질병 검색 서비스를 위한 디렉토리 시스템 설계 및 구현

Design and Implementation of a Directory System for Disease Retrieval Services

여명호, 이윤경, 노규종, 박형순, 김학신, 박준호,
강태호, 김학용, 유재수
충북대학교

Myung-ho Yeo, Yoon-kyeong Lee, Kyu-jong Rho,
Hyoung-soon Park, Hak-sin Kim, Jun-ho Park,
Tae-ho Kang, Hak-yong Kim, Jae-soo Yoo,
Chungbuk National University

요약

생명 공학 분야의 연구는 대용량의 데이터 처리를 요구한다. 과거 실험을 통해 접근하던 방식에서 벗어나 최근 IT 기술의 결합을 통해 다양한 실험 데이터를 공유하고, 연계함으로써 연구를 가속화하고 있다. 질병에 대한 연구는 생명 공학의 큰 테마 중 하나이다. 질병 데이터를 분류하고, 웹을 통해 데이터를 제공하는 다양한 서비스가 존재한다. 하지만, 기존 서비스들은 각기 다른 분류 방법을 가지고 있으며, 고차원 처리를 요구하는 신규 서비스와 연계하기 위한 인프라의 부재는 생명 공학 연구의 발전을 저해하는 요소로 작용하기도 한다. 본 논문에서는 이종의 질병 데이터베이스를 통합하기 위한 데이터 구조를 제안하고, 신규 서비스와 연계하기 위한 인프라로서 질병 디렉토리 시스템을 설계하고 구현한다.

Abstract

Recently, biological researches are required to deal with a large scale of data. While scientists used classical experimental approaches for researches in the past, it is possible to get more sophisticated observations easily with convergence of information technologies and biology. The study on diseases is one of the most important issues of the life science. Conventional services and databases provide users with information such as classification of diseases, symptoms, and medical treatments through web. However, it is hard to connect or develop them for other new services because they have independent and different criterions. It may be a factor that interferes the development of biology. In this paper, we propose an integrated data structure for the disease database, and design and implement a novel directory system for diseases as an infrastructure for developing other new services.

I. 서론

1990년대 처음 등장한 바이오인포메틱스는 현재 생명 과학의 연구에 있어서 중요한 분야로 인식되고 있다. 연구의 시작은 1950년대 단백질 서열에 대한 연구이다. 당시 많은 과학자들이 단백질의 아미노산 서열을 분석하는 실험을 수행하면서 아미노산 서열들에 대한 많은 정보를 축적하고, 그들은 축적된 정보들을 통합하고 정리하여 데이터베이스를 구축해 연구에 사용함으로써 최초의 바이오인포메틱스 데이터베이스가 만들어졌다. 초기 단백질 아미노산 서열 데이터를 축적해 데이터베이스화하고, 그 서열들을 분석하는 도구를 개발하게 되면서 바이오인포메틱스라는 새로운 분야로 자리매김하였고, 지금은 아미노산이나 유전자 서열뿐만 아니라 다양한 종류의 생명 공학 분야의 대용량 데이터에 IT 기술을 결합해 연구하는 모든 분야를 일컬어 바이오인포메틱스라고 한다[1][2].

질병에 대한 연구는 많은 과학자들의 중요 연구테마이자, 일반인을 포함한 모든 사람들의 큰 관심사이기도

이 논문은 교육과학기술부와 한국산업기술재단의 지역혁신인력양성사업과 2009년 교육과학기술부로부터 지원받아 수행된 연구임(지역거점연구단육성사업/충북BIT연구중심대학육성사업단).

