호운용성에 기반하며 생물학적 의미를 확장시키는데 이용된다.

[표 1]은 초파리 단백질에 대한 추출된 9가지 속성에 대한 데이터 비율을 보여주고 있다. 현재 로컬 데이터베이스는 주어진 11,595 단백질 중 40%에 해당하는 단백질, 약 5,000정도의 단백질에 대한 밝혀진 기능 정보, 그리고 약 11,000 단백질에 관한 유전인자 정보를 보유하게 되었다. 이런 사실은 유전인자 정보가 단백질 정보보다 풍부함을 시사하고 있다.

표 1. 초파리 단백질 상호작용에 대한 관련 데이터 추출 결과

내용	데이터 수	웹
Gene informatiom	11,595	GenBank, FlyBase
Gene synonym	3,955	GenBank, FlyBase
Molecular function	4,980	GenBank, GeneOntology
Biological process	3,255	GenBank, GeneOntology
Cellular component	3,502	GenBank, GeneOntology
Protein name	3,235	GenBank, Swiss-Prot
Protein synonym	2,253	GenBank, Swiss-Prot
mRNA length	11,595	GenBank, GadFly
Protein sequence	11,595	GenBank, FlyBase

표 2. Yeast 단백질 상호작용에 대한 관련 데이터 추출 결과

내용	데이터 수	웹
Gene	2,944	MIPS
Chromosome	11,585	MIPS
Viability	5,010	MIPS
Class	4,513	MIPS
Complex	5,751	MIPS
Phenotype	7,610	MIPS
Motif	6,855	MIPS
Function	11,129	MIPS
Protein sequence	2,944	MIPS, GenBanK

[표 2]는 Yeast 단백질에 대한 MIPS[17] 웹 데이터베이스로부터 관련 데이터를 가져온 결과를 보여준다. 초파리 단백질 관련 데이터의 추출은 5개의 웹 데이터베이스로부터 데이터 추출이 이루어 졌으나 MIPS의 CYGD가 대부분의 관련 데이터를 가지고 있으므로 단순 탐색 경로를 이용한 단백질의 키워드만을 이용하여 데이터가 추출되었다. 그러나 서열 데이터의 경우는 FlyBase의 접근번호를 가지고 GenBank로부터 추출된다. 표에서 약 3,000여 개의 각 Yeast 단백질에 대한 9개의 속성과 관련된 단백질의 수를 보여준다. 추출된 데이터 수에서 속성이 아직까지 밝혀지지 않은 unknown의 값은 계산되지 않았다. 유일한 단백질은 약 2~4개의 속성 데이터를 가지고 있으므로 추출된 데이터의 수가 유일한 단백질의 수보다 몇 배수의 양이다.

5. 결 론

본 논문은 생명정보학 관련 정보를 서비스하는 웹 데이터베이스로부터 목표 정보를 위한 데이터 추출기를 생성하는 기법들에 관해 검토하였다. 이를 바탕으로 웹 데이터베이스로부터 단백질 관련 데이터의 지식 확장을 위한 시스템의 구조를 제안하였다. 제안하는 특정 유기체의 단백질 데이터베이스는 웹 데이터 추출기 및 중재자의 역할이 XML 데이터를 기반으로 명시되며 추출된 데이터 변형 및 통합, 사용자 질의 처리는 데이터 중재자의 역할을 세분화 시켜 구성하였다. 다양한 웹 데이터베이스 간의 데이터 참조 관련성을 이용하여 여러 개의 웹 데이터 추출기들을 사용하였다. 제안된 방법은 반 자동 데이터 추출 방식으로 분류되며 유기체 초파리 및 Yeast 단백질의 상호작용 데이터베이스를 구성하는데 적용되었다. 제안된 시스템은 단백질-단백질 상호작용을 파악하기 위해 사용되는 분산된 웹 데이터베이스에서 효율적으로 자료를 추출하여 통합된 형태로 제공을 한다. 실제로 생명정보학에서 사용되는 응용 프로그램의 효율성은 자료 추출 및 통합에 많이 영향을 받는 이유로 제안된 시스템은 응용 분야에 대한 공헌을 한다고 할 수 있다.

