제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

온톨로지를 이용한 서열정보분석 데이터베이스 구축 시스템 설계

이선아*, 전중남*, 이건명*
*충북대학교 컴퓨터과학과, 첨단정보기술 연구센터
e-mail : bluebird@aicore.chungbuk.ac.kr,
joongnam@cbucc.chungbuk.ac.kr, kmlee@aicore.chungbuk.ac.kr

System Design for Building Sequence Information Analysis Databases using Ontology

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

요 약

인터넷과 첨단 기술의 발달로 생물학적 정보에 대한 온라인 데이터베이스들이 급증하고 있으나, 데이터가 방대하고 형식이 다양하여 생물학자들이 정보를 얻는데 많은 어려움이 있다. 본 논문에서는 단백질과 핵산 정보를 제공하는 NCBI에서, 사용자가 질의에 따른 웹 문서로부터 정보를 추출하여 사용자의 특성에 따른 데이터베이스를 구축, 관리하여 주는 시스템을 제안한다. 온톨로지를 이용하여 질의의 모호성을 보완한다. 웹 문서로부터 추출된 데이터를 저장하는 단계에서도 데이터의 특징, 사용빈도에 따라 테이블을 분류하여 관리함으로써 검색과 관리의 효율성을 높인다. 본 논문은 서열정보 분석을 하는 생물학 연구자들에게 데이터베이스를 쉽게 구축하고 서열정보를 분석하기 좋은 인터페이스를 제공하는 것을 목적으로 한다.

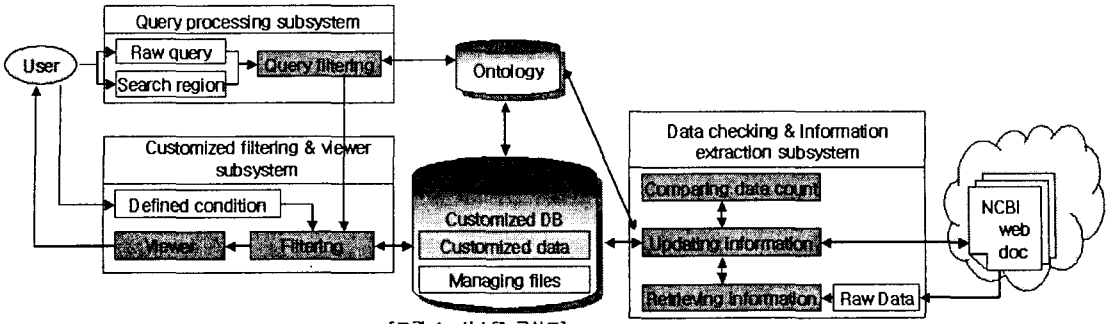
1. 서론[1)]

오늘날 인터넷의 확산과 첨단 기술의 발달로 서열구조, 분자 상호작용, 표현 패턴과 같은 생물학적 정보에 대한 온라인 데이터베이스들이 급증하고 있으나, 생물학자들이 이런 정보를 다루기는 쉽지 않다[1,2]. 현재 생물학 정보를 제공하는 데이터베이스에는 PDB, GenBank, NCBI[3], EMBL, SWISS-PROT, DDBJ,등이 대표적이다. 이들은 데이터가 방대하고 서로 제공하는 데이터의 형식이 다르기 때문에 적절한 데이터를 얻기 위해서는 많은 시간과 노력이 필요하다. 데이터베이스들은 생물학자들과 다른 생물학 관련 연구자들을 위해서 이러한 데이터베이스를 공개하여 제공하고 있으나, 생물학 연구자들이 제공되는 방대한 데이터베이스를 일반 컴퓨터에 구축하여 연구하기란 많은 어려움이 따른다. 지속적으로 증가하는 생물학 데이터들에 대해서도 고려하여 업데이트해야하는 번거로움 또한, 중요한 문제이다. 웹 데이터베이스에서 제공되는 정보는 질의를 통하여 제공되기도 한다. 질의로 제공되는 경우, 방대한 양과 질의의 모호성으로 인하여 추출된 데이터의 정확성이 떨어지는 문제를 들 수 있다.