하다. 현재 질병 데이터를 분류하고, 웹을 통해 데이터를 제공하는 다양한 서비스가 존재한다. 하지만, 서로 다른 목적을 위해서 생성된 데이터베이스이기 때문에 각기 다른 분류 방법을 가지고 있으며, 동일한 질병이라 할지라도 다른 이름을 사용하는 경우가 빈번하다. 또한, 고차원 처리를 요구하는 신규 서비스와 연계할 수 있는 IT 인프라의 부재는 생명 공학 연구의 발전을 저해하는 요소로 작용하기도 한다. 따라서 질병에 대한 효율적인 연구를 위해서 개별적으로 이루어지는 연구 결과를 통합하고 체계적으로 정리해 공유하는 시스템이 필요하다. 이러한 데이터 통합 및 공유 시스템은 중복 연구를 방지함으로써 연구자들의 시간적, 경제적 손실을 막아 질병의 연구에 있어서 큰 성과를 얻는데 도움이 된다.

본 논문에서는 이종의 질병 데이터베이스를 통합하기 위한 데이터 구조를 제안한다. 또한 신규 서비스와 연계하기 위한 인터페이스를 제공하는 질병 디렉토리 시스템을 설계하고 구현한다.

본 논문의 구성은 다음과 같다. II장에서는 관련연구를 기술한다. III장에서는 제안하는 시스템 구조와 제안하는 시스템의 주요 기능을 기술한다. IV장에서는 제안하는 시스템의 구현 환경과 예제 페이지를 구현한 내용을 기술한다. 마지막 V장에서는 논문의 결론에 대해 기술한다.

Ⅱ. 관련연구

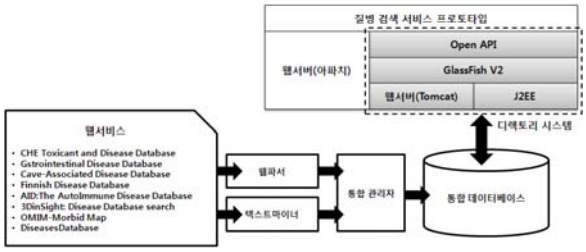
대표적인 질병관련 데이터베이스로는 CHE[3], Gastro net[4], Findis[5], AID[6], 3DinSight[7], OMIM-Morbid Map[8], DiseaseDatabase[9]가 있다. CHE는 약물과 질병데이터베이스로 화학적 약물에 대한 정보와 약 180여 가지의 인간 질병에 대한 간단한 정보를 제공한다. 각 질병들을 유발시키는 물질들의 리스트에 대한 정보를 얻을 수 있고, 그 물질들이 질병을 유발하는 정도는 3단계(Strong, Good, Limited Evidence)로 나누어 표현하고 있다. Strong evidence는 의학단체에서 인정하고 있고, 교과서에도 실릴 정도로 증거가 충분한 경우에 표시한다. Good evidence는 사람들에게 나타난 약간의 증거를 갖고 있거나, 동물실험을 통해 강하게 증명된 경우 표시한다. limited/conflicting evidence는 사람들에게서 약하게 나타나거나, 동물실험에서 약하게 나타나는 경우 표시한다. Gastro net은 환자와 의료 전문가들에 의해 온라인으로 제공되는 의료 정보 사이트이다. 주로 위장과 소화기 계통의 질병에 대한 정보를 제공한다. 질병의 질병에 대한 간략한 설명과, 증상, 치료법, 관련 정보들에 대해 알 수 있다. Findis는 과학자들과 의사들이 논문의 내용을 직접 검토해 신뢰성 있는 데이터만을 수집한 데이터베이스이다. 각 질환의 전이, 발생률 및 임상의 증상 등에 대한 간략한 설명과 유전자 변이에 대한 정보 를 제공한다. AID는 MEDLINE에 수록된 500000개 이상의 관련 논문들을 분석하여 자가 면역과 관련한 정보를 얻어 구축한 데이터베이스로서 자가면역 질병과 관련된 유전자 정보를 제공하고, Entrez Gene, Ensemble, SwissProt와 링크되어 있어 관련 유전자에 대한 추가 정보를 확인할 수 있고 PMID를 통해 관련 정보에 대한 논문을 검색할 수 있다. 3DinSight에서는 키워드나 단백질 이름, 유전자 이름 혹은 질병 이름을 사용해 정보를 검색해 볼수 있다. 이 사이트는 생물분자들의 기능, 특징, 구조, 돌연변이, 질병 등의 정보에 대한 통합 데이터베이스로서 특정 질병과 관련된 분자들에 대한 정보를 제공한다. OMIM-Morbid Map은 OMIM에 있는 질병과 관련된 유전자들의 정보와 그 유전자들의 세포유전학적 위치 (cytogenetic map location) 정보를 제공한다. 또한 OMIM과 링크되어 있어, 유전자에 대한 생물학적 추가 정보 및 관련 논문에 대한 정보를 얻을 수 있다. Disease Database는 인간 질환, 증상, 징후 등에 대한 정보를 제공하는 데이터베이스로서, 여러 가지 질병 및 의학적 용어에 대해 의료 교과서에서 사용하는 것과 동일한 색인법을 사용하여 서비스 한다. 이처럼 기존에 제공되는 질병관련 데이터베이스들은 각기 다른 목적으로 서로 다른 특징과 형태를 갖는 데이터들을 축적하여 제공하고 있다. 이 때문에 특정 질병에 대한 정보를 얻기 위해서는 여러 데이터베이스들을 검색해야하는 오버헤드가 발생한다. 따라서 본 논문에서는 이들 질병 관련 데이터들을 하나로 통합하고, 필요한 정보들을 정형화된 형태로 제공할 수 있는 검색 시스템을 제안하고자 한다.

