

## 품질기반의 웹 서비스 검색을 위한 확장 UDDI 개발

박성수<sup>○</sup> 이종근 윤지희  
 한림대학교 컴퓨터공학과  
 {sspark<sup>○</sup>, jeikei, jhyoon}@hallym.ac.kr

### Development of an Extended UDDI for Quality based Web Service Retrieval

SungSoo Park<sup>○</sup>, JongKeun Lee, JeeHee Yoon  
 Dept. of Computer Engineering, Hallym University

#### 요약

최근 이질 분산형태를 갖는 정보를 통합하는 방법으로서 웹 서비스 기술을 이용한 바이오 정보 시스템이 개발 구축되고 있다. 이러한 웹 서비스 기반 바이오 정보 시스템으로 Bio-MOBY, DDBJ, MyGrid Project 등을 들 수 있다. 그러나 이들 기존 시스템에서는 선택한 DB에 대한 accession 번호 검색을 지원하거나, 시스템에 등록된 서비스의 선택만이 허용되는 등 이용형태가 매우 제한적이다. 또한 서비스의 품질 평가 기능이 제공되지 않아 서비스의 관련성을 판별하지 못하며, 심지어 링크가 바르게 연결되지 않았거나, 작동하지 않는 서비스의 분별조차 불가능한 실정이다. 본 논문에서는 이러한 문제점을 해결하고자 서비스 검색과정에서 웹 서비스의 품질을 평가하고 평가된 품질을 기반으로 웹 서비스를 순위화해 사용자에게 제공하는 품질기반 UDDI를 제안한다. 이를 위해 우리는 Gene Ontology를 이용한 연관 키워드 검색 방식과 키워드 기반의 서비스 품질 평가 방법을 제안하고, 본 방식의 유용성을 보인다.

#### 1. 서론

생물정보학 분야의 활발한 연구 개발로 인하여 생물학 데이터가 기하급수적으로 증가하고 있으며, 생물학, 전산학 전문가들은 공동으로 이 들 분산 이질의 데이터를 분석, 관리하기 위한 통합 바이오 정보 시스템 개발에 많은 시간과 노력을 투자하고 있다[1]. 웹 서비스(Web Service)는 DCOM, CORBA등의 분산객체의 뒤를 이어 등장한 기술로 SOAP(Simple Object Access Protocol), WSDL(Web Service Description Language), UDDI(Universal Description Discovery Integration)를 기반으로 각종 이질 분산 정보를 통합하는 방법으로 최근 주목 받고 있다.

웹 서비스의 핵심 기술인 UDDI는 동적으로 생성되어 등록되는 다수의 웹 서비스를 효율적으로 저장, 관리하여 사용자들에게 유연하고 편리한 검색 환경을 제공하여야 한다. 그러나 현재의 UDDI는 다음과 같은 몇 가지 문제점을 가지고 있다. UDDI는 등록된 웹 서비스의 품질을 평가하여 사용자에게 제공하는 기능을 제공하지 않으므로, 사용자는 서비스 선택에 어려움을 겪게 된다[2,3,4,5]. 또한 UDDI 레지스트리는 IBM의 UDDI business 레지스트리와 SAP의 sap@uddi.business 레지스트리 등과 같이 e-비즈니스 분야 위주로 개발되어 있어, 바이오 정보 시스템 분야에 적합한 UDDI 레지스트리 개발이 미흡한 상황이다. 기존 바이오 정보 시스템 중 웹 서비스를 기반으로 하는 대표적인 실험적 시스템으로서 Bio-Moby[6]와 DDBJ[7,8]등을 들 수 있다. 그러나 이 들 시스템은 UDDI 레지스트리 검색의 기본 방식인 키워드 검색 방식을 지원하지 못하며, 시스템 이용 형태가 매우 제한적이다.