이러한 문제들을 해결하기 위해서 본 논문에서는 온톨로지를 이용한 서열정보 데이터베이스 구축 시스템을 제시한다. 제시된 시스템은 단백질과 핵산 정보를 제공하는 대표적인 데이터베이스인 NCBI에서, 질의를 통해 해당 정보를 추출하고 추출된 정보를 바탕으로 서열정보 데이터베이스를 구축한다. 온톨로지를 이용하여 질의처리와 데이터베이스 구축, 데이터베이스 관리를 한다. 방대한 데이터 중에서 질의 형식의 사용자 요구를 받아들이고 이를 만족하는 데이터만을 추출하여 데이터베이스를 구축한다. 온톨로지를 통해 질의의 모호성을 보완한다. 또한 이미 저장한 질의항목에 대한 업데이트 기능을 제공하여 지속적으로 증가하는 데이터에 대한 방안을 모색한다.

정보를 추출하는 대상이 되는 NCBI는 미국의 국립보건원(NIH) 산하 기관으로 분자생물학 정보의 가공

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



[그림 1. 시스템 구성도]

및 처리를 목적으로 하며, '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[3]. 통합적인 데이터를 제공하기 때문에 연구 대상으로 많이 이용되며 특히 MEDLINE의 데이터에 대해서 여러 접근법들이 시도되고 있다. 하지만 핵산과 단백질 정보에 대해서는 연구가 미비한 편이다. 이는 해당 데이터를 데이터베이스로 제공하기 때문이다. 제안된 시스템은 효율과 사용자의 의도에 맞도록 데이터베이스의 구조를 지정함으로써 좀 더 사용자 측면을 고려한 데이터베이스 구축 시스템을 설계한다.

본 논문은 2장은 제안된 시스템의 구성과 과정을 간략히 소개한다. 3장에서 온톨로지를 통한 질의처리, 4장에서 구축된 데이터베이스를 어떻게 관리하는지를 보인다. 마지막으로 5장은 향후 과제 및 결론을 논한다.

2. 시스템의 구조

제안된 시스템은 크게 다섯 부분으로 구분된다. 질의 처리 서브시스템, 맞춤형 정제 및 시각화 서브시스템, 데이터 확인 및 데이터 추출 서브시스템, 온톨로지, 맞춤형 데이터베이스이다[그림 1]. 사용자에 의해 입력된 질의는 온톨로지를 통해 모호성이 보완되며, 정제된 질의는 맞춤형 데이터베이스에서 해당 내용을 검색한다. 사용자는 업데이트 시간을 지정하여 추출된 정보 인덱스를 가지고 새로이 등록된 데이터를 가져와 업데이트한다.

2.1 시스템의 전제조건

시스템의 처리과정에서 전제조건이 되어야 하는 것이 있다. 시스템의 전제조건을 정의함으로써 무한정 확장될 수 있는 검색내용을 축소시킬 수 있다. 다음은 시스템의 전제조건을 보인 것이다.

① 입력은 검색영역과 검색어 두 가지이다.
② 검색어는 종(種)과 서열정보 두 가지 종류이다.
③ 검색어는 영어만 가능하다.
④ 사용자는 질의된 내용을 보여주는 결과창에 대해 시각화에 필요한 항목들을 정의해야 한다.
⑤ 사용자는 맞춤형 데이터베이스의 구조를 설정할 수 있다.
⑥ 사용자는 업데이트를 할 경우, 질의할 때마다 할 것인지 모든 작업을 마친 후 할 것인지를 결정해야 한다.

2.2 시스템의 구성

2.1.1 질의 처리 서브시스템

사용자가 입력한 질의를 정제하여 맞춤형 정제 및 시각화 서브 시스템에 전달한다. 질의 정제할 때에 온톨로지를 사용한다. 검색 영역을 설정하는 이유는 질의를 좀더 정확히 추론하기 위함이다.