Ⅲ. 제안하는 시스템

1. 시스템 구조

기존 질병 데이터에 대한 정보를 서비스하는 형태를 살펴보면, 많은 질병 데이터를 분류하고, 웹을 통해 데이터를 제공하는 다양한 서비스가 존재한다. 하지만 이러한 서비스들은 각기 다른 방법을 통해 데이터를 분류하고 있으며, 고차원 처리를 요구하는 신규 서비스와의 연계를 위한 방법이 존재하지 않는다. 본 논문에서는 기존 질병 데이터들을 통합한 아래 그림 1과 같은 새로운 시스템을 제안하여 기존 서비스의 단점을 해결하고자 한다.

그림 1은 본 논문에서 제안하는 디렉토리 시스템의 구조를 나타낸다. 기존에 서비스 되고 있는 질병 데이터의 통합을 위해 기존 질병 데이터를 서비스하는 각 웹 사이트 데이터를 수집하였다. 웹 서비스의 각 데이터는 웹파서와 텍스트 마이너를 통해 시스템에서 필요로 하는 자료의 형태로 수집하였다. 웹파서와 텍스트마이너를 통해 수집된 데이터는 통합관리자를 통해 통합 데이터베이스의 스키마에 따라 저장하였다. 통합된 데이터는 어플리케이션 서버 Glass Fish V2를 통해 데이터의 검색, 추가, 수정, 삭제 등의 서비스가 Open API를 제공한다.



