제40회 한국정보처리학회 추계학술발표대회 논문집 제20권 제2호 (2013.11)

텍스트마이닝 기법과 구글데이터를 이용한 질병관련 유전자 식별

김정우[1], 김현진[1], 박상현[1,*]
[1]연세대학교 컴퓨터과학과
* 교신저자
email : {jwkim2013, chriskim, sanghyun}@cs.yonsei.ac.kr

Disease related Gene Identification Using Literature and Google data

김정우[1], 김현진[1], 박상현[1,*]
[1]Department of Computer Science, Yonsei University
*Correspondence sanghyun@cs.yonsei.ac.kr
email : {jwkim2013, chriskim, sanghyun}@cs.yonsei.ac.kr

요 약

텍스트마이닝은(Text mining) 바이오분야에서 사용되는 도구 중 하나이다. 본 논문에서는 전립선암(Prostate cancer)과 관련된 질병 유전자(Disease gene)를 찾기 위해 텍스트마이닝을 이용하여 유전자 네트워크(Gene-network)를 구축하였다. 추가적으로 구글(Google) 검색을 통해 네트워크 내의 유전자 노드(Node)들 사이의 간선(Edge)에 새로운 가중치(Weight)를 추가하고 네트워크를 재구성하였다. 구축된 네트워크에서 노드와 노드 사이의 가중치를 기반으로 전립선암과 관련된 질병 유전자를 추출하였다. 본 논문의 방법은 성공적으로 네트워크를 구축하고 질병 유전자를 찾았으며, 구글 데이터를 사용하지 않고 네트워크를 구축하는 경우보다 더 높은 정확성을 입증했다.

1. 서론

1990년 게놈 프로젝트(Genome project) 이후 유전자(Gene)에 관한 새로운 연구들이 진행되었으며, 그 중 하나가 인간의 질병과 관련된 유전자를 찾는 것이었다. 빠른 연구 속도에 비례하여 생물학 관련 분야에 많은 양의 정보들이 OMIM(Online Mendelian Inheritance)과 같은 데이터베이스에 저장 되었다. 이렇게 방대한 양의 자료를 이용하여 필요한 정보를 얻는 것에는 많은 노력과 시간이 소모 되었지만, 텍스트마이닝이라는 방법이 급속하게 발전함에 따라 자료를 활용하는데 필요한 노력과 시간이 급격하게 감소하였다.

텍스트마이닝은 생물학적 문헌들에서 유기체(Organisms)명, 단백질(Protein)명, 유전자명 등과 같은 생물학적 개체명을 인식하고, 인식의 결과를 기반으로 이들 간에 존재하는 생물학적으로 중요한 의미를 갖는 관계들을 추출한다. 이러한 기법을 이용하면 질병(Disease), 약(Drug), 유전자, 증상(Symptom) 등의 다양한 상호관계에 대한 정보를 생산 할 수 있고, 얻고자 하는 정보를 빠르고 손쉽게 얻을 수 있어 생물학적 상호관계를 추출하는데 효율적이다.

※이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2012R1A2A1A01010775).
논문접수 : 2013년 9월 29일
심사완료 : 2013년 10월 7일

텍스트마이닝 등장 이후에 *Adamic*[1]이 제시한 텍스트마이닝을 이용한 질병과 유전자 사이의 관계를 찾는 방법, *Lee*[2]가 제시한 텍스트마이닝을 이용하여 약과 유전사이의 관계, 그리고 유전자와 질병 사이의 관계를 도출하고, 도출된 결과를 바탕으로 약과 질병 사이의 관계를 도출하는 방법 등과 같은 텍스트마이닝을 이용한 다양한 방법론들이 제안되었다.

기존의 텍스트마이닝을 이용한 방법들은 많은 후보 질병 유전자(Candidate disease gene)들을 제안하고, 제안된 후보 질병 유전자들을 실험적으로 검증하기 위해서는 많은 노력과 시간이 필요했다. 이러한 문제점을 보완하기 위해 본 논문에서는 기존의 텍스트마이닝 방법에 구글 데이터를 추가로 이용하여 더 정확한 유전자 네트워크를 구축하고, 더 정확한 후보 질병 유전자들을 찾는 방법을 제안한다.