2.2.2 맞춤형 정제 및 시각화 서브시스템

사용자가 지정한 조건에 맞추어 질의에 해당하는 데이터를 맞춤형 데이터베이스에서 가져와 사용자에게 보여준다. 보여주는 형태는 사용자가 보이도록 지정한 항목에 대해서 보여준다.

2.2.3 데이터 확인 및 데이터 추출 서브시스템

제안된 서브시스템은 업데이트와 관련된 부분이다. 질의에 대해 맞춤형 데이터베이스에 있는 데이터와 실제 NCBI에서 추출되는 정보의 양에 차이가 있을 경우, 빠진 부분을 식별하여 해당 웹 문서를 추출한다. 추출된 정보는 맞춤형 데이터베이스에 저장되게 된다.

2.2.4 온톨로지

모든 서브시스템에 자원을 제공하는 데이터 사전이다. 질의처리 서브시스템에서 질의 정제를 하는데 자료를 제공하고 데이터베이스에서 관리하는 데이터들 중 업데이트를 위한 관리 파일들(managing files)의 파일명을 지정한다. 또한 계통도의 정보를 가짐으로써 사용자가 계통에 대해 모호함을 가지고 있을 경우, 시각화하여 보여줄 수 있도록 정보를 제공한다.

2.2.5 맞춤형 데이터베이스

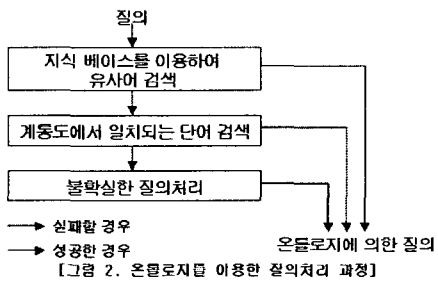
사용자가 정의한 테이블 구조를 지니며 웹 문서에

서 추출한 정보를 저장한다. 또한 사용자의 질의에 따른 정보를 맞춤형 정제 및 시각화 서브시스템에 제공하여 사용자가 볼 수 있도록 한다.

3. 온톨로지를 통한 질의처리

온톨로지는 기본 용어, 주제에 따른 어휘비교로 얻어지는 관계뿐만 아니라 용어와 관계를 비교하여 얻어지는 규칙을 정의한다[4]. 온톨로지는 어떤 서로 다른 대상에 대해 통합이라는 목적을 가지고 접근하기에 적절한 방법중 하나이다. 기본적으로 서로 다른 도메인에 속하는 두 요소에 대해 연관성을 추론하여 같은 의미를 지니는지를 판별할 때에 온톨로지를 사용한다.

제안한 시스템에서 사용하는 온톨로지는 질의 내용에 대한 통일성을 제공한다. 질의의 형태는 사용자의 특성 혹은 찾고자 하는 대상에 따라 한 단어 혹은 여러 단어들이 결합된 문장의 형식을 지닌다. 예를 들어, 'rat'라는 입력을 하면 'rat'이라는 단어가 존재하는 문서를 추출한다. 'rat'이라는 단어에 대해서 'rat'이라는 단어 외에도 'rats'혹은 'Buffalo rat' 등의 입력으로도 가능하다. 다양한 입력이 가능한 질의에 대해 통일된 질의를 사용함으로써 데이터베이스 관리의 효율성을 높일 수 있도록 한다. 제안한 시스템은 질의처리를 하기 위해 NCBI에서 제공하는 계통(taxonomy) 항목을 참조하여 온톨로지를 제공한다. 온톨로지를 이용한 질의 처리의 과정은 다음의 [그림 2]와 같다.



