바이오메디컬 데이터베이스 및 텍스트마이닝 기술의 비교 분석 및 전망

조태원*, 이규범*, 강재우*†
*고려대학교 정보통신대학, †교신저자
e-mail:{twjoh, icebomb, kangj}@korea.ac.kr

Comparative analysis of Biomedical Databases and Text mining Technologies

Taewon Joh*, Kyubum Lee*, Jaewoo Kang*†
*College of Information and Communications, Korea University
†Corresponding author

요 약

분자 생물학을 통한 연구가 심화되면서, 생물학 정보는 기하급수적으로 늘어나고 있다. 그에 따라 바이오메디컬(생물학, 의학) 관련 논문들의 출판 및 등록 건수도 해마다 증가하고 있다. 그러나 바이오메디컬 문서들에서 유용한 정보를 추출하는 기술은 이러한 분야의 전문가 큐레이터(curator)에 의존한 경우가 많아서, 그 작업의 속도와 양적인 면에서 한계를 가지고 있다. 이러한 이유 때문에 바이오메디컬 문서를 기계학습을 통하여 분석하는 기법이 도입되기 시작하였다. 아직까지는 기계학습을 이용하여 구축된 데이터베이스가 소수에 불과하지만, 점차 증가하는 추세에 있다. 이러한 현 추이를 분석하고 향후의 추세를 예측하고자 텍스트마이닝 기술이 생물학과 의학 분야에서 어떻게 사용되며, 그 정보들이 어떻게 관리되는지 연구, 조사 하게 되었다.

현재 바이오메디컬 관련 데이터베이스들이 여러 기관 및 단체에 의해 구축 및 관리되고 있으며, 국가적인 프로젝트로서 이러한 데이터베이스들을 통합하는 과정을 진행하고 있다. 이처럼 국가기관의 주도하에 데이터베이스를 통합하여 관리하고자 하는 노력들이 계속되고 있어, 앞으로는 바이오메디컬 자료들을 검색하기가 보다 용이해질 것으로 생각된다.

텍스트마이닝을 이용하여 바이오메디컬 정보들을 추출하는 기술은 초기에는 공동 발생(co-occurence)과 같이 단순한 통계적 방법을 이용하였지만, 최근에는 다른 문서에서 추출된 정보와 기존의 정보들을 연계하여 새로운 정보를 추출해 내는 기법이 확산되고 있음을 알 수 있었다.

1. 서론

생물학과 관련한 연구에서 다양한 변화가 있어왔다. 특히, 분자생물학 분야에서의 발전은 생물학이 전산기술에 의존할 만큼 많은 데이터를 양산하였다. 생물학에서 중요하게 여겨지는 정보들로는 유전자 및 유전자의 기능, 단백질 및 단백질의 기능, 단백질-단백질 상호작용, 단백질의 패스웨이 등이 있는데[1], 이와 같은 정보들의 양이 많아지면서 이것을 저장하고 검색하기 위한 시스템들의 개발이 진행되고 있다. 데이터마이닝의 기술을 이용하여 새로운 정보들을 예측하는 방법의 사용도 많아지고 있다.

우선 살펴볼 것으로는 생물학 관련 정보를 저장하는 데이터베이스들이 있는데, 그 데이터베이스들은 주로 많은 유전자들의 명칭과 유전자들의 패스웨이 정보, 단백질-단백질 상호작용 정보 등을 담고 있다. 뿐만 아니라, 이미 많은 생물학 실험을 통해서 알려진 유전자 관련 정보들과 단백질의 상호작용 정보들이 여러 연구소 및 국가 지정기관들의 데이터베이스에 의해 관리되고 있다[2]. 이러한 정보들의 관리 방식에는 몇 가지 문제들이 있는데, 그중 첫번째는 다르게 이름 붙여진 유전자가 결국 같은 종류의 유전자임이 밝혀지는 경우가 많다는 것이다. 유전자의 별칭이 많은 경우에 이런 현상이 발생하며, 이 경우에 대한 관심이 2000년대 초반에 이르러 심화되기 시작했다. 두번째는, 통합된 데이터베이스의 부재로 인해 원하는 정보를 찾기 위해 많은 수고가 필요하다는 점이다. 이에 대한 문제의 인식으로 인해 각국에서는 중복된 정보의 저장을 피하고, 쉽고 빠르게 다양한 정보에 접근할 수 있도록 데이터베이스의 통합을 시작했다. 유럽에서 시작된 GALEN 프로젝트와 같은 것들이 이러한 노력의 일환이라고 할 수 있다[3].

