웹기반 문헌분석 및 생물학적 네트워크 분석시스템 개발

Web based Text-mining and Biological Network Analysis System

서 동 민, 조 성 훈*, 안 광 성*, 유 석 종, 박 동 일**
한국과학기술정보연구원, 과학기술연합대학원대학교,
(주)PDXen*, 강북삼성병원**

Dongmin Seo, Seok Jong Yu, Sung-Hoon Cho,
Kwang-Sung Ahn, Dong-Il Park
Korea Institute of Science and Technology Information,
Univ. Science & Technology,
PDXen Biosystems Inc.*,
Kangbuk Samsung Hospital**

요약

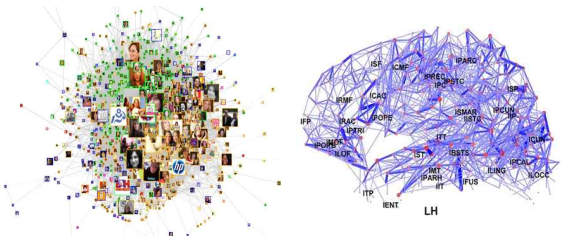
다양한 위상학적 관계(topological relation)를 분석하는 네트워크 분석은 복잡한 데이터에서 숨어있는 특성과 사실을 발견하는 기술로 최근 빅데이터 분야에서 데이터 분석 핵심 기술로 급부상하고 있다. 본 연구에서는 질병연구에 핵심적인 생물학적 네트워크의 생성 및 사용자 친화적인 네트워크 분석시스템을 개발하였다. 개발한 시스템은 PubMed에서 특정 질병과 관련있는 논문 요약 정보를 자동 수집후 텍스트마이닝을 통해 질병 관련 화합물, 유전자 그리고 상호작용 정보를 추출해 생물학적 네트워크를 생성하는 기능을 제공한다. 또한, 연구자가 손쉽게 생성된 네트워크에 대한 검색 및 다차원 분석을 수행할 수 있는 기능을 제공한다. 마지막으로 개발한 시스템의 우수성을 입증하기 위해 크론병(Crohn's Disease)에 대한 적용사례를 소개한다.

I. 서론

컴퓨터와 인터넷 등의 IT 기술들의 발달로 과거에 수행되지 못한 빅데이터 구축 및 분석이 가능해지면서 빅데이터 관련 연구가 급속히 발전하고 있다. 최근에는 유전체학의 발전, 웨어러블 디바이스의 확산, IT/NT의 발전 등에 따라 방대한 양의 바이오-메디컬 데이터가 생산되고, 이에 따라 빅데이터를 활용한 헬스케어 산업이 급속히 발달하고 있으며, 이와 관련된 빅데이터 기술은 국민의 건강 증대와 건강한 고령 삶을 제공하는 핵심 기술로 급부상하고 있다[1, 2].

그림 1의 (a)는 페이스북에서 인맥 구조를, 그림 1의 (b)는 뇌를 구성하는 중요 화합물, 유전자/단백질 구조를 가시화 한 것인데, 모두 네트워크 구조를 보이고 있다. 네트워크는 하나 이상의 데이터를 데이터 간 관계를 기반으로 연결시킨 자료구조를 의미하고, 네트워크 구조를 분석하면 밝혀지지 않은 데이터 간 특성 및 패턴을 발견할 수 있기 때문에 대부분의 빅데이터 분석은 네트워크 분석을 기반으로 한다. 최근 빅데이터가 다양한 분야에서 활용되면서 방대한 양의 네트워크 데이터가 생성되고 있고, 이에 따라서 대용량 네트워크 데이터를 효율적으로 분석할 수 있는 네트워크 분석 시스템의 중요성이 증가하고 있다[3].

