생물학 온톨로지를 이용한 XQuery 확장 시스템 설계 및 구현

김정진[0] 양경아[1] 양재동[1] 배명남[2] 정명근[2] 임명은[2]
전북대학교 전자정보공학부
{kimjj81[0], kayang[1], jdyang[1]}@chonbuk.ac.kr

{mmbae[2], aobo[2], melim[2]}@etri.re.kr

Design and Implementation of a XQuery Expansion System using Bio-Ontology

Jeongjin Kim[0], Kyungah Yang[1], Jaedong Yang[1], Myungnam Bae[2], Myunggeun Chung[2], Myungeun Lim[2]
[0,1]Division of Electronics and Information Engineering, National Chonbuk University
[2]Electronics and Telecommunications Research Institute

요 약

본 논문에서는 온톨로지를 활용하여 생물학 데이터를 효율적으로 통합 검색하기 위한 XQuery 확장 시스템을 설계하고 구현하였다. 이를 위해 본 논문에서는 먼저 공개 생물학 온톨로지 등인 GO, UMLS들로부터 의미 있는 정보를 추출하기 위한 생물학 온톨로지 API를 온톨로지별로 정의하였다. 정의된 온톨로지 API는 본 시스템에서 사용하는 XQuery의 사용자 정의 함수로써 포함되며, 이 XQuery는 본 시스템에 내장된 XQuery Expander에 의해 확장되어 처리된다. 확장된 XQuery는 온톨로지를 이용함으로써 이질적인 구조와 용어로 이루어진 생물학 데이터들을 통합 검색 할 수 있으며, 온톨로지에 정의되어 있는 지식과 관계들을 확장검색에 활용함으로써 재현율을 획기적으로 높일 수 있다. 본 논문에서는 또한 XQuery의 작성을 용이하게 할 수 있도록 지원하는 GUI 환경도 구현하였다.

1. 서 론

생물정보학(Bioinformatics)은 컴퓨터를 이용해 생물학 정보를 저장, 분석, 검색, 계산하는 학문이다. 최근 생물정보학 연구가 크게 증가함에 따라 GenBank, Swiss-Prot, PDB(Protein Data Bank) 그리고 BIND(the Biomolecular Interaction Network Database)와 같은 많은 생물학 데이터베이스가 구축되었다.

생물학 데이터베이스들은 여러 개별적 연구로 구축되었기 때문에 이질적인 관점과 서로 다른 데이터 구조를 가진다[1]. 이는 여러 생물학 데이터베이스의 효과적인 통합 검색을 어렵게 하는 원인이 된다. 이러한 문제를 해결하기 위해서 [2][5][6] 등의 시스템들이 개발되었다. 이들 시스템은 생물학 데이터베이스 간의 이질적 구조에 대해 일정한 데이터 접근 방식을 제공함으로써 통합 검색의 투명성(transparency)을 제공한다.

그러나 이 시스템들을 이용하여 생물학 데이터베이스들을 검색할 경우, 생물학 용어들 사이에 존재하는 의미적 불일치 때문에 재현율이 떨어진다는 단점이 있다. 이에 대한 한 해결책은 온톨로지를 생물학 데이터베이스 검색에 도입하는 것이다. 온톨로지는 지식 도메인에 관한 개념적인 명세로, 도메인에 대한 전문적인 지식을 용어와 용어간의 관계로 정의한다. GO(Gene Ontology)[3]와 UMLS(Unified Medical Language System)[4] 같은 공개 생물학 온톨로지들은 생물정보 데이타베이스의 주석(annotation) 표현이나 세부적인 생물용어들의 계층적 관계표현과 같은 생물학 지식을 표현하고 있기 때문에 이들을 검색에 적절히 활용한다면, 이 시스템들의 재현율을 획기적으로 높일 수 있다.

본 논문에서는 생물학 온톨로지를 이용한 XQuery 확장 시스템을 설계하고 구현하였다. 본 시스템에서는 XML Wrapper을 이용하여 생물학 온톨로지와 생물학 데이터베이스를 XML 문서로 처리하며, 이러한 XML 데이터를 검색하기 위해서는 XQuery를 이용한다.

본 논문에서는 ETRI의 XML Wrapper[2]를 이용하여 공개 생물학 온톨로지의 데이터를 획득하고, 이 데이터로부터 의미 있는 지식을 추출하는 API를 구현했다. 정의된 온톨로지 API는 본 시스템에서 사용하는 XQuery의 사용자 정의 함수로써 포함되며, 이 XQuery는 본 시스템에 내장된 XQuery Expander에 의해 확장되어 처리된다. 또한 생물학 온톨로지를 이용해서 XQuery의 작성을 용이하게 할 수 있도록 지원하는 GUI환경도 구현하였다.

