1)에이전트를 이용한 사용자 중심의 개인용 생물학 검색시스템

김영억° 정광수 류근호
충북대학교 데이터베이스연구실
{kimuks°, ksjung, khryu}@dblab.cbnu.ac.kr

User-Centric Personal Biological Retrieval System Using Agents

Young Uk Kim°, Kwang Su Jung, Keun Ho Ryu
Database Laboratory, Chungbuk National University

요 약

생명정보 분야의 발전과 더불어 과거 축적되어 온 방대한 양의 생물학 데이터들이 이질적인 형태로 데이터베이스화 되어있다. 특히, 인간게놈프로젝트의 완료 후에 유전자 및 단백질의 기능을 밝히기 위한 지노믹스 및 프로테오믹스 연구가 활발해졌다. 새로운 생물학적 과정을 탐색하기 위해서는 기존에 존재하는 생물학 데이터베이스의 데이터를 수집하기 위한 기술적인 검색 능력이 필요하다[1]. 전산지식이 부족한 대부분의 생물학자들은 공개용 데이터베이스로부터 필요한 정보를 획득하는데 어려움을 격고 있다. 각 분야의 생물학자들이 공개용 데이터베이스로부터 자신의 분야에 관련된 데이터를 검색·추출하는 작업을 수월하게 해 줄 검색 시스템이 필요하다.

따라서, 에이전트를 이용하여 공개용 데이터베이스로부터 정보를 수집하는 사용자 중심의 개인용 검색 시스템을 제안하고자 한다. 또한, 검색시스템을 이용하여 생물학자가 지노믹스와 프로테오믹스의 실험적인 접근을 위해 원하는 많은 양의 특정 도메인의 데이터를 검색하고 질의된 결과를 개인 컴퓨터에 2차 데이터베이스를 만들어 저장한다. 사용자에 의해 생성된 특정 분야의 도메인인 2차 데이터베이스를 통해 데이터의 접근의 편리성과 생물학 정보의 분석의 용이성을 얻을 수 있다.

1. 서 론

생명정보학에서 관심을 갖는 데이터 집합은 데이터를 작성한 사람이나 기관, 연구 목적 등의 차이로 인해 구조적으로나 내용적, 그리고 의미적으로 볼 때 이질적인 면들이 많이 존재한다[2]. 생물학 데이터는 데이터 종류가 다양하고 이질적인 포맷을 가지며 데이터가 분산되어 있고, 데이터 자체가 생물학적 변이를 포함하고 이러한 데이터가 실험을 통해 생산되기 때문에 데이터의 상태가 항상 변화하고 유동적인 특징이 있다.

그러나, 각 분야의 생물학자는 자신의 연구 분야에 관련된 데이터만을 필요로 한다. 축적된 공개용 생물 데이터베이스로부터 연구에 필요한 특정 생물학 데이터를 검색·추출·분석하는 작업은 실험생물학자들에게 어려운 과제로 대두되고 있다[3].

사용자가 처리해야 할 생물학 데이터의 양이 많아짐에 따라 이러한 사용자의 수고를 효과적으로 보조해 주는 에이전트의 필요성이 증대되고 있다[4].

따라서, 이 논문에서는 사용자에게 필요한 특정 도메인의 생물학 데이터를 각각의 공개용 데이터베이스로부터 에이전트를 이용하여 가져오는 검색시스템을 개발하였다. 또한, 추출한 생물학적 정보는 개인용 컴퓨터에 2차 데이터베이스로 구축하여 새로운 생물 정보 생성을 위해 사용된다. 2차 데이터베이스에 저장된 데이터는 정보 공유를 위하여 BSML 기반의 생물학 데이터 변환기[5]를 통해 사용자가 원하는 형태로 변환할 수 있다.

이 논문은 2005년도 교육인적자원부 지방연구중심대학 육성사업의 지원에 의하여 연구되었음

2. 관련연구

2.1 생물학 검색시스템

SUISEKI[1]는 정보 추출 시스템으로써, PubMed의 문헌을 검색하여 gene과 protein, interaction을 찾는다. 그러나 문헌의 초록을 단순한 텍스트 마이닝 기법을 통해 gene과 protein의 이름을 사용하므로, false positive가 발생한다. BioRAT[6]는 공개용 DB인 PubMed의 초록과 full-length 논문을 검색하는 시스템으로써, 문서의 패턴을 찾아 사용자가 원하는 생물 정보를 보여주는 기능이 있다. 그러나 검색한 결과의 저장과 관리를 위한 DB가 부재하다. SRS[7]는 생물학 정보의 데이터 분석을 위한 데이터 검색 및 어플리케이션 기능을 포함하는 NCBI Entrez와 유사한 생물학 정보 통합검색관리 서버 시스템이다. 다양한 검색 및 분석 기능을 제공하지만, 시스템 설치와 관리가 복잡하고 고사양의 하드웨어가 필요하다.