특히, 연간 7조원이 넘는 생물정보 시장에서 최근 정밀의료(맞춤형 치료)의 등장에 따라 지식기반의 연구 수요가 증가하고 있지만, 질병 관련 화합물, 유전자 그리고 상호작용 정보에 대한 생화학적 기작을 손쉽게 분석할 수 있는 시스템이 갖추어져 있지 못한 실정이다. 그래서 본 연구에서는 PubMed에서 특정 질병과 관련있는 논문 요약 정보를 자동 수집후 텍스트마이닝을 통해 질병 관련 화합물, 유전자 그리고 상호작용 정보를 추출해 생물학적 네트워크를 생성하고, 연구자가 손쉽게 생성된 네트워크에 대한 검색 및 다차원 분석을 수행할 수 있는 기능을 제공하는 네트워크 분석시스템을 개발하였다. 또한, 개발한 시스템의 우수성을 입증하기 위해 크론병(Crohn's Disease)에 대한 네트워크 생성 및 분석을 수행하였다.



(a) Facebook network (b) Brain network

▶▶ 그림 1. 대용량 네트워크 데이터 예

(출처: http://www.wired.com/2012/04/facebook-disease- friends/, https://commons.wikimedia.org/wiki/File:Network_representation_ of_brain_connectivity.JPG)

Ⅱ. 개발한 생물학적 네트워크 분석시스템

1. 질병 정보 데이터베이스 구축

PubMed는 미국 국립 보건원의 미국 국립 의학 도서관(NLM)이 정보 검색 Entrez 시스템의 일부로서 생명과학 및 생물의학 주제에 대한 참조 및 요약을 담고 있는 MEDLINE 데이터베이스로 전세계 많은 연구자들이 사용하고 있다. 크론병은 설사, 복부 통증, 체중 감소, 피로 등 여러 불편한 증상을 일으키는 만성 장염으로 현재 미국에만 약 56만 5,000명의 환자가 있지만, 의사들은 크론병의 발병 원인을 여태까지 정확히 밝혀내지 못하고 있다. 그래서 최근 많은 연구자들이 유전체 빅데이터 분석을 통해 크론병의 원인이 되는 중요 요소, 요소의 특징 그리고 요소들 간의 상관관계를 발견해 효과적인 치료법을 찾으려는 연구를 시도하고 있다.

본 연구에서 개발한 시스템은 크론병 데이터베이스 구축을 위해 먼저, 미국 국립생물정보센터에서 제공하는 eUtils API[4]를 통해 PubMed에서 약 250,0000건의 문헌정보를 수집했다. 수집된 문헌정보는 XML 포맷으로 SAX 파서를 통해 제목(ArticleTitle), 요약(AbstractTitle), 중심어(KeywordList) 내에서 “Crohn’s Disease”를 포함하는 931건의 문헌정보에 대해 PubMed ID(PMID), 제목, 요약, 게재일(PubDate) 정보를 별도로 수집했다. 그리고 화합물 정보를 전문적으로 추출할 수 있는 개체명인식 도구인 OSCAR4[5], 유전자와 단백질 개체명 인식 도구인 ABNER[6]을 활용하여 총 1,719개의 생물학적 개체를 추출했다. 또한, 추출된 개체를 기반으로 관계 인식 도구인 MKEM[7]을 활용하여 총 2,024개의 생물학적 상호작용 정보를 추출하였다. 마지막으로, 추출된 정보들에 대한 정확도를 높이기 위해서 관련 분야 전문가를 통해 수작업으로 정제 작업을 진행하였다. 그림2는 크론병 관련 PubMed(Pubmed_ID)로부터 추출된 개체1(Entity1, Entity1_Type)과 개체2(Entity2, Entity2_Type)에 대한 관계(Effect, Effect_ Type) 일부를 보여준다.

pubmed_id	entity1	entity1_type	entity2	entity2_type	effect	effect_type
27643741	STAT6	Protein	acetylcholine	Small molecule	negative	suppressed
27638904	LLL12	Protein	STAT3	Protein	negative	attenuates
25923510	GM-CSF	Protein	matrix metalloprotease-9 MMP-9	Protein	negative	blocked
27602757	IL-6	Protein	STAT3	Protein	negative	attenuated
27602757	STAT3	Protein	STAT3	Protein	negative	attenuated