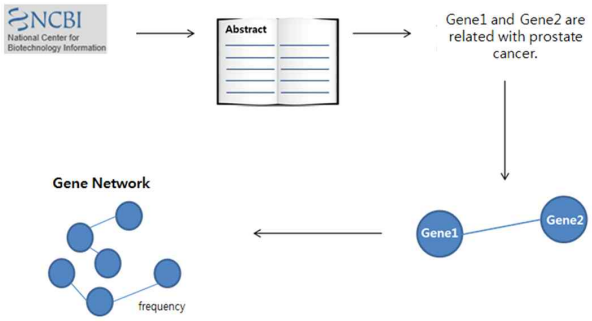
본 논문에서 제안하는 방법은 3단계로 이루어져 있다. 첫 번째 단계에서는 기존의 텍스트마이닝 방법을 이용하여 유전자 네트워크를 구축한다. 두 번째 단계에서는 구글 데이터를 사용하여 앞의 단계에서 구축한 유전자 네트워크에서 노드와 노드 사이의 간선에 가중치를 추가하여 네트워크를 재구성한다. 마지막 단계에서는 새로 구축된 네트워크를 분석하고, 노드들의 가중치를 계산하여 질병 관련 유전자를 추출(Extraction)한다. 기존의 텍스트마이닝

방법에 추가적으로 구글 데이터를 사용함으로써, 기존의 방법보다 더 정확하게 질병 유전자를 찾는 결과를 도출하였다.

본 논문은 다음과 같은 구성으로 이루어져 있다. 2장에서는 본 논문에서 제안하는 질병관련 유전자 식별 과정에 대한 설명을 한다. 3장에서는 본 논문의 방법론을 검증하기 위한 실험 및 결과, 4장에서는 결론과 추후연구 방향에 대해서 설명한다.

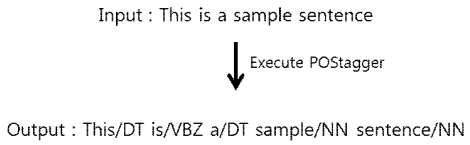
2. 질병관련 유전자 식별 과정

2.1 텍스트마이닝을 이용한 유전자 네트워크 구축



(그림 1) 유전자 네트워크 구축 과정

(그림 1)과 같이 4단계로 유전자 네트워크를 구축 할 수 있다. 먼저 PubMed[3]에서 전립선암과 관련된 문헌들의 Abstract를 다운받는다. 다운받은 Abstract내의 한 문장에 동시에 2개 이상의 유전자가 등장할 경우 유전자들 사이에 관련이 있다고 판단하고 두 노드 사이의 간선을 만든다. 문장에 등장하는 유전자를 식별하기 위해 아래와 같은 방법을 사용한다.



(그림 2) POStagger 처리 과정

(그림 2)와 같이 모든 문장은 POSTagger[4]를 이용하여 품사별로 구분하고, 구분된 품사 중에 명사만을 추출한다. 명사들 중에 유전자만을 선택하기 위해서 앞의 과정에서 추출된 명사들을 HGNC[5]에서 제공하는 유전자 심볼(Symbol)과 비교 하여 유전자를 구별한다.

이와 같은 방법으로 추출한 유전자 노드와 간선을 기반으로 네트워크를 구축한다. 노드와 노드 사이의 간선 가중치는 두 유전자가 얼마나 많은 문장에서 동시에 등장하였는가를 기반으로 빈도(Frequency)를 계산하여 설정한다. 구축한 네트워크는 전립선암과 관련된 유전자 네트워크이며, 노드는 유전자를 나타내고, 간선의 가중치는 두 유전자가 동시에 등장하는 문장의 개수이다.

2.2 구글 데이터를 이용한 새로운 가중치 설정

더 정확한 유전자 네트워크 구축을 위해 빈도라는 가중치 외에 구글 데이터라는 새로운 정보를 추가하였다. 텍스트마이닝을 통해 구축된 네트워크에서 두 노드 사이에 간선의 가중치를 구글 검색이라는 방법을 통해 새롭게 계산하였다.

