정보추출 기법을 이용한 서열정보분석 데이터베이스 구축 시스템 설계

이선아[0] , 전중남 , 이건명
충북대학교 컴퓨터과학과, 첨단정보기술 연구센터
bluebird@aicore.chungbuk.ac.kr, joongnam@cbucc.chungbuk.ac.kr, kmlee@aicore.chungbuk.ac.kr

System Design for Building Sequence Information Analysis Databases using Information Extraction Techniques

Sun-a Lee[0] , Joong-nam Jun , Keon-Myung Lee
Dept. of Computer Science, Chungbuk National University and AITrc

요 약

인터넷의 확산과 첨단기술의 발달로 생물학 정보에 대한 온라인 데이터베이스 집합이 급속히 증가하고 있으나, 데이터의 양이 방대하고 이질적인 형태로 제공되기 때문에 실제 현장의 생물학 연구자들이 쉽게 이용하는데는 여러 가지 어려움이 있다. 이 논문에서는 단백질과 핵산 정보를 제공하는 대표적인 온라인 데이터베이스인 NCBI에, 질의를 하여 얻어진 데이터를 포함한 웹 문서로부터, 정보를 추출하여 사용자의 목적에 적합한 맞춤형 데이터베이스를 구축하는 시스템을 제안한다. 온톨로지를 이용하여 질의 처리를 하며, 웹 문서에 대한 정보추출 기법과 계층구조에 따른 유형별 저장방식을 통해 데이터베이스를 구축한다. 한편, 데이터 추출을 위해 식별 및 분류 작업을 수행한다. 제안한 시스템은 서열정보를 분석하는 생물학자들에게 관심대상 정보를 추출하여 맞춤형 데이터베이스를 구축함으로써, 손쉽게 서열정보 분석을 지원하도록 하는데 목적이 있다.

1. 서론[1)]

오늘날 인터넷의 확산과 첨단 기술의 발달로 서열구조, 분자상호작용, 표현 패턴과 같은 생물학적 정보에 대한 온라인 데이터베이스들이 급속히 증가하고 있으나, 생물학자들이 이러한 정보에 손쉽게 접근하는 데에는 여러 가지 어려움이 많다[1, 2]. 생물학 정보를 제공하는 데이터베이스에는 PDB[3], GenBank, NCBI[4], EMBL[5], DDBJ[6], SWISS-PROT 등이 대표적이다. 이 데이터베이스들은 방대한 데이터를 제공하지만 각각 제공하는 데이터형식이 다르기 때문에 이용하기 쉽지 않다.

이러한 온라인 데이터베이스들에 접속하여 관련된 모든 데이터들을 분석하기 위해서는 많은 시간과 노력이 필요하므로, 방대한 데이터에 대해 목적에 따른 데이터 추출이 필요하게 되었다. 더불어, 생물학 데이터의 급속한 증가에 따른 대처방안이 요구되고 있다. 이 논문에서는 단백질과 핵산정보를 제공하는 NCBI에 질의를 하여 검색된 데이터를 포함한 웹 문서에서, 정보를 추출하여 맞춤형 데이터베이스를 생성하고, 손쉬운 분석 인터페이스를 제공하는 시스템을 제안한다. NCBI는 미국의 국립보건원(NIH) 산하 기관으로 분자생물학 정보의 가공 및 처리를 목적으로 1988년에 설립되었으며, '데이터베이스 공개', '유전자 데이터 분석용 소프트웨어개발', '타정보 처리' 등의 업무를 수행하고 있다. NCBI에서 유전자 및 단백질 관련 데이터가 온라인 데이터베이스로 공개되고 있으며, 온라인을 통해서 질의를 하고 결과가 웹 문서 형태로 제공된다[4]. 제안한 시스템을 통해 맞춤형 데이터베이스를 구축하게 되면, 사용자의 목적에 맞는 데이터들의 추출로 인해 메모리를 절약할 수 있고 업데이트를 자동화함으로서 새로운 정보에 대해 적극적으로 대처할 수 있다. 또한, 손쉬운 분석 인터페이스를 제공하여 검색의 정확도를 높이고 검색시간을 단축할 수 있다. 기존의 NCBI는 다량의 웹 문서를 검색하여 주지만, 해당 문서간의 정보에 대해 비교, 분석할 수 있는 인터페이스를 제공하지 않는다. NCBI에서 정보를 비교, 분석하기 위해서는 다른 툴을 이용하여야 하는 어려움이 있다. 또한, NCBI는 플랫 파일 형식의 문서에서 검색하기 때문에 검색 영역이 불분명하여 불확실한 데이터의 검출로 정확도를 떨어뜨린다. 제안한 시스템은 온톨로지를 이용한 질의 처리를 통하여 검색 효율을 높인다.