▶▶ 그림 1. 제안하는 디렉토리 시스템의 구조

2. 데이터베이스의 통합

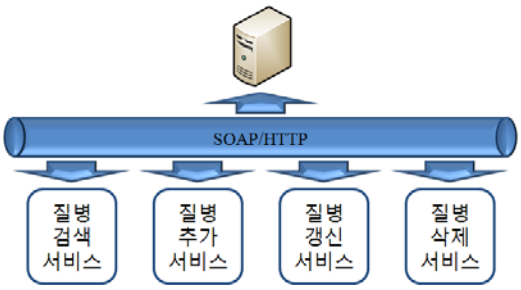
데이터 통합을 위하여 먼저 기존 질병 검색 서비스에서 제공하고 있는 질병의 속성에 대한 분석이 필요하다. 각 서비스 별로 중복된 속성에 대해서 서비스를 제공하는 부분도 있고, 서비스별로 다른 속성에 대한 서비스를 제공하는 부분도 있다. CHE는 질병에 대한 카테고리, 정확도, 유발물질, 원인 그리고 관련논문 등에 대해서 서비스를 제공하고 있다. 통합 관리를 하기 위해서 각 서비스로부터 해당 데이터를 추출해야하며, 추출을 위하여 각 서비스에 적합한 웹파서와 텍스터 마이너를 구현하였다. 그 다음, 통합 관리자를 통하여 질병의 식별자를 생성하고 식별자를 기준으로 통합 데이터베이스를 구성한다. 통합 데이터는 OpenAPI를 통하여 XML 문서로 사용자에게 제공된다. 그림 2는 통합 데이터의 XML DTD를 나타낸다. 각 웹서비스별로 얻어진 데이터의 속성을 분류하여 관리함으로써 데이터의 관리를 용이하게 하였다. 또한, 특정 질병 정보에 대해 다수개로 존재할 수 있는 속성에 대해 데이터베이스의 키 값을 이용하여 테이블별로 나누어 관리하기 때문에 불필요한 데이터 공간의 낭비를 줄였다. 질병에 대한 자료 검색시에는 XML 구조를 바탕으로 하나의 내용으로 표현하여 질병의 각 속성별 내용 파악이 용이하도록 하였다.

```
<disease>
   <id>식별자</id>
   <diagnosis>진단</diagnosis>
   <treatment>의학적처방</treatment>
   <symptom>증상</symptom>

   <alternativenames>
      <alias>대체질병명</alias>
      <alias>대체질병명</alias>
   </alternativenames>

   <categories>
      <category>카테고리1</category>
      <category>카테고리2</category>
   </categories>

   <causesList>
      <causes strength="unknown">유발물질</causes>
      <causes strength="unknown">유발물질</causes>
   </causesList>

   <geneList>
      <gene>관련GENE</gene>
      <gene>관련GENE</gene>
   </geneList>

   <linkList>
      <ensemble>Ensemble Link</ensemble>
      <omim>OMIM Link</omim>
      <entrez>Entrez Gene</entrez>
      <swissprot>SwissProt Link</swissprot>
   </linkList>

   <researchList>
      <research>관련논문정보</research>
      <research>관련논문정보</research>
   </researchList>

</disease>
```

▶▶ 그림 2. 통합 데이터의 XML DTD

3. 질병 검색 서비스를 위한 웹 서비스

본 논문에서는 인간 질병 검색 서비스를 위해 필요한 웹 서비스 시스템을 제안한다. 제안하는 웹 서비스 시스템은 크게 질병 검색을 위한 웹 서비스, 질병 추가를 위한 웹 서비스, 질병 갱신을 위한 웹 서비스, 질병 삭제를 위한 웹 서비스로 나누어진다. 각 웹 서비스는 개발자가 데이터 질병 검색 서비스를 위해 필요한 API를 제공한다. 제안하는 시스템은 웹 서버에 질병 검색 서비스를 제공하기 위한 통신 기능을 제공하며 웹 서버와 SOAP/HTTP을 통해 통신을 수행한다.



▶▶ 그림 3. 제안하는 웹서비스의 구성

질병 검색 서비스는 질병의 질병 식별자, 질병 이름, 질병 범주, 의학적 진단, 치료법, 증상, 유발 물질, 유전자, 논문 등으로 검색이 가능하다. 또한, 검색된 질병의 정보는 제안하는 데이터 구조 형태로 제공한다. 질병 추가 서비스는 현재 새로운 질병 정보가 유전학자들에 의해 꾸준히 발견되고 있기 때문에, 새로운 질병이 발견될 경우, 질병의 대한 정보를 쉽게 추가할 수 있는 웹 서비스다. 질병 갱신 서비스는 기존 서비스마다 다른 분류 방법을 사용하므로, 각기 다르게 분류된 기존의 질병 정보를 표준화된 질병 정보로 변경 가능하도록 하는 서비스다. 마지막으로 질병 삭제 서비스는 각기 다른 분류 방법으로 인해 동일한 질병 정보가 다르게 분류된 경우, 동일한 질병의 정보를 통합하고 불필요한 질병 정보를 삭제하기 위한 서비스다.