본 논문에서는 이와 같은 문제점을 분석하고 이를 해결할 수 있는 방안으로서 키워드 중심의 품질 평가 기능을 갖는 웹 서비스 바이오 정보 시스템의 설계 및 구현 방식을 제안한다. 본 시스템은 키워드 검색 방식을 지원하며, GO(Gene Ontology)[9]를 활용하여 입력된 키워드와 의미적으로 유사한 혹은 상위 개념의 키워드를 찾아 이 들을 연관 키워드로 추출하고, 이들 간의 순위

를 설정하여 순서화된 키워드 리스트를 생성한다. 다음 이 들 키워드를 이용하여 UDDI의 등록 서비스를 검색하여 사용자에게 제시하게 되며, 이 때 키워드의 출현 빈도 및 서비스 실행 시간을 기반으로 각 서비스의 품질 평가를 수행하여 서비스 품질 순위에 따라 서비스 검색 결과를 사용자에게 제시한다. 따라서 사용자는 각 서비스의 품질 평가 순위에 따라 제시된 서비스에 대하여 보다 선택적인 서비스 활용이 가능해지게 된다.

#### 2. 관련연구

##### 2.1 웹서비스

웹 서비스는 각종 분산 이질 형태를 갖는 정보를 통합하기 위한 XML 기반 기술이다[10,11]. 웹 서비스는 서비스 제공자(Service Provider), 서비스 요청자(Service Requester), 서비스 대리자(Service Agency)로 구성된다. 서비스 제공자는 웹 서비스를 구성하여 서비스를 제공하는 소프트웨어 시스템이고, 서비스 요청자는 서비스를 받기위해 웹 서비스를 호출하는 시스템이며, 서비스 대리자는 서비스 요청자가 자신이 원하는 웹 서비스를 찾을 수 있도록 하는 레지스트리이다. 관련 기술은 다음과 같다.

○ WSDL[12] : 웹 서비스의 IDL(Interface Definition Language) 버전이며, 웹 서비스의 방법과 프로토콜, 데이터 포맷 등을 정의하는 표준 스크립트이다. 서비스 요청자가 서비스를 호출하는 방법과 사용할 수 있는 매개변수들의 목록을 포함한다.

○ SOAP[13] : 클라이언트의 작업 요청과 시스템의 응답을 XML 문자열로 구성하고 전송하는 표준 프로토콜이다.

○ UDDI[14] : 웹 서비스에 관한 정보를 등록, 탐색하기 위한 레지스트리로, 서비스 설명과 서비스 발견에 대한 표준 스펙을 제공한다. UDDI는 각 표준과 프로토콜별로 가상의 모든 인터넷 페이지를 사용하여 매우 유연한 서비스를 구성한다. 그림 1.는 UDDI Version 3.0.2의 표준 데이터 구조를 보인다. UDDI의 각 구성요소는 XML 형태로 표현되어 저장되며, BusinessEntity, BusinessService, BindingTemplate, tModel로 이루어진다. BusinessEntity는 제공자에 대한 정보를 담고 있으며, BusinessService는 제공자가 제공하는 웹 서비스에 대한 다양한 타임을 기술하

본 연구는 과학기술부 과학재단 목적기초연구(R01-2006-000-11106-0)지원으로 수행되었음.

고, BusinessTemplate은 서비스 진입점에 대한 정보를 담고 있다. 마지막으로 tModel은 웹 서비스와 어떻게 상호작용할 수 있는지에 대한 방법을 담고 있다.