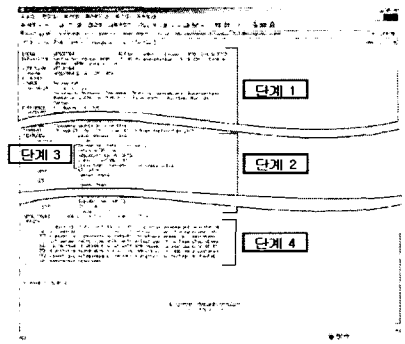
[그림 2. 온톨로지를 이용한 질의처리 과정]

먼저 온톨로지를 이용하여 질의와 유사한 단어가 있는지를 검사한다. 만일, 유사한 단어가 존재한다면 서로 매핑하여 원시 검색어를 온톨로지를 이용해 일관성을 지닌 유사한 단어로 변환한다. 만약 유사한 단어가 없을 경우에는 계통도에서 검색하여 본다. 각 단계별 계통도를 참고로 하여 유사어를 검색한다. 계통도 내에 유사어가 존재한다면 매핑한다. 만약 두 번째 단계에도 유사어가 존재하지 않는다면 불확실한 질의처리 단계를 통해 일관성을 지닌 질의어를 추론한다.

불확실한 질의로 판별이 되어 세 번째 단계에 온 경우, 먼저 한 단어인지 여러 단어가 결합된 문장인지를 확인한다. 만약 여러 단어들이 연결되어 있는 경우라면 매 마지막에 위치하는 단어를 대표단어로 지정하여 유사성을 파악한다. 한 단어로 이루어진 경우에는 단어가 지니는 확률값을 부여한다. 확률값은 길이에 의한 일치여부를 확인함으로써 가능한다. K를 (단어길이)/2 라고 할 경우, K에 대해 일치여부를 확인하고 K의 위치에서 한 단어씩 덧붙여 읽으면서 유사여부를 확인한다. 유사여부를 확인하면서 전체 계산 결과에 대해 유사한 경우가 어느 정도 발생하는지를 확률값으로 측정하여 가장 높은 값을 지니는 것으로 매핑하여 질의를 만든다. 이는 질의의 불확실한 부분을 1/2 이상의 위치부터라는 가정 하에 문제를 해결하려 한 것이다. 이때, K에 대해 한 문자씩 증가하여 확률을 결정하는 과정에서 모음(a, e, i, o, u)이 나올 경우에는 확장하여 모음에 해당하는 것을 몇 가지 대입하여 유사도를 검색한다. 사용자가 단어에 대해 확실하게 모를 경우, 자음은 발음의 주요한 역할을 담당하므로 오류가 적지만 모음의 경우에는 유사모음으로 잘못 입력할 수 있기 때문이다. 때문에 여러 경우를 고려하여 질의를 추론한다. 이러한 방법을 통해 가장 높은 확률값을 가지는 단어를 입력 질의와 매핑하여 온톨로지에 의한 질의로서 반환한다.

4. 데이터베이스의 구축과 관리

이 논문에서 제안한 시스템은 온톨로지를 이용하여 서열정보 데이터베이스를 구축한다. 웹 문서로부터 정보를 추출하여 사용자가 정의한 맞춤형 데이터베이스에 저장한다. 정보가 포함되어 있는 웹 문서는 다음과 같은 형태를 지닌다[그림 3].



[그림 3. 정보를 추출하는 웹 문서의 형태]

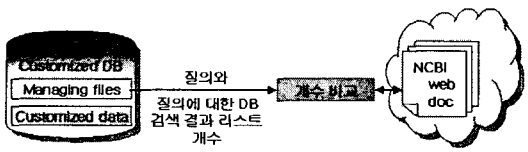
웹 문서에서 추출된 데이터들은 크게 6부류로 나뉘어 저장된다[표 1].