그동안의 연구를 통해서 얻어진 수많은 논문들에는 단백질-단백질 상호작용, 단백질 패스웨이 정보들처럼 유용한 정보들이 많다. 전문가를 큐레이터(curator)로써 사용하여, 문서들에서 유용한 생물학 정보들을 추출하여 데이터베이스에 저장하였는데, 오늘날에는 문서의 양이 너무 많아서 큐레이터가 하나하나 찾아내어 입력하기도 힘들고, 큐레이터를 고용하는 것도 쉽지 않다고 한다. 때문에, 자연어처리기법(NLP)을 이용, 자동적으로 문서상에서 원하는 정보를 추출하고, 예측하는 기술이 발전하였는데, 이와

같은 것을 바이오메디컬 텍스트마이닝(Biomedical Text mining)이라고 한다.

한편, 의학 분야에 적용할 수 있는 기술도 많이 등장하고 있는데, 그 예로 임상기록 및 기계 학습을 이용한 발(foot) 진찰 소견의 자동 분류에 관한 연구[4], 바이오메디컬 연구 네트워크에서 데이터의 안전한 공유에 관한 연구[5], 바이오메디컬 및 임상기록으로 부터 질병-약에 관한 지식을 자동으로 획득하는 것에 대한 연구[6], 전문가의 전산 환자 기록에서 당뇨병 치료의 질에 대한 정보의 추출[7], 임상방사선학 보고서의 자동 오류 발견에 대한 새로운 접근법, X-ray 보고서의 자동화된 분류법을 이용하여 손목 골절 환자를 확인하는 연구, 개인용 건강 기록 시스템을 통한 데이터 분석 및 저장, 임상 방사선학 보고서의 정확성을 위해 보고서에 등장하는 용어를 UMLS(Unified Medical Language System) 색인을 통해 확인하는 연구 등이 있다. 이처럼 의학 분야 역시 자동화된 시스템의 적용에 관한 연구에 관심이 높아지고 있다.

위에서 설명한 것과 같이 바이오메디컬 분야에서 자동화 및 전산화에 대한 노력이 특히 바이오인포매틱스(Bioinformatics)를 연구하는 사람들에 의해 이루어지고 있다. 본 논문은 위에서 제시한 바이오메디컬 분야의 유용한 도구들을 제시하고, 앞으로 이 분야에서 이루어질 발전에 대해서 생각해 보고자 한다. 제2장에서는 바이오메디컬 관련 정보를 저장하고 있는 데이터베이스의 종류 및 특징을 소개하였고, 제3장에서는 이러한 데이터베이스의 구축을 위해 바이오메디컬 정보를 추출하는 기술 및 시스템을 소개하고 있다. 그리고 제4장에서는 바이오메디컬 텍스트 마이닝의 전망에 대하여 고찰하였다.

2. 바이오메디컬 데이터의 집적

2.1 생물학 문서에서의 데이터베이스

생물학 문서들은 많은 유전자 및 단백질의 이름, 단백질-단백질 상호작용 정보, 단백질 패스웨이 정보들을 담고 있다. 이러한 정보들은 여러 연구기관에서 데이터베이스를 구축하여 운영하고 있다. 그 중 단백질-단백질 상호작용 정보는 생물학 분야에서 가지는 유용성이 크고, 다양한 종류의 데이터베이스가 구현되어 운영되고 있다. 또한, 각 데이터베이스마다 다루고 있는 단백질의 종류가 조금씩 달라서, 연구의 목적에 따라 적절한 데이터베이스를 선택하여 보다 나은 정보를 얻을 수 있다. <표 1> 에서는 이러한 데이터베이스들의 특징을 간략히 정리하였다[8].