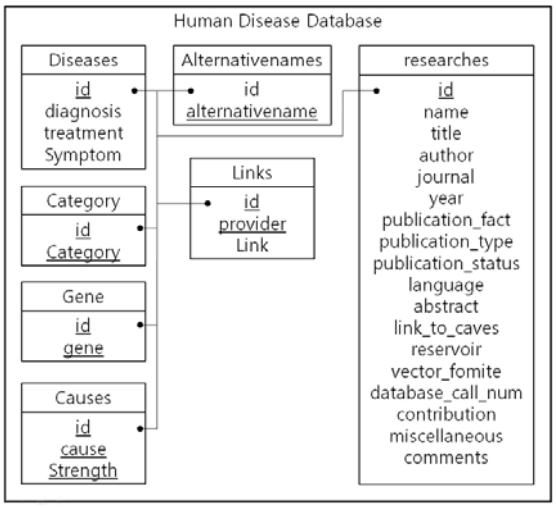
Ⅳ. 디렉토리 시스템 및 질병 검색 서비스의 설계 및 구현

1. 구현환경

제안하는 디렉토리 시스템은 CentOS 5.2 서버 환경에서 J2EE 1.4와 J2SDK 1.5를 이용하여 구현하였으며, 데이터베이스 관리 시스템으로 MySQL 5.0을 사용하였다. 또한, 디렉토리 시스템의 활용을 보이기 위해서 Ajax와 PHP를 이용하여 질병 검색 서비스를 구현하였다.

2. 통합 데이터베이스

그림 4는 Disease Database의 릴레이션 구성도를 보여준다. Disease Database는 질병에 대한 정보를 관리하기 위해서 7개의 테이블로 구성되어 있다. Diseases 테이블은 질병에 대한 정보를 관리하는 테이블로 질병의 고유한 ID와 질병을 판별할 수 있는 진단 정보 그리고 질병에 대한 의학적 처방과 증상 정보를 관리 한다. Category 테이블은 질병의 ID와 질병의 카테고리 정보를 관리 한다. Gene 테이블은 질병의 ID와 gene의 정보를 관리한다. Causes테이블은 질병의 ID와 질병에 대한 유발물질 그리고 유발물질의 강도 정보를 관리한다. Alternativenames 테이블은 질병의 ID와 대체 질병정보를 관리 한다. Links 테이블은 질병의 ID와 링크 제공자 그리고 Link 정보를 관리 한다. researches 테이블은 질병의 ID와 논문의 이름, 제목, 저자, 논문지, 개제 년도 논문 언어에 대한 정보 등을 관리 한다.



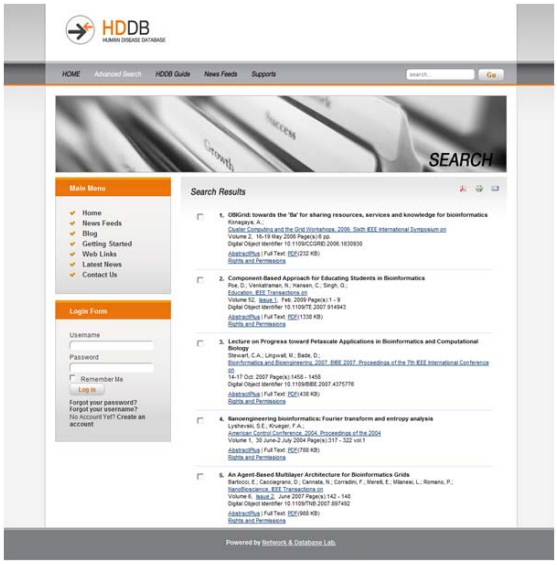
▶▶ 그림 4. 통합 데이터베이스의 릴레이션 구성도

표 1은 DB 테이블에 포함되어 있는 데이터의 수를 나타낸다. 질병의 수는 24개 정도이고, 대체질병명의 수

는 19644개 이다. 관련 Gene의 수는 22864개이며, Category의 수는 392개이다. 유발물질의 개수는 2760개 이며, link의 수는 8336개이고 Researchs는 209개의 정보를 포함한다.