기존의 시스템은 어느 정도의 편리성을 가지고 있으나, 검색된 결과의 저장·관리 및 사용의 복잡성을 가지고 있다. 이 논문에서 제안한 시스템은 다양한 생물학 데이터베이스로부터 검색된 결과를 관리하고 특정 생물학적 문제에 적합한 데이터로 재구성하는 문제를 해결하고자 하였다. 전산 지식이 부재한 생물학자들이 손쉽게 공개용 생물학 데이터베이스에 접근하며, 요구사항들을 신속

하게 처리해 줄 수 있도록 하였다. 따라서, 기존에 구축된 공개용 생물학 데이터베이스를 분석 가공하여 새로운 정보를 추출하여 유용한 생물학적 정보를 저장할 수 있는 2차 데이터베이스를 구축하는데 응용이 될 수 있다.

2.2 공개용 생물학 데이터베이스 : NCBI

NCBI는 DNA 서열 데이터베이스인 Nucleotide, 단백질 서열 데이터베이스인 Protein, 문헌정보 데이터베이스 PubMed, 특정 개체의 진화적 관련을 분석하기 위한 DNA 서열집합 데이터베이스인 Popset 등의 이질적인 데이터베이스를 유지하고 있다. 이러한 데이터베이스로부터 서열 및 유전자 데이터를 검색하기 위한 검색시스템으로 Entrez를 이용하고 문헌 검색시스템으로 PudMed를 제공한다. 많은 기능을 제공하기 때문에, 이용의 복잡성이 존재한다. 그러므로, 이 논문에서 제안한 시스템은 사용자 편의 단일 인터페이스를 제공하여 사용자들이 손쉽게 접근할 수 있도록 하였다.

2.3 검색 에이전트

에이전트란 인간처럼 자치적이고 지능적으로 동작하는 시스템이다[8]. 특히, 검색 에이전트는 사용자를 대신해서 사용자가 원하는 검색 작업을 자동적으로 해결하여 주는 소프트웨어라고 할 수 있다. 주어진 질의를 통해 처리된 정보를 축적하고 이를 바탕으로 영역 지식의 확장을 도모하는 기능을 하게 된다. 공개용 생물학 데이터베이스는 사용자에게 다양하고 방대한 양의 생물학 정보를 제공하기 때문에 본 시스템에서는 특정 영역에 적합한 자료만을 대상으로 저장·관리할 수 있는 데이터베이스를 구축하여 사용하고 있다.

3. 시스템 구조

에이전트를 이용한 사용자 중심의 개인용 생물학 검색 시스템의 전체적인 구성은 그림 1과 같다. 먼저 사용자 뷰어를 통해 생물학 검색어를 입력하면, DB검색 모듈을 통해 공개용 데이터베이스로부터 검색어와 관련된 다양한 유전체 및 단백질 정보를 XML이나 HTML과 같은 웹 문서로 결과를 가져온다. 문서파싱 모듈은 웹 문서로부터 필요한 정보를 파싱하여 필드 및 데이터를 추출하여 사용자에게 보여 준다. 사용자 뷰어를 통해 사용자는 원하는 자료만을 저장관리 모듈을 통해 개인용 생물학 데이터베이스에 저장하게 된다. 그리고 효율적으로 생물학 데이터들 간의 정보를 공유하기 위해서 사용자가 포맷 변환 모듈을 통해 원하는 포맷으로 결과를 얻을 수 있다.

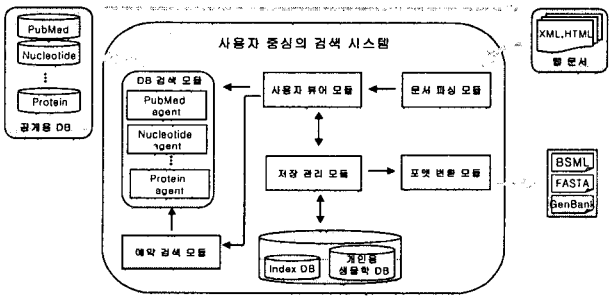

그림 1. 시스템 구조

3.1 사용자뷰어 모듈

사용자 뷰어 모듈은 사용자 중심의 검색 시스템의 메인 부분으로 사용자의 질의를 생성하고, 질의 결과로 얻어진 다양한 유전체 및 단백질 정보를 저장·관리하기 위해 이질적인 포맷간의 맵핑과정을 통하여 BSML 형태로 변환하여 보여준다. 또한, 사용자는 뷰어모듈을 통해 선택한 정보를 저장 관리 모듈을 통해 개인용 생물학 DB에 저장하게 된다.

