온톨로지를 적용한 생명정보 검색 시스템

이동훈°, 양정진
가톨릭대학교
{rephaser°, jungjin}@catholic.ac.kr

Towards Bio-Information Retrieval using Ontology

DongHun Lee°, Jungjin Yang
School of Computer Science and Information Engineering
The Catholic University of Korea

요 약

OMIM 이나 MedLine과 같은 바이오 정보를 포함한 대규모의 데이터베이스에서의 바이오 정보검색이나 추출은 단계별 처리가 상호적인 운용에 의한 것이 아니라 수동 또는 Perl script에 의한 것이 대부분이다. 바이오정보(bio-information)의 효과적 추출과 유추를 위해서는 현재 상이한 스키마로 이루어진 다양한 데이터베이스들 간의 이질적 데이터(heterogeneous data)를 체계적이고 효율적으로 통합하는 표준 방안이 필요하다. 이를 위해서는 서로 다르게 표현된 다양한 개념간의 관계를 표현하는 지식체계가 필요하다. 본 연구에서는 이러한 지식체계인 온톨로지(ontology)와 자원 메타정보(metadata)를 표현하기 위한 국제적 표준안으로 대두되고 있는 시맨틱 웹(semantic web)에서 제공하는 온톨로지, 메타정보, 스키마 통합 안을 발전적으로 적용하여 이질적 바이오 정보의 효율적 통합처리 방안을 제시하고자 하였다.

1. 서 론

OMIM[1] 이나 MedLine[2]과 같은 바이오 정보를 포함한 대규모의 데이터베이스에서의 바이오 정보검색이나 추출은 단계별 처리가 상호적인 운용에 의한 것이 아니라 수동 또는 Perl script에 의한 것이 대부분이다. 바이오정보(bio-information)의 효과적 추출과 유추를 위해 현재 상이한 스키마로 이루어진 다양한 데이터베이스 간의 이질적 데이터(heterogeneous data)를 체계적이고 효율적으로 통합하는 표준 방안이 필요하다. 이를 위해 서로 다르게 표현된 다양한 개념간의 관계를 표현하는 지식체계가 필요하다. 본 연구에서는 이러한 지식체계인 온톨로지(ontology)와 자원 메타정보(metadata)를 표현하기 위해 국제적 표준안으로 대두되고 있는 시맨틱 웹(semantic web)에서 제공하는 온톨로지, 메타정보, 스키마 통합 안을 적용하여 이질적 바이오 정보의 효율적 통합처리 방안을 제시한다.

대부분의 사용자는 사전에 준비되어진 세세한 의학적 지식이나 이 분야의 검색을 위한 특정 환경이 없는 상황에서 검색목적에 적합한 정형화된 질의를 준비하는 데에 어려움을 나타내고 있다. 효과적인 검색을 위해서는 대부분 부분적으로 결과를 확인하고 반복적으로 질의를 고쳐가며 필요한 정보를 찾고 있는 것이다. 본 연구의 목적은 이러한 사용자의 부담을 덜고 사용자의 의도를 적절히 파악하여 관련된 정보를 제공하여 정보검색의 질을 높이는 데에 있다. 특히, 의료정보와 같이 일반사용자에게 부과되는 특정 용어나 개념들로 인한 부담을 줄이기 위해 온톨로지 기반의 적극적인 사용자 인터페이스 에이전트(proactive user interface agent)를 통한 의미적인 정보검색에 초점을 두고 있다.

이는 자원이나 특성(property)에 관한 메타정보, 개념과 특성들의 계층적인 관계를 나타내는 온톨로지와 이를 바탕으로 하여 추론을 통한 정보추출용의 논리적인 온톨로지를 포함한다. 또한 의미검색의 결과에 대한 내용 분석을 통하여 선별된 정보를 제공한다. 이러한 시스템의 혜택은 의학지식이 없는 사용자에게만 있는 것이 아니라 의학 온톨로지(medical ontology)와 질의 모델이 있는 타스크 모델 온톨로지, 바이오 인포메틱스 온톨로지에 전문지식을 부여 하는 입장의 의학 전문가에게도 MedLine을 포함한 최신의 웹 정보를 제공하는 등의 혜택을 제공한다.

2. 관련연구

2.1 시맨틱 웹 기술 적용의 온톨로지 기반 지식 처리

인터넷상에는 요구된 질의에 적합한 웹 페이지를 찾아주는 일반적인 목적을 가진 검색 엔진들이 많이 있다. 반면에 MedLine과 같은 특정분야의 문헌정보를 찾는 PubMed라는 전문적인 분야의 검색 엔진도 존재하지만 적절한 정보를 얻기 위해서는 그에 맞는 전문지식을 필요 한다. 이는 사용자에게 부담되는 요소가 크다.