이 논문은 다음과 같이 구성된다. 2장에서 시스템의 구성요소별 기능을 설명하고 3장에서는 웹 문서에서 정보추출을 위한 데이터 식별과 분류에 대한 방안을 제시한다. 4장은 맞춤형 데이터베이스의 구축 방법을 기술한다. 5절에서는 시스템의 향후 개발 방향을 제시하고 관련 연구에 대한 기여도를 논한다.

2. 시스템 구성

시스템은 크게 질의 처리 서브시스템, 맞춤형 정제 및 시각화 서브시스템, 맞춤형 데이터베이스, 데이터 확인 및 추출 서브시스템으로 나뉜다. 사용자의 질의를 맞춤형 데이터베이스 내에서 검색하여 결과를 보여준다. 검색 결과는 사용자가 정의한 분석환경에 맞게 보여준다. 맞춤형 데이터베이스에서 검색이 이루어짐과 동시에 NCBI에 질의하여, 새로이 등록된 내용을 업데이트한다. 시스템 구성도는 [그림 1]과 같다.

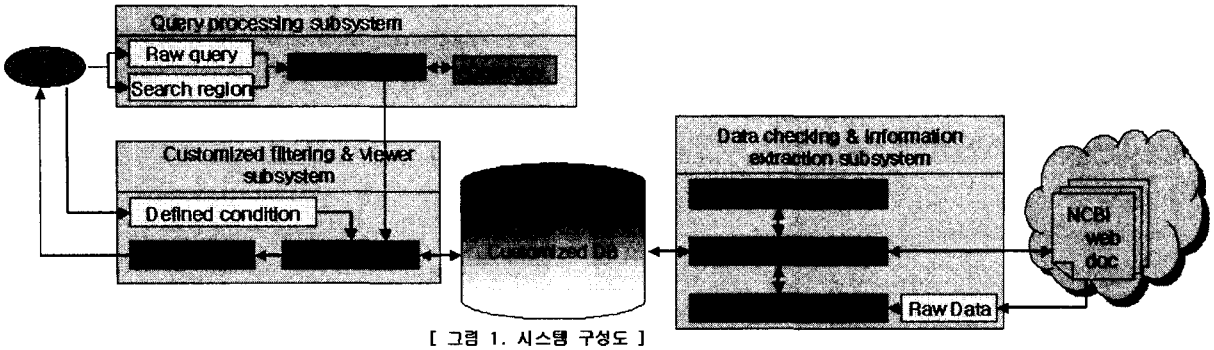
2.1 질의 처리 서브시스템

질의 처리 서브시스템은 사용자가 입력하는 질의를 정제함으로써 정보검색의 정확도를 높이고 검색 시간을 줄인다. 자연어 형태의 질의는 부정확한 정보를 요구하는 모호성을 지니기 쉽기 때문에 온톨로지를 통해 일관성을 지니게 된다. 온톨로지의 데이터는 기본적으로 NCBI에서 제공하는 정보를 포함한다. 만약, 질의가 불확실하여 온톨로지를 사용할 수 없는 경우에는 유사도를 판별하여 적절한 질의로 변환한다. 질의시 검색영역의 설정을 통해 검색효율을 높일 수 있도록 한다. 질의 처리 서브시스템은 정제된 질의를 맞춤형 정제 및 시각화 서브시스템으로 전달한다.