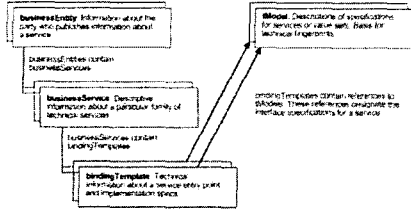


그림 1. UDDI Version 3.0.2 표준 데이터 구조

### 2.2 온톨로지

온톨로지는 공유된 개념에 대한 정형화되고 명시적인 명세이다. 온톨로지는 단어와 관계들로 구성된 일종의 사전으로서 생각할 수 있으며, 그 속에는 특정 도메인에 관련된 단어들이 계층적으로 표현되어 있고, 추가적으로 이를 확장할 수 있는 추론 규칙이 포함되어 있어, 웹 기반의 지식 처리나 응용프로그램 사이의 지식 공유, 재사용 등이 가능하도록 되어 있다. 본 연구에서는 생물정보학에서 사용되고 있는 대표적인 GO를 지식 데이터베이스로 활용하여 연관 키워드 추출에 활용하였다. GO는 단백질, 유전자 등과 같은 유전자 산출물(gene product)들이 각 생물체 내에서 수행하는 작용 및 기능 등을 구조화하여 체계적으로 기술하고 있다. GO는 Molecular Function(MF), Biological Process(BP), Cellular Component(CC)의 세 가지 범주로 나누어져 있다. 각 범주의 GO는 term이라 부르는 노드(node)들의 집합으로 이루어지며, 이들 노드는 *is\_a* 관련(혹은 *has\_part*의 관련)으로 연결된 DAG(Directed Acyclic Graph) 형태를 갖는다. 여기에서 각 term은 생물학적 작용 및 기능 등을 표현하고 있으며, 각 term과 연관된 단백질, 유전자 정보가 association 정보의 형태로 링크되어 있는 형태를 갖는다.

### 3. 확장 UDDI 모델

#### 3.1 UDDI의 문제점

현재의 범용 UDDI는 단일 키워드 검색 등 제한된 검색 방식에 의존하여 효율적인 정보 검색 환경을 지원하지 못하며, 등록된 서비스의 신뢰성과 검색 결과에 대한 품질 평가를 지원하지 못하는 등의 문제점을 가지고 있다[2,3]. 즉, 기존 UDDI는 제공자 명(Business Name), 서비스 명(Service Name), tModel 명 등에 대한 단순 키워드 기반 질의만을 허용하고 있으며, 서비스의 품질 평가 기능이 제공되지 않아 서비스의 관련성을 판별하지 못하며, 심지어 링크가 바르게 연결되지 않았거나, 작동하지 않는 서비스의 분별조차 불가능한 실정이다.

또한 기존의 범용 UDDI 레지스트리는 e-business 분야 위주로 개발되어 있어, 바이오 정보 시스템 분야에 적합한 UDDI 레지스트리 개발이 미흡한 상황이다. 기존 바이오 정보 시스템 중 웹 서비스를 기반으로 하는 대표적인 시스템으로 Bio-Moby[6]와 DBJ[7]등을 들 수 있다. 그러나 Bio-MOBY는 서비스 검색 대상의 데이터베이스를 미리 설정하고 accession 번호만을 이용한 검색을 지원하며, DBJ는 UDDI 레지스트리가 아닌 제작자의 설계를 가지며, BLASTDemo, Blast, ClustralW, DBJ, Ensembl 등 15개의 정해진 WSDL 문서를 제공하여 정보 검색을 지원하고 있다. 또한 이들 시스템은 레지스트리에 저장되어 있는 정보의 신뢰도를 가능하기 어려우며, 예를 들어 Bio-MOBY의 경우, 시스템에서 매뉴의 기본 값으로 설정되어 있는 "NCBI\_gi"의 "431260"을 이용해 검색을 수행하면, 서비스 16건 중 결과를 돌려주는 서비스는 3건에 불과하고 링크가 깨진 서비스가 11건, Bio-MOBY Cross Ref.[15]를 돌려주는 서비스가 2건으로 대부

분의 서비스들이 검색 결과를 못 돌려주고 있다.

#### 3.2 UDDI 개선 전략

이 들 문제점을 해결할 수 있는 방안으로 본 연구에서는 GO를 이용한 연관 키워드 추출 기법과 서비스 품질 평가 방식을 제안한다.