온톨로지는 최근에 지능적인 정보통합, 협동적인 정보시스템, 정보검색, 전자상거래, 지식관리 등의 연구분야에서 웹 정보에 관한 사람과 응용시스템사이의 공유된 지식과 공통된 해석이라는 차원에서 더욱 관심을 받고 있다.

2.2 생명정보 분야의 데이터 표준화를 위한 시스템 통합 문제

생명정보학의 다양한 데이터베이스들을 용도별로 크게 구분해 보자면 아래와 같은 그룹으로 나누어 볼 수 있다.

- annotation search : search for keywords, authors, features
- similarity (homology) searches : search for similar sequence
- pattern searches : search for occurrences of patterns
- predictions : using the databases as knowledge DB
- comparisons : gene families, phylogenic trees

이러한 문제와 관련된 데이터베이스들을 타입별로 구분해 보면 아래와 같이 나누어 볼 수 있다.

- sequence DB : EMBL, Genbank, DDBJ, Swissprot, PIR,

TREMBL
- genome DB : GDB(Genome Data Bank), OMIM(Online Medelian Inheritance in Man)
- pattern DB : Prosite(Protein sequence motif), Blocks, TFD(Transcription Factor DB)
- family DB : Gene families, pfam(Protein families)
- structure DB : PDB(Protein Data Bank-3D structure)
- expression patterns DB : Fly view-gene expression
- literature DB : Medline, Current Contents
- trends in DB : Cross reference, graphical interface, java & applets

본 논문에서는 기존의 전문가 시스템에서의 지식 분석 처리 과정을 일반적인 지식과 전문적인 지식이 표현된 온톨로지 기반과 이를 표준화된 시맨틱 웹 언어로 나타내어 지식 처리에 적용하였다. 웹 상에 존재하는 다른 바이오 정보 데이터베이스 등과 연계하여 이를 바이오 관련 지식을 내포하고 있는 문헌정보 데이터베이스인 MedLine에서 보다 효과적이고 적절한 관련자료를 찾아주는 정보검색 시스템 개발에 적용하였다.

3. 방법론

시맨틱 웹 분야는 RDF, RDF Schema, DAML+ OIL 및 OWL과 같은 마크업 언어를 통하여 공유되는 지식인 온톨로지를 개발하고 이를 데이터에 적용하여 추론하는 주제들을 중점적으로 다룬다. 국제적 표준화 작업에 해당하는 W3C(World Wide Web Consortium)를 중심으로 계속 개발, 변화, 진화하여 온 시맨틱 웹 언어를 지속적으로 연구, 학습하였으며 이를 의료정보와 의료정보 문헌검색 시스템에 적용하였다. 서로 상이한 애플리케이션들에 적용되는 규칙 시스템 또한 서로 다른 구조와 형태로 되어 시맨틱 웹 기술의 다음단계로 상호운용성 (Interoperability)을 향상시키는 것이 관건이다. 이를 위한 규칙 기반 비즈니스 인텔리전스를 적용하기 위하여 다음의 추론 엔진과 상이한 규칙들의 변환을 아래 시스템들의 적용으로 꾀하였다.

- Jess (Java Expert System Shell) : Rete 알고리즘을 활용하는 규칙 엔진 시스템 [3]
- RuleML : XML 규칙들을 표현하기 위하여 W3C에서 추구하는 산업표준화 마크업 언어
- Jena API: facts와 규칙을 표현하기 위하여 Java에서 제공하는 APIs[4]
- SweetJess: 상호운용적인 지식처리를 위하여 ruleML, DAMLML등 다른 형태의 규칙들을 변환하는 시스템

4. 시나리오

마약중독자가 어느 순간부터 귀가 잘 들리지 않아서, 인터넷을 통하여 원인을 파악하려 한다. 이때 마약중독자는 MedLine 등의 의학 데이터베이스에 PubMed 인터페이스를 통하여 접속하고 'Deafness'라는 검색어로 검색을 하게 된다. PubMed만을 사용하여 검색한 경우 3만개 이상의 검색결과로 정확한 확인이 불가능한 반면 본 시스템에서는 'Deafness'라는 검색어가 들어오게 되면 UMLS등의 온톨로지 라이브러리를 통하여 Deafness와 관련된 온톨로지를 추출하고, 미리 정의되어 있는 규칙 중 병명과 유전자관계에 있는 규칙을 UMLS에 적용하여 얻은 결과로부터 'Cockayne Syndrome'이 Deafness와 관련이 있다는 사실을 도출한다.[그림 1.] 이를 통해 마약중독자는 자신의 귀가 들리지 않는 증상이 마약과 관련이 있다는 사실을 알게 되고, 질의를 재생성[그림 4.]함으로써 자신이 필요로 하는 결과를 보다 정확하게 MedLine등으로부터 얻을 수 있다.