본 논문의 순서는 다음과 같다. 2장에서는 기존에 개발된 생물학 통합 검색 시스템의 장단점을 살펴본다. 3장에서는 본 연구에서 설계한 시스템에 관해서 소개하고 4장에서는 프로그램으로 구현해서 실질적인 적용 방식을 살펴본다. 마지막으로 5장에서는 결론 및 향후 연구 방향에 대해서 정리한다.

2. 관련 연구

생물학 데이터의 통합 검색을 위해 [5][6]같은 생물학 데이터베이스 통합 검색 시스템이 개발되었다.

[5]는 질의 변환기(Query Reformulator), 질의 처리기(Query Engine), 중개자(Mediator)로 구성되어 있다. PQL(a Path Based Query Language)을 이용해 생물정보 검색을 위한 질의를 작성한다. PQL은 질의 변환기에서 XQuery로 변환되며, 중개자에서 데이터 소스 정보 저장소(Source Knowledge Base)를 참조해 전역 스키마(Global schema)와 목적 스키마(Target schema)의 맵핑을 통해서 각각의 생물학 데이터베이스에 맞는 질의를 생성해서 결과를 가져온다. 이 시스템은 데이터 통합을 전역 스키마와 목적 스키마의 맵핑을 통해서 해결을 시

도했다. 이 시스템의 장점은 사용자들이 PQL이라는 질의언어만 알고 있으면 여러 생물학 데이터베이스를 투명하게 검색 할 수 있다는 것이다. 또한 검색할 데이터베이스가 추가되면 단순히 대상 데이터베이스에 대한 맵핑 스키마를 추가하면 검색 대상이 확장된다.

[6]은 [5]처럼 생물학 데이터베이스들의 통합 검색을 지원한다. [6]은 [5]와 유사한 면이 많이 있다. 연구에서 정의한 질의 언어에 맞춰 질의를 작성하면 각 데이터 소스에 맞는 질의로 변환해서 각각의 생물학 데이터베이스를 검색한다.

그러나 [5][6]의 연구에서는 온톨로지를 전역 스키마와 대상 스키미간의 연관에만 사용하였다. 그 결과 온톨로지의 지식을 이용한 검색 방식을 제공하지 못하였다. 즉, 이질적인 생물학 데이터베이스들의 투명한 이용은 가능해도 각 데이터베이스간의 일관되지 못한 용어 사용으로 인한 검색 실패는 해결하지 못하였다.

3. 생물학 온톨로지 API와 XQuery 확장 시스템

3.1 시스템 구조

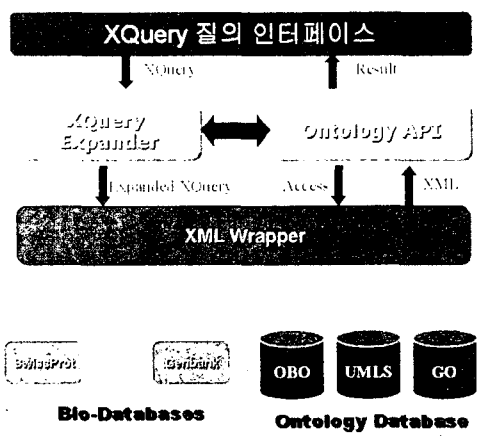

그림 1. 시스템 구조도

이 장에서는 생물학 온톨로지를 이용하여 XML 문서로 작성된 생물학 정보를 검색하는 시스템의 설계 내용을 기술한다. 그림 1은 본 XQuery 확장 시스템의 구조도이다. 온톨로지 API는 Gene Ontology, UMLS 등 공개 생물학 온톨로지에서 유용한 지식을 추출한다. XQuery Expander는 XQuery를 확장하는 모듈로 동작은 다음과 같이 작동된다. 입력받은 XQuery를 파싱하고, XQuery에서 본 연구에서 정의한 온톨로지 API와 대응되는 XQuery 사용자 정의 함수가 사용되는지 분석한다. 그리고 해당 XQuery 사용자 정의 함수와 대응되는 온톨로지 API를 호출하여 온톨로지 API의 결과 값을 이용하여 XQuery를 확장한다. 본 연구에서는 ETRI에서 개발한 XML Wrapper를 이용해서 온톨로지 소스와 생물학 데이터베이스 등에 접근한다. XML Wrapper는 생물학 데이터베이스에 대한 프로토콜, 데이터 구조, 접근 경로등의 정보를 관리하여, 사용자에게 일정한 데이터 접근 방식을 제공함으로써, 생물학 데이터베이스의 투명한 접근을 지원한다.

3.2 생물학 온톨로지 API

생물학 온톨로지 API는 크게 온톨로지 용어 정보 탐색, 온톨로지 용어에 관련된 경로 확장, 온톨로지 용어들 사이의 추론 연산 3가지로 추상화 했다.