한편, 이와 같이 각각의 데이터베이스로 흩어져 있는 단백질의 정보들을 쉽게 관리하기 위해서 국가적 차원의 통합 데이터베이스 구축을 위한 프로젝트들이 진행되었는데, 대표적인 것이 유럽의 GALEN 프로젝트와 UniProt 프로젝트이다. 국내에서도 KISTI의 바이오인포매틱스 센터에서 이러한 통합에 대한 시도를 하고 있다.

<표 1> 단백질-단백질 상호작용(PPI)정보 데이터베이스들

명 칭	PPI data	내 용
HPRD	Experimental, Curated	각 단백질 관련 주석 정보 (예, PTMs, substrate information, tissue expression, disease association, protein complexes, subcellular localization 등). 신호전달경로(Pathways) 검색
BIND	Experimental, Curated	단백질 중합체, biological pathways, 단백질 외의 상호작용, 1473종 이상의 정보
DIP	Experimental, Curated	여러 종의 PPI, 단백질 중합체
MINT	Experimental, Curated	여러 종의 PPI, 단백질 중합체, 단백질 외의 상호작용, MINT viewer
MIPS	Experimental, Curated, Functionally Predicted	여러 종의 PPI
GRID	Experimental, Curated	여러 종의 PPI
IntAct	Experimental, Curated	단백질 중합체, 여러 종의 PPI, 단백질 외의 상호작용, 웹 상에서 다양한 어플리케이션 제공, ProViz와 Hierarchy View
STRING	Experimental, Physically & Functionally Predicted	Physically, Functionally 예측된 자료들로 인해 방대한 자료의 양.

3. 바이오메디컬 텍스트에서의 정보추출 기술

3.1 바이오메디컬 연구에서의 시소러스 사용

이번 장에서는 다양한 생물학 및 의학과 관련된 문서로부터 정보들을 추출하기 위한 다양한 기술들을 설명하고자 한다. 우선 바이오메디컬 텍스트에 사용되는 수많은 전문용어와 그 용어들이 특수한 사용때문에 파싱(Parsing)과정에서 문제가 되어 왔다. 또한, 일반적인 영어 단어가 바이오메디컬 문서에서는 다른 의미로 사용되기도 한다. 때문에, 바이오메디컬 문서용 시소러스(thesaurus)의 필요성이 보다 높아졌다.

WordNet과 NLP 전문가 어휘집(NLP Specialist Lexicon)은 바이오메디컬 문서에서 사용하기에는 적당하지 않다. WordNet은 추가적으로 생물학 용어를 포함할 수 있지만 생물학 영역에서의 사용을 목표로 한 것이 아니기 때문이다[9].

바로 분자 생물학 용어와 같은 생물 특성화 용어의 집적을 위해서 생물용어사전(BioLexicon)과 같은 것을 사용한다. 오늘날에는, UMLS와 MeSH라는 미국 국립의학도서관(National Library of Medicine)에 의해서 집약된 시소러스가 유용하게 사용되고 있다[10].

3.2 생물 및 의학 문서에서 사용되는 기본적인 기능들

3.2.1 바이오메디컬 텍스트 마이닝의 초기 기술

바이오메디컬 텍스트 마이닝 문서에서 정보를 추출하기

위해 초기에는 공동발생(co-occurrence)에 기초한 접근법과 규칙기반(rule-based)에 기초한 접근법을 사용하였다. 이 두가지 방법은 각각 한계를 가지고 있는데, 공동발생에 기초한 방법은 추출된 유전자 혹은 단백질 관계들의 정확성을 보장할 수 없다는 것이고, 규칙기반에 기초한 방법은 문장단위로 프로세스가 처리되어 문맥 단위의 정보를 놓치게 된다는 데 있다[11]. 때문에 오늘날은 이 두 방법을 결합하여 사용한다. 특히, 규칙기반의 방법을 사용하기 위해서는 자연어 처리 기법(NLP)이 뒷받침 되어야 한다. 이 방법은 유전자 및 단백질의 이름을 확인하기 위해서 Part-Of-Speech(POS) 태깅(tagging) 기술을 사용한다. 구체적으로 설명하면 시소러스에 속하는 단어들을 글 속에서 찾은 후 문맥과 정의에 기초한 언어적 역할(Part of Speech)에 따라 표시(mark)하는 것이다. 이 때, 언어적 역할이란, 예를 들면, 명사, 동사, 형용사, 전치사와 같은 품사 또는 주어, 목적어, 수식어와 같이 정의된 역할을 의미하는 것이다. 이 때, 유전자 및 단백질의 이름을 인식하기 위해 적당한 시소러스를 사용한다.