##### 3.2.1 연관 키워드 추출 기법

바이오 정보 검색을 위하여 사용자가 단일 키워드를 입력하여 웹 서비스 검색을 수행하는 경우, 돌아오는 서비스 검색 결과가 전혀 없거나 사용자의 요구를 충분히 수행한 서비스를 찾지 못하는 경우가 발생할 가능성이 매우 높다. 따라서 이와 같은 이유로 기존의 웹 서비스 기반 바이오 정보 시스템에서는 제한적인 사용자 환경만을 제공하고 있는 것이 사실이다. 따라서 본 연구에서는 온톨로지를 사용한 입력 키워드 확장 방식을 사용하여 입력 키워드로부터 연관 키워드를 다수 추출하여 이를 서비스 검색에 활용한다. 또한 연관 키워드 추출 시 이들 사이에 순위(rank)를 설정하여 서비스 검색에 활용하며, 이는 서비스 검색 성능을 높이는 역할을 담당한다. 입력된 키워드로부터 연관 키워드 리스트를 추출하는 기본적인 방법은 다음과 같다.

- (1) 유전자명, term name, 일반 전문 용어 등으로 이루어진 입력 키워드의 유사 키워드를 추출한다. : 본 시스템에서는 GO에서 제공하는 정보를 관계형 데이터베이스 MySQL 4.에 구축하였다. 이들 GO 테이블로부터 통칭 동의어(synonym)/약어(acronym) 정보를 추출한다. 예를 들어 유전자명의 입력 키워드를 위한 유사 키워드는 "gene\_product" 테이블, "gene\_product\_synonym" 테이블을 조인한 결과로부터 얻어진다. 또한 용어들의 형태소 및 구조 분석 등을 통하여 용어 간의 연결성을 높인다.
- (2) 입력 키워드의 상위 개념의 키워드를 추출한다. : 입력 키워드의 상위 개념을 추출하기 위하여 GO를 사용한다. 입력 키워드가 유전학적 기능 등을 나타내는 term name의 경우, GO의 계층 구조로부터 직접 상위 개념의 name을 추출한다. GO는 노드라 부르는 term이 *is\_a* 관련(혹은 *has\_part*의 관련)으로 연결된 DAG(Directed Acyclic Graph) 형태를 가지므로 *is\_a* 관련으로 연결된 term 사이에 상위 개념 추출이 가능하다. 한편, 입력 키워드가 유전자명의 경우, 역시 GO로부터 상위 개념 추출이 가능하다. GO 상에서 각 유전자 정보는 term에 association 정보로 링크되어 있으며, 각 유전자가 링크되어 있는 term이 각 유전자의 기능을 표현하는 상위 개념을 나타낸다고 볼 수 있다. 입력 키워드가 복수개의 유전자명으로 이루어진 경우, 이들 유전자들의 상위 개념을 나타내는 term을 추출하기 위하여 유전자 클러스터링 기법을 사용할 수 있다[16,17,18]. 본 연구에서는 참고 문헌 [16]의 방식을 사용하여 유전자들의 상위 개념을 나타내는 term을 추출, 이들을 순위화한다.

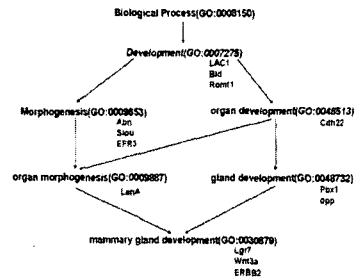


그림 2. GO에서 유전자와 상위 키워드와의 관계

예를 들어 그림 2의 GO에 대하여 입력 키워드로서 K = {ERBB2, dpp}가 주어진 경우, 이들의 상위 개념으로 'gland development', 'organ development', 'Development'가 추출되며,

'gland development'가 가장 순위가 높게 추천된다. 여기에서 'organ development'는 GO의 term 이름을 나타내며, GO:0048513은 해당 term의 GO 식별자를 나타낸다. 또한 'Cdh22'는 'organ development' term과 관련된 유전자명을 나타내며, association 정보의 형태로 term과 링크되어 있음을 나타낸다.