현실적으로 웹 정보 데이터의 추출기를 생성하는데 있어 추출 규칙의 일반화에 많은 노력이 요구된다. 이러한 어려움을 극복하기 위하여 기계학습을 통한 추출 규칙의 생성에 대한 연구가 활성화 되었으나 특정 웹 데이터베이스에 활용이 제한되고 있다. 생물정보학 응용에서 관심있는 웹 데이터를 하나로 통합하는 과정은 복잡도가 높으며, 또한 비 표준화된 용어의 사용이 빈번하기 때문이다. 따라서 데이터 통합의 기능이 강조된 확장된 추출기에 대한 방법론과 효율성 검증을 필요로 한다. 이를 위해서 다음과 같은 연구의 진행이 필요하다. 첫째, 본 논문에서 제안하는 것과 유사한 데이터 통합과 관련된 정보를 모두 포함하는 XML 맵과 비슷한 메타데이터를 이용하는 것이다. 이 메타데이터는 일종의 다양한 표현의 이질적인 웹 소스들간의 연계를 위한 정보를 정의하고 있을 뿐만 아니라 개별 추출기에 필수적인 데이터를 포함할 것이다. 둘째, 여러 웹 소스들간의 의미론에 관한 단일화된 개념 규약이 없으므로 서로간의 연계에서 어려운 상황이 발생하는 것은 당연한 사실이므로 각 개념들에 관해 정형화된 표현의 온톨로지(ontology)에 대한 필요성이 높아지고 있다. 온톨로지를 이용한 시스템이 구축된다면 이러한 의미론적 문제의 일반적인 해결이 될 것으로 기대된다.

참고문헌

[1] Arnaud Sahuguet andFabien Azavant, Building light-weight wrappers for legacy Web data-sources using W4F, VLDB, 1999.

[2] R. Baumgartner, Sergio Flesca and Gottlob, Supervised Generation with Lixto, VLDB, 2001.

[3] Muslea, S. Minton, and C. A. Knoblock, Wrapper Induction for Semistructured, Web-base Information Sources Conference on Automated Learning and Discovery, 1998.

[4] N. Kushmerick, Wrapper Induction: Efficiency and expressiveness, Artificial Intelligence Journal 118, 15-68, 2000.

[5] Wei Han, David Buttler and Calton Pu, "Wrapping Web Data into XML," SIGMOD Record, Vol.30, No.3, pp.33-45, 2001.

[6] Frederic Achard, Guy Vaysseix and Emmnauel Barillot,

"XML, Bioinformics and data integration," Bioinformics Review, Vol.17, No.2, pp.115-125, 2001.

[7] T. Critchlow, K. Fidelis, M. Ganesh, R. Musick and T. Slezak, "DataFoundry : Information Management for Scie-ntific Data," Processdings of IEEE Advances in Digital Libraries, 2000.

[8] C. A. Goble R. Stevens, G. Ng, S. Bechhofer, N. W. Paton, P. G. Baker, M. Peim and A. Brass, "Transparent access to multiple Bioinformics information sources," IBM Sys-tems Journal, Vol.40, No.2, pp.532-551, 2001.

[9] IBM Life Science Solution Team, IBM Life Science Solu-tions : Turing Data into Discovery with Discovery-Link, Redbooks, 2002.

[10] Hector Garcis-Molina, Yannis Papakonstantinos, Dallan Qu-ass, Anand Rajaraman, Jeffrey Ullman, Jennifer Widom and Vasilis Vassalos, "The TSIMMIS Approach to Mediation : Data Models and Languages," Journal of Intelligent In-formation Systems, Vol.8, No.2, pp.117-132, 1997.

[11] Mary Tork Roth and Peter Schwarz, "A Wrapper Architecture for Legacy Data Sources," IBM, Technical Report RJ10077, 1997.

[12] FlyBase Consortium, http://flybase.org/.

[13] GadFly, http:// flybase.bio.indiana.edu/annot/.

[14] GENE ONTOLOGY Consortium, http://www. geneon tology.org.

[15] GenBank, http://www.ncbi.nlm.nih.gov/GenBank.

[16] SWISS-PROT, http://www.expasy.ch/sprot/sprot-top.html.

[17] MIPS CYGD, http:// mips.gsf.de/proj/yeast/.

황 두 성(Doosung Hwang) **[정회원]**

- 1986년 2월 : 충남대학교 계산통계학과(이학사)
- 1990년 2월 : 충남대학원 계산통계학과(석사)
- 1993년 2월 : 한국대학교 전자공학과(공학박사)
- 2003년 5월 : Wayne State University(이학박사)
- 2003년 9월~현재 : 단국대학교 컴퓨터과학과

<관심분야>
데이터 마이닝(data mining), 머신 학습(machine learning), 바이오인포매틱스(bioinformatics)