3.2.2 의학 문서에 대한 정보 추출 기술

의학 분야에서 텍스트마이닝 기술은 주로 환자의 임상기록을 분석하는데 사용된다. 이러한 임상기록은 서술적인 특성을 가진 원문(Free Text)으로부터 의미있는 데이터를 추출하기 위해 자연어 처리 기법을 사용한다. 이것을 처리하기 위해서 어휘를 보충하고 표준 어휘로 바꾸어 주는 작업을 하게 되는데, 이 때 대표적으로 사용되는 시소러스가 UMLS(Unified Medical Language System) 이다. UMLS는 Metathesaurus, Semantic Network, SPECIAL Lexicon and Lexical Tools의 세 가지 파트로 이루어져 있다. 이렇게 표준화된 바이오메디컬 문서는 BioMedLEE와 같은 툴에 의해서 구조화 되고 인코딩된다. 또한, 표준화된 문서들은 Levenshtein difference 와 같은 Edit-distance를 이용하는 유사도 측정법을 사용하여 문서의 속성에 따라 군집화 되기도 한다.

3.3 정보 추출 기술의 변천

바이오메디컬 텍스트에서 추출하고자 하는 데이터들은 단백질-단백질, 유전자-유전자, 질병-약 등의 텍스트 기반의 정보들이 주류를 이룬다. 이를 위한 기술들의 단계별 과정을 살펴보면, IR (Information Retrieval), ER (Entity Recognition), IE (Information Extraction), 텍스트마이닝 (Text Mining), 통합(Integration)의 단계로 요약될 수 있다[13].

IR 시스템은 오랜 기간 사용되어져 온 기술이며, 특별한 주제어가 관련된 쿼리문을 이용하여, 해당되는 문서를 검색한다. 많이 알려진 IR 시스템은 MedMiner(1999), XplorMed (2001), E-BioSci(2003), Textpresso(2004), PubMed(2004), GoPubMed(2005), PubFinder(2005), EBIMed(2006) 등이 있다.

<표 2> 바이오메디컬 정보 추출 시스템의 URL 주소

분류	정보추출시스템	정보추출시스템 URL
IR	MedMiner	http://discover.nci.nih.gov/host/1999_medminer_abstract.jsp
	XplorMed	http://www.ogic.ca/projects/xplormed/
	E-BioSci	http://www.ebi.ac.uk/citexplore
	Textpresso	http://www.textpresso.org/
	PubMed	http://www.ncbi.nlm.nih.gov/pubmed/
	GoPubMed	http://www.gopubmed.com/web/gopubmed/1?WEB10O00h001000900001
	PubFinder	http://www.glycosciences.de/tools/PubFinder
	EBIMed	http://www.ebi.ac.uk/Rebholz-srv/ebimed
	Google Scholar	http://scholar.google.co.kr/
	RefMed	http://dm.postech.ac.kr/refmed
ER	MetaMap	http://mmtx.nlm.nih.gov/
	OpenGALEN	http://browse.opengalen.org/browser.aspx
	iHOP	http://www.ihop-net.org/UniPub/iHOP/
IE	SymText	Not Found
	Genies	Not Found
	PubGene 1.1	http://www.pubgene.com/public.html
	PreBIND	http://bond.unleashedinformatics.com/Action?
	iProLINK	http://pir.georgetown.edu/iprolink/
	BioMedLEE	http://lucid.cpmc.columbia.edu:8080/hux_test/index.html
	MedLEE	http://zellig.cpmc.columbia.edu/medlee/
	MutationFinder	http://mutationfinder.sourceforge.net/
	PhenoGO	http://www.phenogo.org/
TM	EDGAR	Not Found
	Arrowsmith	http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html
	PIE	http://pie.snu.ac.kr/
	PPI Finder	http://liweilab.genetics.ac.cn/tm/
IG	ProLinks	http://mysql5.mbi.ucla.edu/cgi-bin/functionator/pronav
	G2D	http://coot.embl.de/g2d/
	BITOLA	http://ibmi.mf.uni-lj.si/bitola/
	OpenDMAP	http://opendmap.sourceforge.net/
	CoreMine Medical	http://www.coremine.com/medical/
	STRING 8	http://string-db.org/