(그림 3) 구글 검색 과정

(그림 3)과 같이 검색란에 두 노드 즉, 두 유전자의 이름을 검색한다. 검색 결과로 나온 관련 검색 결과 수를 두 노드 사이의 새로운 가중치로 설정하는데 사용한다. 새로운 가중치를 구하는 식은 다음과 같다.

$$New\,weight = Nor(FRE) + Nor(GSR) \quad (1)$$

$Nor(A)$: 임의의 수 A에 관하여 $Min-\max\ Normalization$ 을 수행한 결과 값

FRE : $Abstract$ 내에서 두 노드가 동시에 등장하는 문장의 빈도수($Frequency$)값

GSR : 두 노드를 동시에 검색창에 검색했을 때 나오는 구글 검색 결과 수

2.3 점수 설정

위의 식과 같은 방법으로 계산된 New weight를 노드 사이의 가중치로 설정하여 네트워크를 재구축한다. 실험을 위해 유전자 네트워크에 존재하는 모든 노드들에 대해 가중치를 기반으로 점수를 계산하고, 계산된 점수를 바탕으로 유전자 순위(Rank)를 정하였다. 각 노드의 점수는 다음과 같이 계산한다.

$$Score(A) = \sum_{n=1}^{N(A)} Weight(A, A_n^+) \quad (2)$$

A_n^+ : 노드 A와 인접한 n번째 이웃

$N(A)$: 노드 A와 인접한 모든 이웃 노드의 개수

$Weight(A, B)$: 노드 A와 노드 B 사이의 $New weight$ 값

$Score(A)$: 노드 A의 점수

3. 실험

점수가 높은 상위 20개의 유전자에 대해서 검증을 하였으며, Frequency와 New weight를 가중치로 사용하는 각각의 경우로 나누어서 수행하였다.

3.1 실험환경

OS	RAM	CPU
Window 7(64bit)	8G	Intel i5-3570 3.4GHz

3.2 실험 방법

계산된 각 노드의 점수를 기반으로 내림차순 정렬하고, 상위 20개의 유전자에 대해 실제로 전립선암과 관련이 있는가를 검증한다. 빈도만을 가중치로 사용했을 때와 New weight를 가중치로 사용했을 경우를 비교하여 실험하였다. 검증을 위해서 전립선암과 관련된 유전자 정보를 담고 있는 다양한 데이터베이스를 이용하였다. PGDB[6], Sanger[7], DDPC[8], KEGG[9]의 전립선암과 관련된 유전자 정보 데이터베이스들을 사용하여 검증을 수행 하였으며, 몇몇 유전자들은 추가적으로 문헌 검증을 수행하였다.

3.3 실험 결과

<표2>빈도를 사용하였을 때 유전자 순위

Rank	Gene symbol	Evidence
1	AR	PGDB
2	T	None
3	PC	None
4	ERG	Sanger
5	TMPRSS2	PGDB
6	GSTP1	PGDB
7	PTEN	KEGG
8	BRCA1	PGDB
9	EGFR	DDPC
10	BRCA2	PGDB
11	EGF	DDPC
12	GSTM1	PGDB
13	GSTT1	PGDB
14	APC	PGDB
15	CS	None
16	SRD5A2	PGDB
17	VDR	PGDB
18	CD4	None
19	MS	None
20	HR	None

<표2>는 빈도를 가중치로 사용하여서 각 유전자의 점수를 계산하고 점수가 높은 유전자부터 차례로 20번째까지 나타낸 표이다. 20개의 유전자 중에 14개의 유전자가 실제로 전립선암과 관련이 있었으며, 70%의 정확도를 가진다.

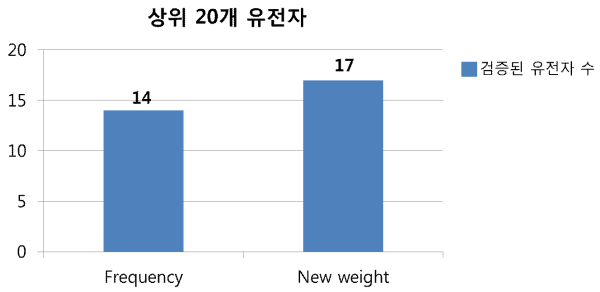
$$\text{정확도}(\%) = \frac{\text{질병과 관련있는 유전자 수}}{\text{전체 유전자 수}} \times 100 \qquad (3)$$

정확도는 식(3)과 같은 방법으로 계산하였다.