표 1. DB 테이블에 포함된 데이터의 수

DB 테이블	데이터의 수
Alternativename	19644
Category	392
Causes	2760
Diseases	24
Genes	22864
Links	8336
Researchs	209



▶▶ 그림 5. 질병 검색 서비스 검색 결과 화면

3. 질병 검색 서비스 예제

본 절에서는 제안하는 질병 검색 서비스를 이용한 예제 페이지를 구현하여 제안하는 서비스의 유용성을 기술한다. 제안하는 서비스의 서비스 제공 페이지는 사전에 구축 된 통합 데이터베이스와 연동 및 XML 엘리먼트를 효율적으로 처리하여 결과를 제공하기 위해 AJAX (Asynchronous Javascript and XML) 와 PHP를 이용하여 구현하였다. 질병을 검색하기 위한 검색 폼 및 질병 검색 서비스에 새롭게 추가 된 데이터베이스나 서비스 변경 사항을 제공한다. 또한 해당 정보를 RSS 피드 및 이메일 등의 경로를 통해서 이용이 가능하도록 하이퍼링크를 제공한다. 그림 5은 통합 데이터베이스 검색 결과를 제공하는 화면을 나타낸 것이다. 상단의 검색 폼에 입력한 검색어에 대한 통합 데이터베이스 검색 결과를 제공하며, 해당 결과에 대한 상세 정보에 대한 열람이 가능하다. 사용자에 의해 요청 된 질의는 통합 데이터베이스로부터 XML 문서를 결과로 반환한다. 서비스 페이지는 XML데이터를 처리하여 동적으로 페이지를 생성한다.

V. 결론 및 향후 연구

본 논문에서는 이종의 질병 데이터베이스를 통합하기 위한 데이터 구조를 제안하고, 신규 서비스와 연계하기 위한 인프라를 제공하는 질병 디렉토리 시스템을 설계하고, 구현하였다. 이를 위해, 기존 데이터베이스를 분석하여, 다양한 속성을 XML 데이터 구조로 정리하고, 웹파서와 텍스트 마이닝 도구를 이용하여 통합 데이터베이스를 구축하였다. 그리고, SOAP/HTTP 통신을 이용한 웹서비스를 통해 이용할 수 있도록 하였으며, 질병 검색 서비스 프로토타입 예제를 함께 구현하였다. 향후에는 수집된 질병 데이터의 연관성을 바탕으로 질병을 보다 체계적으로 분류하고, 질병에 작용하는 기전, 단백질, 상호작용 등의 다양한 정보들을 수집하고 제공하여 질병단위의 분자 네트워크 연구나 질병 사이의 연계성 분석 등에 활용할 예정이다.

■ 참 고 문 헌 ■

[1] Luscomebe N.M and G. D, G. M., "What is bioinformatics? A proposed definition and overview of the filed", Methods Inf. Med 40:346-358, 2001.

[2] Lesk A. M. "Introduction to bioinformatics", pp.2-20, Oxford Iniversity Press, United Kingdom, 2002

[3] CHE, http://database.healthandenvironment.org/
[4] Gastro net,
http://www.gastro.net.au/gastrodiseases/
[5] Findis, http://www.findis.org/
[6] AID, http://www.uni-rostock.de/aidb/
[7] 3DinSight,
http://gibk26.bse.kyutech.ac.jp/jouhou/3dinsight/
[8] OMIM-Morbid Map,
http://www.ncbi.nlm.nih.gov/Omim/
[9] DiseaseDatabase,
http://www.diseasesdatabase.com/