* IR: Information Retrieval ER: Entity Recognition IE: Information Extraction TM: Text Mining IG: Integration

ER 시스템은 바이오메디컬 문서에서 나타나는 유전자, 단백질 등과 같은 생물학 개체를 온톨로지 데이터베이스 등을 이용하여 찾아내는 기능을 가진다. ER 시스템의 특징으로는 IR과 IE의 중간다리 역할을 한다는 점과 특정 유전자 및 단백질과 관련이 있는 문서들을 인덱스로 상호 연결하는 기능을 한다. 많이 알려진 ER 시스템으로는 MetaMap(2001), iHOP(2005) 등이 있다[12].

IE 시스템은 바이오메디컬 텍스트에서 바이오메디컬 개체 중 이미 그 관계가 정의된 개체들을 추출해 내는 것을 목적으로 한다. IE 시스템은 기본적으로 다른 두 가지 접근법을 가지고 있는데, 앞에서 설명했던 공동발생에 기초한 접근과 규칙기반에 기초한 접근법이다. 대표적인 시스템으로는 iProLINK(2004), BioMedLEE(2004), Mutation Finder(2007), PhenoGO(2008) 등이 있다[13].

텍스트마이닝(Text Mining) 시스템은 이전에 알려지지 않은 정보 등을 추출하는 시스템이다. 텍스트마이닝 시스템의 추출된 정보들은 IE 시스템에 의해서 이미 추출된 정보를 사용하여 검증될 수 있다. 텍스트마이닝에서 사용되는 방식은, 'A leads to B' 와 'B leads to C' 라는 두 개의 문장에서 'A leads to C' 를 유추하며, 이 정보는 새롭게 알려진 정보이다. IE시스템의 대표적인 예는, Arrowsmith(2007), PIE(2008), PPI Finder(2009) 등이 있다.

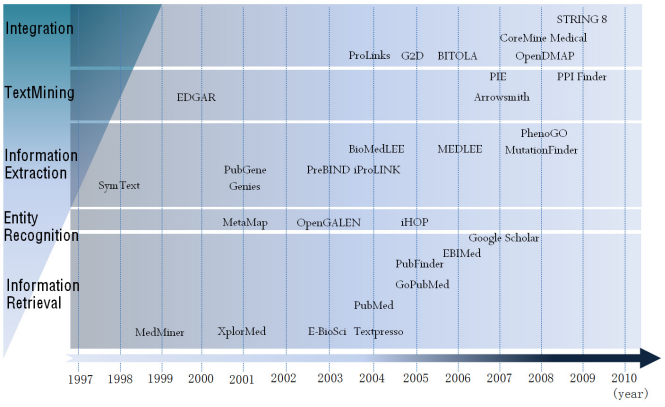
마지막으로 통합(Integration) 방식이 존재한다. 이 방식은 텍스트 형태의 자료와 그 이외의 형태(유전자 및 단백질 아미노산 서열과 같은 정보)의 데이터 분석 결과간의 상관관계를 밝혀 문서 및 단백질, 유전자 간의 연관성을 밝히는 연구방법이다. 예를 들면, ER 방법에 의해 추출되거나 참조되고 있는 데이터베이스로부터 추출된 생물학 개체 정보들과 마이크로어레이(Microarray)데이터를 클러스터링하여 얻은 유전자 혹은 단백질을 같이 비교 분석하여 관련이 있는 Medline 문서들을 검색하고, 해당되는 초록들을 텍스트 내의 키워드를 확인하기 위해 사용한다. STRING ver 8 (2009)과 같은 시스템이 유전자 서열과 Pubmed 에 포함된 문서 정보들을 보여주므로 통합 작업에서 사용하기에 적절하다. 각 기술에 해당되는 시스템은 <표 2>에서 확인해볼 수 있다.