<표3> New weight를 사용하였을 때 유전자 순위

Rank	Gene symbol	Evidence
1	AR	PGDB
2	T	None
3	EGFR	DDPC
4	EGF	DDPC
5	PC	None
6	ESR1	PGDB
7	PTEN	PGDB
8	GSTP1	PGDB
9	ERG	Sanger
10	BRCA1	PGDB
11	TP53	PGDB
12	TMPRSS2	PGDB
13	TNF	PGDB
14	CDK2	None
15	APC	PGDB
16	BRCA2	PGDB
17	IGFBP3	PGDB
18	IL6	literature[10]
19	CYP1A1	literature[11]
20	CXCR4	literature[12]

<표3>은 New weight를 가중치로 사용하여서 각 유전자의 점수를 계산하고 점수가 높은 유전자부터 차례로 20번째까지 나타낸 표이다. 20개의 유전자 중에 17개의 유전자가 실제로 전립선암과 관련이 있었으며, 85%의 정확도를 가진다.



(그림 4) Frequency와 New weight 비교분석

(그림 4) Frequency를 이용하는 기존의 방법은 상위 20개의 유전자 중에 14개가 전립선암과 관련이 있다고 검증되었다. 본 논문에서 제시하는 구글 데이터를 추가하여 네트워크에 가중치를 재계산하는 방법은 상위 20개의 유전자 중에 17개의 유전자가 전립선암과 관련이 있다고 검증되었다. 이는 본 논문의 방법이 기존의 Frequency를 사용하는 방법보다 구글 데이터를 추가함으로써 더 높은 정확성을 가지는 네트워크를 구축했다는 것을 입증했다.

4. 결론

바이오 관련 문헌들의 양이 많아짐에 따라, 문헌들에서 효율적으로 필요한 정보를 얻는 방법이 필요하게 되었다. 그 중 하나가 텍스트마이닝 기법이다. 본 논문에서는 기존에 문헌들을 기반으로 수행하는 텍스트마이닝 방법론에 추가적으로 구글 데이터라는 새로운 정보를 추가하여 접근함으로써 새로운 유전자 네트워크를 구축하였다. 구축된 네트워크에 노드들의 점수를 계산하고 점수가 높은 상위 20개의 유전자들을 검증한 결과, 기존의 방법으로 구축한 네트워크는 20개의 유전자중 14개의 질병 관련 유전자를 찾았다. 반면에 본 논문에서 제안하는 방법론은 기존의 방법보다 3개가 많은 17개의 질병 관련 유전자를 찾았다. 이는 본 논문에서 제시한 방법론이 기존의 방법으로 구축한 네트워크 보다 높은 정확성을 가짐을 입증하였다.

본 논문은 구글 데이터 중에 검색 결과의 수라는 일부분의 정보만을 활용하였으나, 구글 데이터에는 활용할 수 있는 정보들이 많이 존재한다. 많은 정보들을 활용하여 더 정확한 네트워크 구축을 수행 할 예정이다. 또한 네트워크 분석에 있어서 다양한 각도로 해석하고, 점수를 계산하여 최적의 결과를 찾도록 개선할 예정이다.

참고문헌

[1] Admic et al, “A literature based method for identifying gene-disease connections”, IEEE Computer Society Bioinformatics Conference, pp. 109-107, 2002

[2] Lee, “Discovering context-specific relationships from biological literature by using multi-level context terms”, BMC Medical Informatics and Decision Making, 2012

[3] PubMed National Center for Biotechnology Information , http://www.ncbi.nlm.nih.gov/pubmed

[4] Stanford Log-linear Part-Of-Speech Tagger, http://nlp.stanford.edu/software/tagger.shtml

[5] HUGO Gene Nomenclature Committee, http://www.genenames.org/

[6] Human Prostate Gene DataBase, http://www.urogene.org/pgdb/

[7] Sanger institute, http://www.sanger.ac.uk/

[8] Dragon Database of Genes Implicated in Prostate Cancer, http://cbrc.kaust.edu.sa/ddpc/

[9] Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/

[10] Zhang H et al, “The interleukin-6 -174G/C polymorphism and prostate cancer risk: a systematic review and meta-analysis“, International Journal of Urology, vol. 88, No. 4, pp. 447-453, 2012

[11] Abjal Pasha Shaik et al, “CYP1A1 Polymorphisms and Risk of Prostate Cancer”, International Journal of Urology, vol. 6, No. 2, Spring 2009

[12] de Muga S, “CXCR4 mRNA overexpression in high grade prrostate tumors: lack of association with TMPRSS2-ERG rearrangement”, Cancer Biomark, vol. 12, No. 1, pp. 21-30, 2012