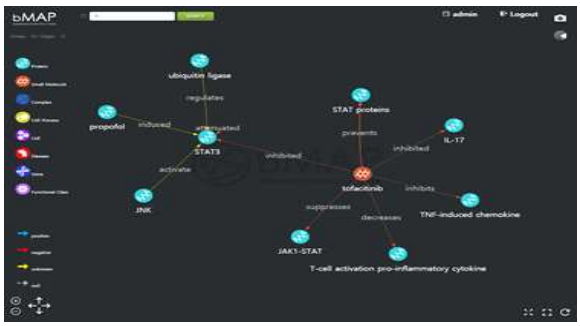
▶▶ 그림 2. 추출된 크론병 관련 생물학적 상호작용 정보 예

2. 다차원 분석 네트워크 가시화 시스템 개발

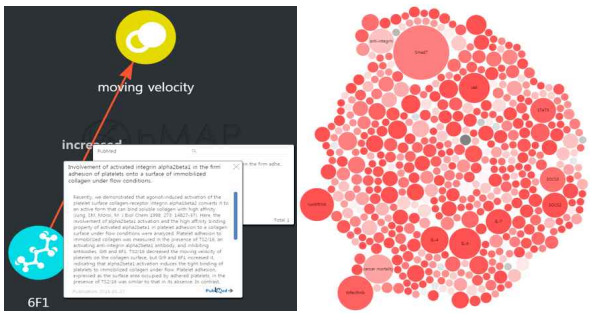
그림 3은 개발한 네트워크 가시화 시스템을 통해 가시화된 크론병 생물학적 네트워크를 보여주고 있다. 개발한 시스템은 HTML5, CSS3, Canvas 등의 최신 웹 기술을 기반으로 개발되었다. 그리고 사용자가 키워드로 입력한 한약재들에 대한 화합물, 유전자/단백질 정보 간의 상호작용 정보를 네트워크로 가시화한다. 특히, 네트워크를 구성하고 있는 노드를 선택하면, 해당 노드로 인해 영향을 받는 다른 화합물, 유전자/단백질들을 조회할 수 있고, 조회된 정보를 기반으로 한 기존 네트워크 확장을 통해 다차원 분석을 용이하게 할 수 있는 기능을 제공한다. 또한, 그림 4와 같이 추출에 활용된 PubMed와 추출된 화합물, 유전자/단백질의 통계 정보를 기반으로 한 크론병 관련 연구 동향 정보를 제공한다.

Ⅲ. 결론 및 향후연구

본 연구에서는 최신의 텍스트마이닝 기술을 통합하여 질병 분야 문헌 분석을 손쉽게 수행할 수 있는 분석 시스템을 개발하였으며, 크론병을 대상으로 관련된 약재 정보와 생물학적 상호작용 정보를 추출하여 통합 데이터베이스를 구축하였다. 또한 연구자가 직관적으로 생물학적 네트워크 정보를 분석할 수 있도록 생물학적 네트워크 가시화 도구를 최신 웹 기술을 활용하여 개발하였다. 향후 연구로는 추출된 개체 및 상호작용 정보에 대한 정확도를 높이기 위한 텍스트마이닝 기술을 고도화할 계획이다.



▶▶ 그림 3. STAT3을 기준으로 확장된 네트워크 탐색 예



▶▶ 그림 4. 추출 정보에 대한 PubMed와 연구동향 제공 예

■ 참 고 문 헌 ■

[1] 서동민, 최윤수, 전선희, 이민호, “바이오 패스웨이 다차원 분석 시스템 개발”, 한국콘텐츠학회논문지, 제14권, 제11호, pp.467-475, 2015.

[2] 윤미영, 권정은, 빅데이터로 진화하는 세상 - Big Data 글로벌 선진 사례, 한국정보화진흥원, 2012.

[3] 서동민, 유석종, 이민호, “대용량 네트워크 압축 기반 클러스터링 알고리즘 개발”, 한국콘텐츠학회 2016 춘계종합학술대회, 제14권, 제1호, pp.53-54, 2016.

[4] https://www.ncbi.nlm.nih.gov/books/NBK25501/

[5] D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy, and P. Murray-Rust, “OSCAR4: a flexible architecture for chemical text-mining”, J. Cheminform, Vol.3, No.1, pp.41, 2011.

[6] B. Settles, “ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text”, Bioinformatics, Vol.21, No.14, pp.3191-3192, 2005.

[7] M. S. D. L. Ali Z Ijaz, “MKEM: a Multi-level Knowledge Emergence Model for mining undiscovered public knowledge”, BMC Bioinformatics, Vol.11, No.2, 2010.