4. 결론

본 조사결과를 바탕으로 하여, 몇 가지를 확인할 수 있었는데, 우선은 바이오메디컬 텍스트마이닝의 발전 방향이다. 다양한 시소러스를 통해 생물학 개념을 추출하고 있으며, 유전자와 유전자, 단백질과 단백질 등의 관계들을 서로 다른 문서를 분석하여 추출하는 기술이 많이 사용되고 있다. 또 연관된 유전자 및 단백질의 서열 정보와 텍스트 정보를 함께 이용하는 통합(Integration)이 많이 사용되고 있다. 향후 과제는 추출된 연관관계들의 신뢰도를 높여가는 것이며, 신뢰도가 높아진다면 생물학 및 의학 연구자들이 이와 같은 시스템에 더 많은 관심을 가질 수 있을 것이다.

한편 의학분야에서는 환자에 대한 진료 기록이 텍스트 마이닝의 중요한 데이터로 여겨지고 있다. 개인의 정보 보호때문에 아직 분석 자료로 사용하는데 많은 제약이 있지만, 이 부분에 대한 대책이 잘 마련된다면, 환자에 대한 진료 기록을 마이닝함으로써 많은 가치있는 정보를 발견해 낼 수 있을 것이다.

<표 3> 바이오메디컬 텍스트에서의 정보 추출 시스템들



참고문헌

[1] Roxana Danger, Paolo Rosso , Ferran Pla, Antonio Molina, "PPIEs: Protein-Protein Interaction Information Extraction system," Procesamiento del Lenguaje Natural, 2008.

[2] Andreas D. Baxevanis, B.F. Francis Ouellette, "Bioinformatics:A Practical Guide To The Analysis Of Genes And Proteins (3rd edition)," WILEYLISS.

[3] Gary D. Bader, Ian Donaldson, Cheryl Wolting, B. F. Francis Ouellette, Tony Pawson and Christopher W. V. Hogue, "GALEN Ten Years On: Tasks and Supporting Tools," IMIA, 2001.

[4] Serguei V.S. Pakhomov, PhD, Penny L. Hanson, Susan S. Bjornsen, Steven A. Smith, MD., "Automatic Classification of Foot Examination Findings Using Clinical Notes and Machine Learning," JAMIA, 2008.

[5] Stephen Langella, Shannon Hastings, Scott Oster, Tony Pan, Ashish Sharma, Justin, Permar, David Ervin, B. Barla Cambazoglu, Tahsin Kurc and Joel Saltz, "Sharing Data and Analytical Resources Securely in a Biomedical Research Grid Environment," JAMIA, 2008.

[6] Elizabeth S. Chen, PhD, George Hripcsak, MD, MS, Hua Xu, MS, Marianthi Markatou, PhD, and Carol Friedman, PhD., "Automated Acquisition of Disease-Drug Knowledge from Biomedical and Clinical Documents: An Initial Study," JAMIA, 2008.

[7] Jaco Voorham and Petra Denig, "Computerized Extraction of Information of the Quality of Diabetes Care from Free Text in Electronic Patient Records of General Practitioners," JAMIA, 2007.

[8] Zhou D, He Y., "Extracting interactions between proteins from the literature," J Biomed Inform, 2008.

[9] 김영택 외, "자연 언어 처리", 생능출판사.

[10] Shanfeng Zhu, Jia Zeng and Hiroshi Mamitsuka, "Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity," BioInformatics, 2009.

[11] Min He, Yi Wang, Wei Li., "PPI Finder: A Mining Tool for Human Protein-Protein Interactions," PLOS ONE, 2008.

[12] Robert Hoffmann, Alfonso Valencia, "Implementing the iHOP concept for navigation of biomedical literature," BioInformatics, 2005.

[13] Lars Juhl Jensen, Jasmin Saric and Peer Bork, "Literature mining for the biologist from information retrieval to biological discovery," Nature reviews, 2006.