5. 구현

본 연구에서는 Protege[5]에 UMLS Tab과 Jess Tab를 Plug-in하여 사용하였다. 먼저 UMLS Tab을 이용하여 Deafness를 사용할 온톨로지로 Deafness 클래스를 등록한다.

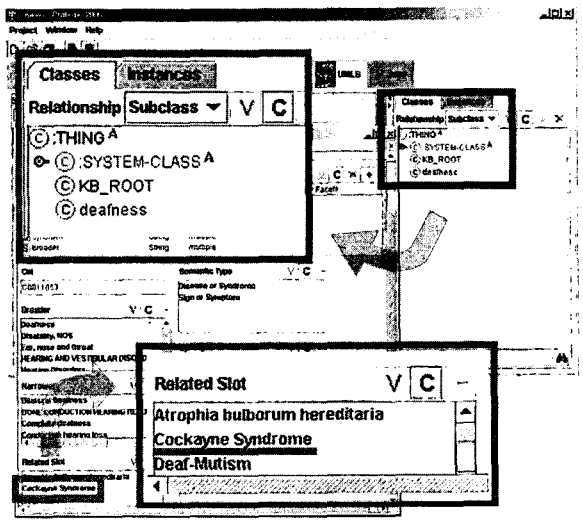

그림 1.

등록된 Deafness 클래스는 Jena를 통해 Jess에서 사용하게 될 Fact 형태로 변환된다. UMLS에 적용하여 얻은 결과로부터 'Cockayne Syndrome'이 Deafness와 관련이 있다는 사실을 도출하는 과정이 그림 1에 해당한다. 다음으로 미리 정의되어 있는 Rule은 SweetJess를 통하여 변환된 JessRule과 Jena를 통해 얻은 Fact를 Jess에 입력하는 과정이 그림 2이다.

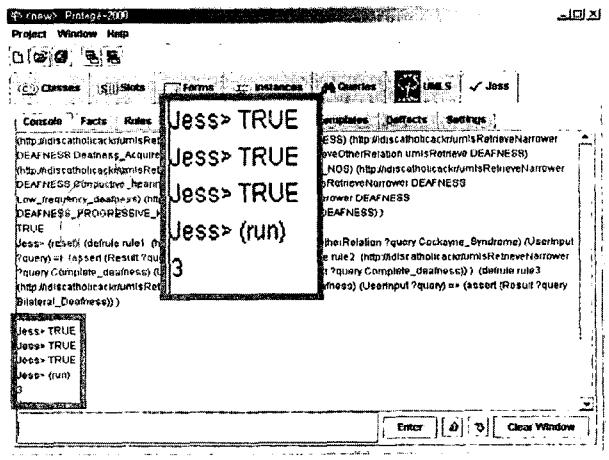

그림 2.

입력된 Fact와 Rule을 이용하여 Jess를 실행하면 새로운 3가지의 질의어를 얻을 수 있는데 그 중 'Cockayne Syndrome'이 포함되어 있는 것을 알 수 있다.

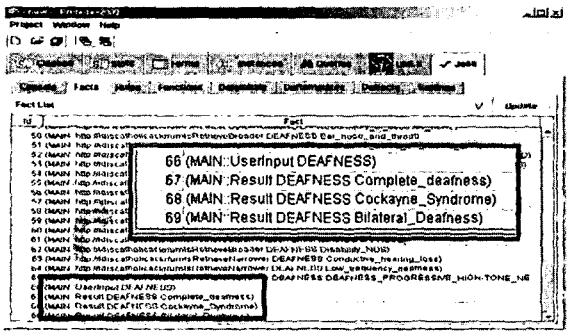

그림 3.

사용자는 그림 3에서와 같이 재생성 된 질의어를 통해 자신이 원하는 문헌을 검색하게 된다.

6. 결과 및 고찰

이번 장에서는 본래의 사용자 질의(질병)와 5장에서 온톨로지를 적용하여 생성된 재질의(질병과 관련유전자)를 다른 바이오데이터베이스에 적용하여 검색된 결과를 비교, 분석하였다. MedLine 검색을 위한 PubMed 인터페이스를 통해 "Deafness"라는 질의를 한 경우 PubMed는 30330개의 관련 문헌을 제공한다.