생물학 온톨로지 정보 탐색은 용어 정보에 관한 탐색으로 용어의 정의, 동의어, 분류, 외부 참조 등에 관한 함수들이다. 생물학 온톨로지 용어 경로 확장은 부모, 자식, 형제, 소유 등 온톨로지의 용어에 대한 관계 탐색을 위한 함수들이다. 생물학 온톨로지 용어들 사이의 추론 연산은 Union, Intersection, LUBS(Least Upper Bound Set), GLBS(Greatest Lower Bound Set) 등의 연산이 있다. 그림 2는 구현된 생물학 온톨로지 API를 열거하고 있다.

정보 탐색	• string getTerms (string Term) • set getID (string Term) • string getDefinition (string Term) • string getAspect (string Term) • set externalReferences (string Term)
	• set getSynonyms(string Term)
경로 확장	• set getParents (string Term) • set getChildren (string Term) • set getPartOfTerms (string Term) • set getOwnerOfTerms (string Term) • set getAncestors (string Term) • set getSiblings (string Term)
용어들 사이의 추론 연산	• set intersection(set IDs1, set IDs2) • set union(set Terms1, set Terms2) • set getLUBSTerms(string Term 1,string Term 2) • set getGLBSTerms(string Term 1,string Term 2)

그림 2. 생물학 온톨로지 API

3.3 생물학 온톨로지 XQuery 사용자 정의 함수

앞에서 정의한 온톨로지 API를 XQuery에서 이용하기 위해서 생물학 온톨로지 XQuery 사용자 정의 함수를 생물학 온톨로지 API와 1:1대응이 되도록 정의한다. 이렇게 정의된 XQuery 사용자 정의 함수는 XQuery에 삽입되어 XQuery Expander에서 처리하게 된다.

3.4 XQuery Expander

XQuery Expander는 생물학 온톨로지 API와 생물학 온톨로지 XQuery 사용자 정의함수를 이용해서 입력된 XQuery를 확장시키는 모듈이다. XQuery가 입력되면 XQuery Expander에서 파싱을 한 후, 본 연구에서 정의한 XQuery 사용자 정의 함수가 존재하면, 그에 대응하는 생물학 온톨로지 API를 이용해서 생물학 온톨로지를 이용한 XQuery로 확장한다.

```
for     $i in doc("Genbank.xml")/genbank_db/genebank_entry
where  $i/keywords = $OBO:GO:getPartofTerms("GO:0031234")
return  <Genbank>
    <keyword >
      { $i/keyword }
      <definition>{$i/definition}</definition>
    </keyword>
</Genbank>;
```

Expaned XQuery

```
for     $i in doc("Genbank.xml")/genbank_db/genebank_entry
where  $i/keywords = "trailing edge membrane"  or
         $i/keywords = "uropod"
return  <Genbank>
    <keyword >
      { $i/keyword }
      <definition>{$i/definition}</definition>
    </keyword>
</Genbank>;
```

그림 3. XQuery 확장 예.

그림 3은 본 연구에서 설계한 XQuery 사용자 정의 함수를 포함한 XQuery가 어떻게 확장 되는지 보여준다. 밑줄 친 $OBO:GO:getPartOfTerms("trailing edge") 라는 XQuery 사용자 정의 함수는 Gene Ontology의 온톨로지를 이용해서, "trailing edge"와 part-of 관계를 가진 용어를 추출한다. 그림 3의 밑줄 친 부분은, 이렇게 추출한 용어를 이용해서 XQuery를 확장한 것을 보여준다.

3.5 생물학 온톨로지 API 사용 예

API 종류	사용 예	결과
1.용어 정보 탐색	getSynonyms("mitochondrial processing peptidase activity")	alpha-mitochondrial processing peptidase , beta-mitochondrial processing peptidase
2.용어 경로 탐색	getAncestors('DNA helicase activity')	DNA helicase activity, helicase activity, catalytic activity,
3.추론 연산	getLUBS('motor axon guidance','wing vein specification')	organ development

그림 4 생물학 온톨로지 사용 예

그림 4는 온톨로지 API의 사용 예이다. 여러 가지 API 중에서 몇 개의 API에 관해서 예를 들어보았다. 첫째, 용어 정보 탐색에서 getSynonyms("mitochondrial processing peptidase activity")는 온톨로지 함수를 이용해서 mitochondria processing peptidase activity와 동의어 관계에 있는 용어를 추출한다. 이렇게 추출된 동의어를 이용하면 검색 재현율을 높일 수 있다. 둘째, 용어 경로 탐색에서는 DNA helicase activity의 조상에 해당하는 용어를 가져온다. 용어 경로 탐색을 통해 특정 관계에 있는 용어를 탐색해 지식을 확장 할 수 있다. 셋째, 두 용어사이의 공통된 조상 중 최단거리에 있는 용어를 추출하는 예이다. 이런 추론 연산을 통해 의미 있고, 고차원적인 결과를 추출할 수 있다.