3.2 검색 모듈

검색 모듈은 DB검색 모듈과 예약검색 모듈로 이루어졌다. DB검색 모듈은 각각의 공개용 생물학 데이터베이스를 검색하는 에이전트들로 구성되어 있다. 그림 2와 같이 에이전트는 생물학 데이터베이스를 접근하여, 사용자가 어떠한 검색이나 분석을 하기 위하여 요청을 할 때에 정보를 검색하고 추론하여 반환하는 역할을 한다.

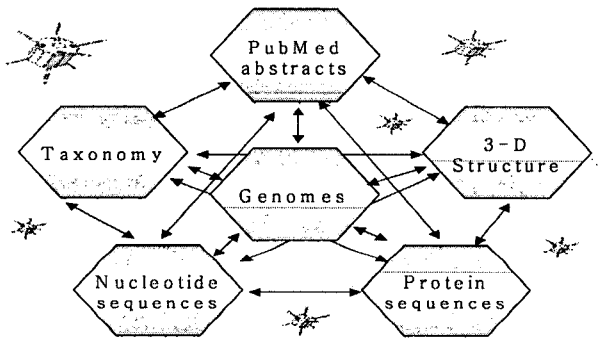

그림 2. 에이전트를 사용한 생물학 DB 검색

예약검색 모듈은 사용자가 검색한 정보를 토대로 검색한 데이터를 자동으로 업데이트할 수 있다. Index 데이터베이스에 저장되어 있는 질의어, 결과 ID 정보를 바탕으로 공개용 DB의 데이터 소스를 주기적·사용자의 요청에 따라 데이터를 갱신하게 된다. 그림 3은 기존 검색 결과를 근거로 한 자동 업데이트 검색 과정이다. 문서 수집기(Crawler)를 통해 웹문서 형태의 질의 결과를 얻어 온다. 색인기(Indexer)는 웹문서로부터 유니크한 ID 목록을 추출한다. 추출한 ID 목록과 Index 데이터베이스에 저장된 ID 목록을 비교하여 새로운 ID들은 검색기(Searcher)에 의해서 공개용 생물학 데이터베이스로부터 ID와 관련된 생물학 정보를 가져온다.

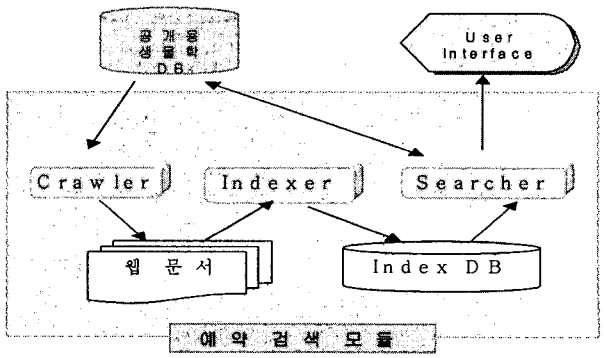

그림 3. 자동 업데이트 검색 과정

3.3 문서파싱 모듈

문서파싱 모듈은 다양한 공개용 생물정보 데이터베이스(Nucleotide, Protein, PubMed 등)로부터 수집된 XML이나 HTML 형식의 웹 문서에서 필요한 정보를 파싱하는 역할을 한다. 다시 말해, XML이나 HTML 형식의 생물정보 플랫 파일로부터 필드 및 데이터 추출을 위해 파서기능을 담당하게 된다.

3.4 저장관리 모듈

저장관리 모듈은 문서파싱의 결과로 얻어진 필드 및 데이터와 질의 내용 및 예약 정보를 통합 관리한다. 문서파싱 모듈로부터 얻어진 필드와 데이터는 개인용 생물학 데이터베이스에 저장을 한다. Index 데이터베이스는 검색모듈이 수집한 생물학 정보의 질의어, 검색 날짜, 검색한 결과들의 각각의 ID, 관련 데이터베이스 및 예약 정보를 담고 있는 데이터베이스이다.