3.2.2 서비스 품질 평가 방법

웹 서비스 품질 평가 요소로 이전 연구[19,20]를 살펴보면 가용성(availability), 실행시간 등 많은 요소로 품질 평가를 수행하고 있다. 이와 같은 다양한 요소를 사용한 품질평가를 원활히 수행하기 위해서는 UDDI의 기본 자료구조 이외에 추가적인 내용들을 기술해야 하기 때문에 서비스 제공자와 사용자가 이 품질 평가 요소를 기술 관리해야 한다. 본 연구에서는 서비스 품질을 결정하는 요소로 서비스 이름, 서비스 설명 등과 같은 레지스트리에 저장된 정보를 이용한 서비스 선호도와 서비스 실행 시간을 사용한다. 서비스 선호도란 입력 키워드와 검색된 서비스의 연관성으로 정의할 수 있다. 즉 서비스를 기술하는 서비스 문서를 평가하여 입력 질의와의 연관성을 추출하는 방식으로 사용한다. 여기에서 서비스 문서는 UDDI에 등록된 각 서비스의 서비스 이름, 서비스 설명 등의 관련 정보를 연결한 가상의 문서로 정의한다. 본 연구에서는 다음과 같은 수식을 이용해 서비스  $S_j(j=1, \dots, n)$ 의 품질 순위를 계산한다.

$$Quality(S_j) = \sum_{i=1}^n \frac{K_i \cdot f_i + w(K_i)}{S_j \cdot f_j}$$

위 수식은 입력 키워드 리스트  $[K_1, K_2, \dots, K_n]$ 를 입력받으며, 이때 입력받은 입력 키워드 리스트는 사용자의 입력 키워드와의 연관성에 따라 순위화 된 리스트로 가정한다. 여기서  $K_i \cdot f_i$ 는 입력 키워드  $K_i$ 가 서비스 문서  $S_j$ 에서 얼마나 출현하는지의 출현 빈도를 나타내며,  $w(K_i)$ 는 입력 키워드  $K_i$ 의 순위에 의한 가중치를 나타낸다. 또한  $S_j \cdot f_j$ 는 서비스 문서  $S_j$ 에 출현하는 단어 수를 의미하며, 실제로는 UDDI에 등록된 서비스  $S_j$ 의 서비스 이름, 서비스 설명 등에 등록된 단어수를 의미한다. 즉  $Quality(S_j)$ 는 서비스 문서  $S_j$ 에 출현하는 입력 키워드들의 출현 빈도와 그들의 가중치 합을 서비스 문서의 단어 수로 나누는 결과를 의미한다.

이러한 서비스 선호도와 더불어 본 연구에서는 서비스 실행 시간을 품질 평가 요소로 사용하였다. 이는 여러 품질 평가 요소 중 서비스가 실제 실행되고 있는지를 가장 정확하게 판단할 수 있는 요소이기 때문이다. 이렇게 평가된 각 서비스들은 평가 결과에 의해 순위화 되어 사용자에게 제공된다. 또한 이렇게 평가된 서비스 품질 자료는 자체적인 품질평가 데이터베이스에 축적되며 이후 이 요소를 이용해 품질 평가 시간을 단축 시키는데 활용한다.

4. 시스템 구조

확장된 UDDI모델의 시스템 구조를 그림 3.에 보인다. 본 시스템은 3개의 모듈로 구성되어 있고 각 모듈의 기능은 다음과 같다.

- 키워드 추출 모듈 : 사용자가 입력한 키워드에 대하여 GO를 이용하여 연관 키워드 리스트를 추출하는 모듈
- UDDI 검색 모듈 : 키워드 추출 모듈을 통해 확장된 질의를 Bio-MOBY와 같은 외부 UDDI를 통해 검색하고 검색된 결과를 품질평가 모듈로 전해주는 모듈
- 품질평가 모듈 : UDDI를 통해 검색된 결과를 각 서비스별로 품질을 계산하고 순위화한다. 품질평가 결과는 추후의 품질 평가에 사용되기 위해 품질평가 데이터베이스에 저장된다.