4. 구현

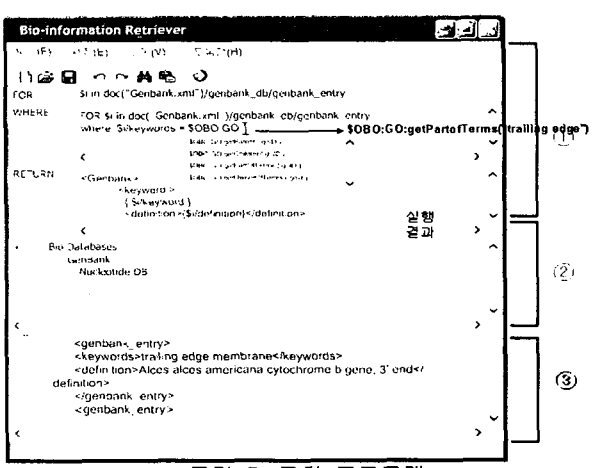

그림 5. 구현 프로그램

지금까지 정의한 생물학 온톨로지 API와 XQuery 사용자 정의 함수를 이용해서 XQuery를 작성하고 생물학 데이터를 검색하는 GUI를 구현하였다. 구현 환경은 JAVA SDK 1.4이고 IDE툴은 Eclipse 3.0, GUI 라이브러리는 SWT[7]를 이용하였다.

그림 5의 ①은 XQuery 입력 부분으로 XQuery FLWR 3부분으로 나눠서 입력의 편의성을 도모했다. 또한, 본 연구에서 정의한 생물학 온톨로지 XQuery 사용자 정의 함수를 입력 할 경우 선택 목록이 나와서 입력의 편의성을 도모했다. ②는 검색할 XML 문서에 대한 정보를 보여준다. 그림 5의 ②는 GenBank의 nuecleotide DB를 대상으로 했다는 것을 의미한다. ③은 XQuery로 검색한 결과가 출력되는 부분이다.

구현한 바이오 정보 탐색기는 XQuery 작성을 도와주고, 생물학 온톨로지를 이용한 XQuery 확장 그리고 확장된 XQuery를 이용한 결과를 한눈에 보여줌으로써 생물학 정보 검색에 편의를 제공하였다.

5. 결론 및 향후 연구

본 시스템은 여러 생물학 데이터베이스간의 통합 검색의 어려움을 해결하였다. 생물학 온톨로지 API를 제공하여 공개 생물학 온톨로지를 검색에 이용하였고, 온톨로지의 유용한 지식을 검색에 활용함으로써 검색의 재현율을 획기적으로 높였다. 또한, XQuery 작성을 효율적으로 도와주고 생물학 데이터를 검색하는 GUI를 구현하였다.

추후 연구 과제로는 이용 가능한 공개 온톨로지 대상을 확대하고, 공개 온톨로지간의 비교를 통해서 더욱 정확하고 의미 있는 지식을 제공하는 연구가 필요하다. 그리고 더욱 고차원적이고, 의미가 있는 API를 추가하는 연구가 필요하다. 또한, 검색 효율성을 높이기 위한 메모리 구조와 검색 기법에 관한 연구도 필요하다.

6. 참고문헌

[1]Francois Bry, Peer Kroger, "A computattional Biology Databse Digest: Data, Data Analysis, and Data Management", Distributed and Parallel Databases, Vol. 13, Issue 1, Page 7-42, 2003.

[2]Myungeun Lim, Myunggeun Chung, Myungnam Bae and Sunhee Park,"Design of a description language for generating wrapper to collect biological data",Bioinformatics Vol. 21 Suppl. 2 , pages ii1-ii3, 2005

[3]Gene Ontology Consortium, http://www.geneongtology.org

[4]Unified Medical Language System,http://umlsks.nlm.nih.gov

[5]Loren Donelson, PeterTarczy-Hornoch, PeterMork Cindy Dolan, Joyce A.Mitchell, M.Barrier, Hao Mei, "The BioMediator System as a Data Integration Tool to Answer Diverse Biologic Queries", Medinfo., Page 768-72, 2004.

[6]Patrick Lambrix, Vaida jakoniene, "Towards transparent access to multiple biological databanks", Conferences in Research and Practice in Information Technology Series; Vol. 33 , Conferences in Research and Practice in Information Technology Series; Vol. 33, Pages: 53 - 60 , 2003

[7]SWT, http://www.eclipse.org/swt/