2.2 맞춤형 정제 및 시각화 서브시스템

이 서브시스템은 질의 처리 서브시스템에서 전달된 질의로

1) 이 논문은 첨단 정보기술 연구센터(AITrc)를 통해서 과학재단의 지원을 받은 것임



[그림 1. 시스템 구성도]

맞춤형 데이터베이스를 검색하여 해당 정보를 사용자가 요구하는 형태로 보여준다. 일반적으로 NCBI에서 제시되는 분석방법은 쉽지 않다. NCBI에서는 반복적인 질의를 통해 최상의 데이터를 얻어야 하는 반면, 맞춤형 정제 및 시각화 서브시스템에서는 정제단계에서 복합질의가 가능하며 정의된 조건에 따라 맞춤형 데이터베이스로부터 추출된 결과를 보여준다. 사용자는 지속적으로 조건을 정의할 수 있다.

2.3 맞춤형 데이터베이스

맞춤형 데이터베이스는 기본적으로 제시하는 데이터들과 개인이 정의한 데이터들로 구성된다. 시스템은 웹 문서로부터 정보를 추출하여 계층구조에 의한 분류를 통해 해당 테이블에 저장한다. 사용자의 질의가 있을 경우, 질의에 적합한 데이터를 추출하여 맞춤형 정제 및 시각화 서브시스템에서 사용자에게 결과를 보여줄 수 있도록 지원한다.

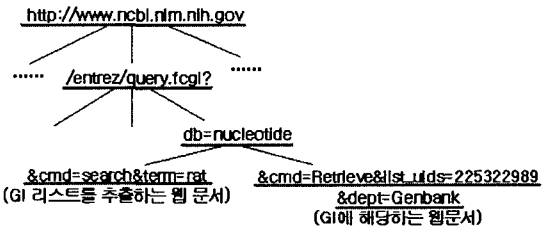
2.4 데이터 확인 및 추출 서브시스템

시스템은 초기 단계에 빈 데이터베이스를 가진다. 질의를 통한 정보추출로 데이터베이스를 구축하고 지속적으로 업데이트하는 것이 제안된 시스템의 특징이다. 사용자의 사용목적에 따른 데이터를 저장함으로써 메모리를 줄인다. 누락된 내용이 존재할 수 있으나 지속적인 업데이트를 통해 데이터를 보완한다.

데이터 확인 및 추출 서브시스템은 업데이트에 필요한 과정이다. 맞춤형 데이터베이스에서 질의에 해당하는 내용을 검색하여 사용자에게 보여준다. 맞춤형 데이터베이스에서 검색이 이루어지는 동안 내부적으로 같은 질의를 NCBI에 전달한다. 질의 결과가 서로 다른 경우, 정보 추출을 통해 맞춤형 데이터베이스를 보완한다.

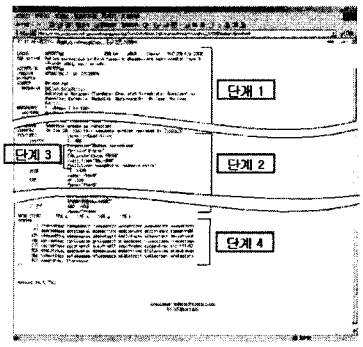
3. 데이터 식별과 분류에 의한 정보추출

정보추출의 대상은 NCBI에서 제공하는 단백질과 염기서열 정보이다. NCBI에서의 정보추출은 웹 문서에 접근하는 문제와 해당 웹 문서 내의 데이터 식별과 분류 문제의 해결을 통해 가능하다. 웹 문서로의 접근 문제는 URL을 통해 해결할 수 있다. 웹 문서에 접근하는 준비단계로 시스템의 시작과 함께 NCBI에 접속하여야 한다. NCBI의 웹 문서가 열리지 않은 상태에서는 URL을 통한 접근이 불가능하기 때문이다. URL을 이용해 추출할 데이터의 GI 리스트와 해당 웹 문서를 획득한다. GI 란 서열의 식별 번호이다. 예를 들어, 검색어가 'Nucleotide'와 'rat'이라면 [그림 2]와 같은 URL 경로를 유도할 수 있다. 웹 문서로의 접근은 'NCBI(http://www.ncbi.nlm.nih.gov) -> GI 리스트 추출 -> GI에 해당하는 웹문서' 순서로 이루어진다. [그림 2]에서 처음 단계와 두 번째 단계를 기본으로 한다. GI 리스트의 추출은 세 번째와 네 번째 단계의 왼쪽 URL을 통해 가능하다. 추출된 GI 리스트를 이용해 GI 각각에 해당하는 웹 문서에 접근하여 해당 정보를 추출한다.