Relevancy level of $t_{(query\ term)}$: L(t)

$$L(t) = (0.5 + 0.5 freq_{t,q}/max freq_{t,q}) * \log n/m$$

TF (Term Frequency) — IDF (Inverse Document Frequency)

$freq_{t,q}$: Raw frequency of term k, in the document d_q
$max freq_{t,q}$: The highest $freq_{t,i}$
n : Total number of documents
m : Number of documents in which the index term k_i appears

같은 질의에 대하여 각 기 다른 주제로 데이터가 저장되어 있는 데이터베이스들을 통하여 병명과 유전자의 관계로 질의 한 결과가 그림 4와 5에 나타나 있다. 각 데이터베이스들을 통해 검색된 결과를 평가하기 위하여 TFIDF(Term Frequency/Inversed Document Frequency)를 위의 L(t)함수로 적용하였다. 그림 5는 LocusLink를 활용한 한 예를 보여주며 이 과정을 통하여 714개의 관련 문헌을 검색할 수 있다.

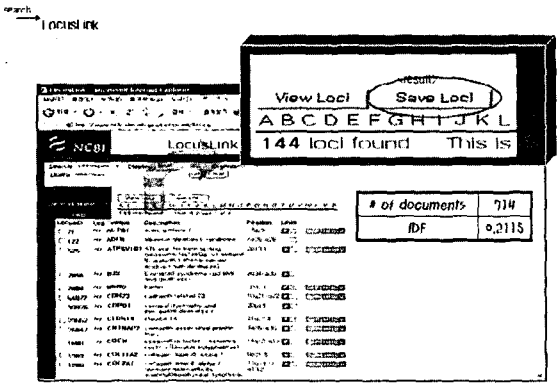

그림 4.

Ensembl을 통한 OMIM은 346개의 문헌이 제공된다.

그림 5은 OMIM, Ensembl, Hugi with LocusLink를 이용한 질의 결과를 측정 함수 L(t)로 비교한 것이다.

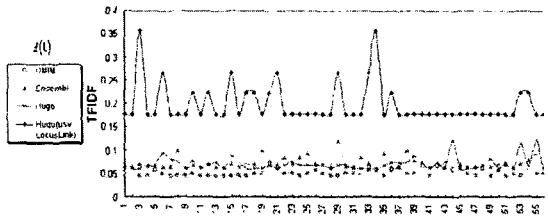

그림 5.

측정 결과는 LocusLink를 이용한 Hugo가 가장 최적의 결과를 나타내었다. 또한 다른 데이터베이스를 활용한 검색 방법들이 PubMed만을 이용한 검색 방법보다 더 효율적임을 알 수 있다.

여러 단계의 관련정보를 연계하여 검색한 경우 각 데이터베이스의 다른 색인 방법과 개별 아이디 등으로 연계된 정보가 많은 것이 바로 관련문헌의 높은 검색율을 보이지는 않았다. 이는 MedLine의 색인 방법과도 관련이 있는 것으로 여겨지며 단계별 연계정보 활용에 있어서 전적인 자동화보다는 전문가의 의견을 겸비하는 휴먼 팩토링(Human Factoring)의 이슈가 고려대상으로 여겨진다.

보다 진보된 검색엔진은 자치적인 학습 능력을 포함하는 것으로서 이러한 시스템들이 학습하는 방법과 이들 방식으로 학습된 내용이 얼마나 정확한 것인가 하는 것은 중요한 연구과제로 여겨진다. 다음 연구단계로는 에이전트 온톨로지에 베이지안 네트워크와 같은 학습능력을 갖춘 지식체계 표현을 함으로써 데이터마이닝적인 요소를 포함하는 것이며 사용자의 흥미, 선호도나 유전적인 요소에 이르기 까지 사용자 개인 정보를 포함한 사용자 프로파일 적용으로 관련된 정보를 연계하여 제공하는 시스템으로 확장하고자 한다.

향후 연구는 이번 연구에서 얻어진 결과와 분석, 경험을 토대로 새로운 메타정보 창출 및 온톨로지 생성에 중점을 두고 있다.

7. 참조문헌

[1] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
[3] JESS: Java Expert System Shell, http://herzberg.ca.sandia.gov/jess/docs/52/api/jess/Rete.html
[4] Jena, http://www.hpl.hp.com/semweb/doc/tutorial/DAML/
[5] Protege, http://protege.stanford.edu/
[6] M. Uschold and R. Jasper, A Framework for Understanding and Classifying Ontology Applications, In Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods.
[7] H. Boley, S. Tabet, G. Wagner, Design rationale of RuleML: A markup language for Semantic Web rules, in Semantic Web Working Symposium, 2001.
[8] TRIPLE Semantic Web, http://triple.semanticweb.org/, March, 2002