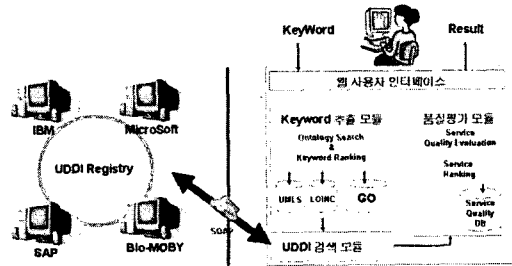


그림 3. 시스템 구조도

5. 결론 및 향후 연구과제

본 연구에서 제안한 시스템은 GO를 이용한 품질 기반 서비스 검색 결과를 제공하는 확장된 UDDI 시스템이다. 현재 프로토타입 시스템을 개발 중에 있으며, 로컬 품질평가 데이터베이스의 확장에 의한 시스템 성능 개선 작업을 수행 중에 있다.

6. 참고문헌

- [1] Stein, L., "Creating a bioinformatics nation," Nature, Vol. 417 pp.119-129, 2002.
- [2] 임의천, 김현수, 김승령, "웹 서비스의 평가, 인증 제도에 관한 탐색적 연구," 정보화정책 Vol. 10, No. 1, pp. 99-119, 2003.
- [3] Clark, M., "UDDI-The Weather Report," <http://www.webservicesarchitect.com>. 2001.
- [4] Maximilien, E. and Singh, M., "Toward Autonomic Web Services Trust and Selection," ICSC'04, pp. 15-19, 2004.
- [5] 윤석현, 김동준, 한상용, "웹 서비스 품질 기반 검색을 위한 UDDI 개선 모델," 정보과학회, Vol. 31, No.5, pp. 511-518, 2004.
- [6] BioMOBY : "<http://www.bioMOBY.org>"
- [7] DDBJ : "<http://www.ddbj.nig.ac.jp>"
- [8] Sugawara, H. and Miyazaki, S., "Biological SOAP Server and Web Services provided by the public sequence data bank," Nucleic Acids Research, Vol. 31, No. 13, pp. 3836-3839, 2003.
- [9] Gene Ontology : "<http://www.geneontology.org>"
- [10] "<http://www.w3.org/TR/2002/WD-ws-arch-20021114/>", W3C Working Draft, 2004.
- [11] "<http://www.w3.org/2002/ws/>", Web Services Activity.
- [12] WSDL : "<http://www.w3.org/TR/2001/NOTE-wsdl-20010315>"
- [13] SOAP : "<http://www.w3.org/TR/2000/NOTE-SOAP-20000508>"
- [14] UDDI : "[www.uddi.org](http://www.uddi.org)", OASIS
- [15] Wilkinson, M., Gessler, D., Farmer, A., and Stein, L., "The BioMOBY Project Explores Open-Source, Simple, Extensible Protocols for Enabling Biological Database Interoperability," Proc Virt Conf Genom and Bioinf, Vol. 3, pp. 17-27, 2003.
- [16] Joslyn, A., Mniszewski, S., Fulmer, A., and Heaton, G., "The Gene Ontology Categorizer," Bioinformatics, Vol. 20, Suppl.1, pp. i169-i177, 2004.
- [17] Lee, S., Hur, J., and Kim, Y., "A graph-theoretic modeling on GO space for biological interpretation of gene clusters," Bioinformatics, Vol. 20 No.3, pp. 381-388, 2004.
- [18] Smid, M. and Dorssers, L., "GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontolog terms," Bioinformatics, Vol. 20, No.16, pp. 2618-2625, 2004.
- [19] Menasce, D., "QoS Issues in Web Services," IEEE Internet Computing Vol. 6 No. 6 pp. 72-75, 2002.
- [20] Zeng, L., Benattallah, B., Dumas, M., Kalagnanam, J., and Sheng, Q., "Quality Driven Web Services Composition," www2003, pp. 411-421, 2003.