[그림 2. 검색어 'Nucleotide'와 'rat'으로 유도한 URL 경로]

해당 웹 문서에 접근한 다음에는 데이터 식별과 분류 문제가 있다. 해당 웹 문서는 형식적으로 HTML의 형식을 지니지만 HTML 태그로는 데이터의 식별이 쉽지 않다. 추출하려는 정보가 <pre>와 </pre> 태그 사이에 존재하기 때문이다. <pre> 태그는 문장을 쓴 그대로 브라우저 안에 나타내게 한다. 정보추출은 <pre>태그 안에 존재하는 데이터들의 위치상의 규칙을 이용한다. [그림 3]은 실제 정보추출의 대상이 되는 웹 문서를 보이고 있다.



[그림 3. 실제 웹 문서에서 보여지는 위치에 따른 단계]

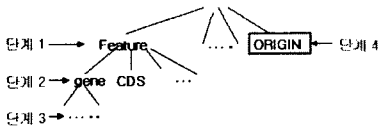
<pre>와 </pre> 태그 내에 포함된 정보는 식별자의 위치를 기준으로 5단계로 구분할 수 있다[표 1]. 1단계부터 4단계까지는 같은 문서이나 5단계는 다른 형식의 문서이다. 각 단계는 식별자와 속성을 지닌다. 각 단계별 식별자와 속성의 규칙은 다음과 같다.

식별자 = { [식별자가 존재하는 위치 범위], {식별자가 되는 값들의 집합} }
속성 = { [속성이 존재하는 위치 범위], {특수 문자(문자 나열은 '&' 이용)}, {식별자의 속성이 되는 값들의 집합} }

단계	키워드와 속성의 특성
1 단계	식별자 = { [1...11(12)],{Locus name,...[선택사항] } 속성 = {[12(13)...80],{'.'(마침표) & ';'},{[임의의 속성값] }
2 단계	식별자 = {[6...21],{[misc RNA],[repeat region],...}} 속성 = { 3 단계 }
3 단계	식별자 = { [22...80] , { depend_on(2) } } 속성 = { [3 단계의 키워드 위치 + 1(2)] , {'/' & ' "" ' & '<' & '>'}, { depend_on(2) } }
4 단계	식별자 = { [1...6] , { ORIGIN } } 속성 = { [4...9, 11...76] , ø , { 각 라인의 첫 번째 염기 번호, 서열기호[a , t , g , c] }
5 단계	식별자 = {[1... 15],{dbEST_Id, ..., [선택사항] } 속성 = { [16... 100],{'.' & '&'}, { 임의의 속성값 } }

[표 1. 단계별 키워드와 속성의 특성]

이때에 depend_on(단계)은 해당 단계에 의존하여 그 속성이 결정됨을 의미한다. [그림 3]에서 보여지는 단계별 식별자에 대한 계층구조는 [그림 4]와 같다.



[그림 4. 단계별 식별자의 층구조]

REFERENCE 식별자를 제외한 1 단계의 식별자는 강제적으로 추출한다. REFERENCE 식별자는 속성의 하위항목의 수가 불규칙하고 목적에 따라서 불필요할 수도 있기 때문에 선택항목으로 설정하였다. 2, 3 단계는 사용자가 정의하는 식별자에 대해서만 정보를 추출한다. 2, 3 단계에 존재하는 식별자들은 발생유무가 불규칙적이기 때문에 강제성을 띄게 되면 메모리의 낭비를 가져오기 쉽다. 4 단계의 ORIGIN 식별자에 해당하는 정보는 강제적으로 추출한다. ORIGIN 식별자는 검색으로 가장 많이 사용되는 부분이므로 반드시 추출하여야 한다. 5 단계의 정보추출이 되는 대상은 [그림 3]과는 전혀 다른 형태를 지닌다. 5단계의 웹 문서는 시스템 초기에 미리 추출할 식별자를 설정하여야 한다.