3.5 포맷변환 모듈

포맷변환 모듈은 개인용 생물학 데이터베이스에 저장된 생물학 정보를 생물학자가 원하는 BSML, FASTA, GenBank 등과 같은 문서포맷으로 생성하기 위한 기능을 한다. 그림 4와 같이 다른 포맷간의 매핑관계에 따라서 사용자가 원하는 형태로 생성하는 역할을 한다.

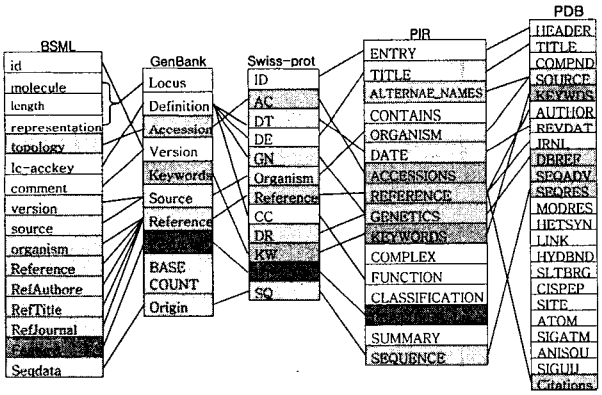

그림 4. 이질적인 포맷간의 매핑정보

4. 구현

구현된 시스템의 구현환경은 Pentium PC 850MHz 시스템에서 MySQL 데이터베이스를 이용하였다. 그리고 플랫폼 독립적으로 시스템을 실현하기 위해 JAVA 구현하였다.

그림 5는 PubMed로부터 검색어가 SOD1이고 PMID가 15843790를 XML 형식의 결과로, Nucleotide로부터 검색어가 SOD1고 GI가 52208053를 Genbank 양식으로 추출하였다.

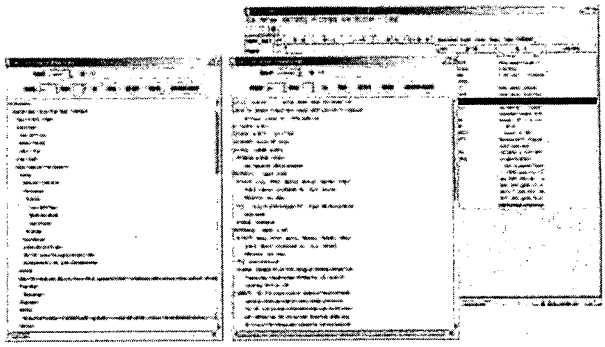

그림 5. 시스템 전체화면 및 에이전트를 이용한 검색화면

5. 결론

이 논문에서는 에이전트를 이용한 사용자 중심의 개인용 생물학 검색 시스템을 제안하였다. 이 시스템은 지능적인 서비스를 위해 에이전트를 적용한 시스템이다. 에이전트를 통해 생물학 DB관리에 보다 효율적이고 지능적인 처리 및 서비스를 받을 수 있다.

향후 연구로는 생물학 정보에 대한 도메인 정의 및 지식의 정형화를 위해 바이오 온톨로지를 추가하여 공개용 생물학 데이터베이스와의 상호운영성과 지능적인 갱신 서비스 및 복잡한 분석이나 검색의 정확성을 높이는 연구도 필요하다.

참 고 문 헌

[1] C. Blaschke, "The Frame-Based Module of the SUISEKI Information Extraction System", IEEE INTELLIGENT SYSTEMS, vol.17, issue.2, pp.14-20, 2002.

[2] A. Silvescu, "Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed Biological Data Sources", Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery.

[3] Andreas D. Baxevanis, "Bioinformatics", Wiley, 2001.

[4] M.N. Huhns, "Personal Assistants", IEEE Internet Computing, vol.2, issue.5, pp.90-92, 1998.

[5] 김영억, 정광수, 정영진, 차효성, 류근호, "정보공유를 위한 BSML 기반의 생물학 데이터 변환기", 한국정보과학회, vol.31, no.2, pp 37~39, 2004.

[6] David P. A. Corney, "BioRAT: extracting biological information fromfull-length papers", bioinformatics, vol.20, no.17, pp.3206-3213, 2004

[7] E. M. Zdobnov, "The EBI SRS server-new features", bioinformatics, vol.18, no.8, pp.1149-1150, 2002

[8] M. Wooldridge, "Intelligent Agents : Theory and Practice", Knowledge Engineering Review, 1994