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

[표 1. 데이터 특성에 따른 저장 테이블의 구성]

기본 테이블은 [그림 3]에서 보여지는 부분에서 단계 1에 해당하는 부분을 저장한다. 단계 1에 해당하는 데이터들은 단백질 혹은 핵산 정보의 기본적인 항목을 담고 있다. 정보의 식별이름(Access number), 서열길이, 분자의 종류, 정보가 등록 혹은 수정된 날짜, 어떤 종에 속하는 생물의 유전정보(Organism)인가 등의 정보가 있다. 식별 유전자 테이블에 저장되는 것은 단계 5이다. [그림 3]에서 보여지는 것은 단계 1부터 단계 4까지이다. 단계 5에 해당하는 웹 문서는 [그림 3]과는 형태가 전혀 다른 형태이기 때문에 따로 분류하였다. 유전자 특성 테이블은 단계 2의 항목 중에서 가장 많이 반복되는 gene항목과 CDS항목을 제외한 나머지 항목에 대한 정보를 저장한다. gene과 CDS 항목은 반복이 많고 데이터 검색 시에 검색 대상이 되기 때문에 따로 구분하여 Gene_CDS 테이블에 저장한다. 전이 테이블은 CDS의 하위 항목 중, 염기들이 전이된 서열항목(translation)에 해당하는 내용만을 따로 저장한 것으로, 이 테이블 또한 검색 대상이 되기 때문에 따로 분류하여 관리한다. 마지막으로 ORIGIN 테이블은 [그림 3]에서 가장 마지막에 존재하는 것으로 염기서열 부분을 가리키고 있다. 염기서열 또한 검색의 대상이 되기 때문에 검색효율을 높이기 위해 분리하여 관리한다.

저장된 데이터를 바탕으로 사용자가 질의하는 내용에 따라 해당 데이터를 제공한다. 본 논문에서 제안한 시스템은 질의에 따른 업데이트 기능을 지원한다. 사용자가 질의한 내용을 리스트로 저장하여 업데이트를 한다. 업데이트 여부는 맞춤형 데이터베이스에 관리용 파일(managing files)를 이용해 판별한다. 관리용 파일에는 현재 사용자가 질의했던 내용들을 목록화하여 관리하고 해당 질의 결과에 대해서도 리스트로 관리한다. 만일 업데이트가 요구될 경우, 시스템은 관리용 파일에서 사용자가 질의한 내용을 순차적으로 선택하여 해당 질의에 따른 결과 리스트의 개수와 NCBI에 같은 질의를 하였을 때에 나오는 리스트의 개수를 비교하여 업데이트 여부를 판별한다. 비교하는 과정은 [그림 4]와 같다.



[그림 4. 개수 비교를 하는 과정]

5. 향후 과제 및 결론

본 논문에서는 단백질과 핵산 정보를 제공하는 대표적인 데이터베이스인 NCBI에서 질의에 의해 얻어지는 정보를 추출하여 사용자의 의도에 따른 데이터베이스를 구축하여 주는 시스템을 제안하였다. 온톨로지를 이용하여 질의를 처리하고 데이터베이스를 구축, 관리하는 방안을 제시하였다. 온톨로지를 통해 NCBI에서 제공하는 단백질과 핵산 정보에 대한 질의를 정제하고 그에 따른 결과물에 대한 관리를 통해 검색의 효율을 증가시키고 업데이트가 가능하도록 하였다. 이를 통해 구축된 데이터베이스를 생물학 연구자들이 좀더 쉽게 분석하는 것을 목적으로 하였다.

현재 이루어지고 있는 생물학 데이터베이스들의 통합문제에 있어서 NCBI라는 데이터베이스에 대해 정보추출이 가능하고 그에 대한 데이터베이스 구축이 가능함을 보였다. NCBI와 같이 생물학 데이터베이스를 제공하는 PDB와 같은 또 다른 데이터베이스에 대한 분석을 통해 데이터베이스들을 통합하는 문제에 좀더 접근하는 기회를 제공한 것이다.

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] NCBI(*National Center of BioTechnology Information*), http://www.ncbi.nih.gov/

[4] Asuncion Gomez Perez, V. Richard Benjamins, Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods, Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends . (IJCAI99). 2 de Agosto. Estocolmo. 1999. Pags.