4. 맞춤형 데이터베이스 구축

이 논문에서 제안한 시스템은 서열정보 데이터베이스를 구축한다. 시스템은 데이터의 식별과 분류 과정을 통해 추출된 데이터들을 사용자가 설정한 조건에 맞추어 정제한 후, 저장한다. 데이터베이스에 저장되는 정보들은 사용자의 질의 특성, 질의 빈도 등에 따라 이용목적에 따른 특징을 지닌다. 질의를 통해 해당 데이터들의 정보만으로 구성되었기 때문이다. 계층구조에 의해 분류된 식별자들의 해당 속성값은 [표 2]에 따라 저장된다.

기본 테이블은 1단계 식별자들의 속성값을 저장하며 기본적인 서열정보를 나타낸다. 유전자 특성 테이블, Gene_CDS 테이블, 전이 테이블, ORIGIN 테이블은 각기 검색 대상이 되는 테이블로, 사용빈도가 높기 때문에 검색효율을 위해 나누어 관리한다. 특히, ORIGIN 테이블은 1단계와 같은 위치 1에서부터 시작하지만 종결점이나 속성의 위치가 달라 1단계와는 구분하여 관리할 필요성이 있다. 선택사항으로 1단계의 'REFERENCE' 식별자가 선택이 될 경우, 새로운 테이블을 만들어야 한다. 'REFERENCE' 식별자는 관련 논문에 대한 기록으로, 사용자의 목적이 논문검색일 경우, 사용빈도가 상당히 높기 때문이다.

테이블	저장되는 추출 정보의 단계
기본 테이블	1단계
식별유전자 테이블	5단계
유전자특성 테이블	2단계 식별자 중 gene, CDS를 제외한 나머지
Gene_CDS 테이블	2단계의 gene, CDS
전이 테이블	2단계 CDS에 대한 3단계 항목중 전이 식별자
ORIGIN 테이블	4단계

[표 2. 계층구조에 따른 저장 테이블 구성]

맞춤형 데이터베이스는 계층구조를 이용해 단계별로 정제된 데이터들을 데이터베이스에 저장하고 사용자의 질의가 있을 때에 해당 정보를 제공한다. 또한, 질의에 의해 추출된 정보의 수를 데이터 확인 및 추출 서브시스템에 보낸다. 데이터 확인 및 추출 서브시스템에서는 이를 이용해 NCBI의 질의 결과와 비교하여 데이터의 업데이트 여부를 결정한다.

5. 결론 및 향후과제

이 논문에서는 단백질과 핵산 정보를 제공하는 대표적인 온라인 데이터베이스인 NCBI에, 질의를 하여 얻어진 데이터를 포함한 웹 문서로부터, 정보를 추출하여 사용자의 목적에 적합한 맞춤형 데이터베이스를 구축하는 시스템을 제안하였다. 제안한 시스템에서는 온톨로지를 통한 질의 처리를 하며, 웹 문서에 대한 정보추출 기법과 계층구조를 이용하여 데이터베이스를 구축한다. 제안한 시스템은 현재 구현 중에 있으며, 구현을 통해 유용성을 확인할 것이다. 이 시스템은 서열분석을 하고자 하는 생물학자들을 위해 생물학 정보가 있는 NCBI 문서로부터 정보를 추출하여 데이터베이스를 구축한다. 이 시스템은 생물학자들이 좀더 정확하고 효율적인 분석을 위한 중간과정이며 이 시스템을 통해 생물학자들이 서열분석을 하기에 좀더 편한 인터페이스를 제공하고자 한다.

6. 참고문헌

1. Chikashi Nobata, Nigel Collier and Jun-ichi Tsujii, Automatic Term Identification and Classification in Biology Texts, *in Proc. of the Natural Language Pacific Rim Symposium (NLPRS'2000), 369-375*
2. Mark Craven, Johan Kumlien, Constructing Biological Knowledge Bases by Extracting Information from Text Sources, *in Proc. of the 7th International Conference on Intelligent Systems for Molecula Biology(ISMB-99)*
3. PDB(*Protein Data Bank*), http://www.rcsb.org/pdb/
4. NCBI (*National Center of BioTechnology Information*), http://www.ncbi.nih.gov/
5. EMBL (*European Molecular Biology Laboratory*), http://www.embl-heidelberg.de/
6. DDBJ(*DNA Data Bank of Japan*), http://www.ddbj.nig.ac.jp/