생물학적 문헌 데이터와 구글 데이터를 활용한 질병 연구

연세대학교 ▌박상현* · 김정우

1. 서 론

1990년 게놈 프로젝트(Genome project) 이후 유전자(Gene)에 관한 새로운 연구들이 진행되고 있으며, 질병(Disease)과 유전자 사이에 관련성이 있다는 사실을 확인하였다. 수많은 연구 결과들은 문헌 데이터로 기록되고 있고, 이러한 문헌 데이터들은 데이터베이스(Database)로 구축되어 저장된다. 생물학적(Biological) 문헌 데이터를 저장하고 있는 가장 유명한 데이터베이스로 PubMed[1]가 있다. PubMed는 1997년 6월부터 개인이 사용 가능한 무료 MEDLINE 검색 서비스를 시작했고, 현재는 약 2,400만 개 이상의 생물학 문헌 데이터를 제공하고 있다.

생물학 문헌 데이터들은 유전자, 단백질(Protein), 화학성분(Chemical compound) 등 질병 관련 연구에 있어서 중요한 내용을 포함하고 있다. 하지만 데이터의 양이 방대하고 산재되어 있어, 연구자들이 일일이 모든 문헌 데이터를 확인하는 것은 거의 불가능하다. 따라서 이와 같은 문제를 해결하는 방법을 찾는 것이 하나의 과제가 되었고, 그 방법론 중의 하나가 텍스트 마이닝(Text-mining)이다.

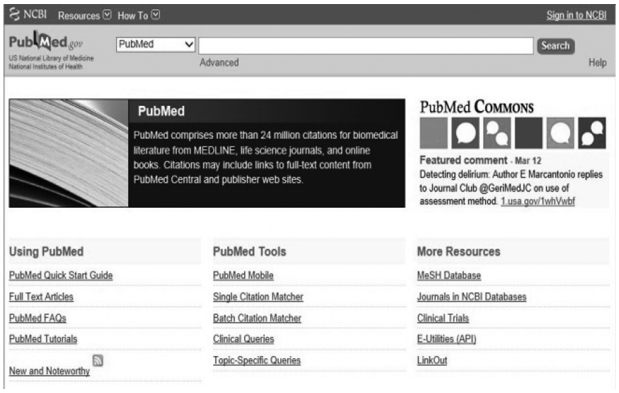


그림 1 생물학 문헌 데이터베이스 PubMed

텍스트 마이닝이란 문헌 데이터에 나타나는 단어들을 분석하여 필요한 지식을 추출하는 방법론을 의미한다. 텍스트 마이닝을 이용하면 모든 문헌 데이터를 직접 확인할 필요 없이, 사용자가 원하는 데이터를 추출할 수 있다. 생물학 분야에서 텍스트 마이닝은 크게 세 가지로 분류된다. 생물학적 문헌들로부터 유기체명, 단백질명, 유전자명 등과 같은 생물학적 개체들을 추출하는 기법과[2, 3, 4], 이들 사이에 존재하는 생물학적으로 중요한 의미를 갖는 관계(Relationship)들을 추출하는 방법[5, 6, 7], 그리고 전체적인 생물학 개체들 사이의 관계를 효과적으로 구성하고 표현하는 생물학적 네트워크(Network) 구축 기술 등이 있다[8, 9, 10].

텍스트 마이닝은 정보를 추출할 때, 단어의 출현 빈도(Frequency)를 사용하는 기법과[11] 동시 출현(Co-occurrence)빈도를 사용하는 기법[12]을 주로 사용한다. 단어의 출현 빈도란 문헌상에서 단어들이 등장하는 횟수를 의미하며, 특정 단어의 출현 빈도가 다른 단어들에 비해 높으면, 해당 단어는 다른 단어들에 비해 더 중요하다고 여겨진다. 동시 출현 빈도란 문헌상에서 하나 이상의 생물학적 개체들이 동시에 등장하는 횟수를 의미한다. 두 개 이상의 개체가 같은 문헌 혹은 같은 문장에 빈번하게 등장하면, 그 개체들 사이에는 생물학적으로 중요한 관계가 있다고 판단하게 된다. 출현 빈도를 사용하면 상대적으로 중요한 생물학적 개체를 추출할 수 있으며, 동시 출현 빈도를 사용하면 개체들 사이의 관계를 추출함에 있어서 연구가 많이 진행되어 검증된 관계들을 추출 할 수 있다.

문헌으로부터 연구들의 결과로 밝혀진 정보를 찾는 목적을 위해서는 출현 빈도와 동시 출현 빈도는 유용하게 사용 될 수 있다. 하지만 생물학 분야에서의 또 다른 목적인 새로운 개체나 개체들 사이의 관계를 찾는 것에서는 출현 빈도와 동시 출현 빈도를 사용하는 것이 오히려 단점으로 작용 될 수 있다. 출현 빈도와 동시 출현 빈도를 기반으로 추출하게 되면, 기존에 많이 연구되고 알려진 개체와 관계들만이 추출될 수 있

* 종신회원

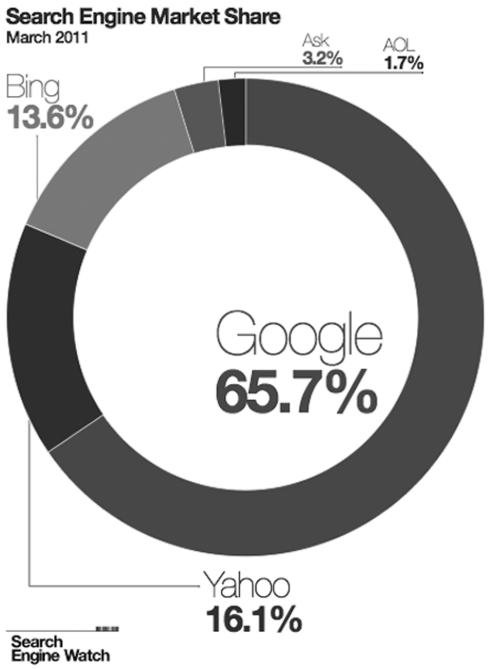
† 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2012R1A2A1A01010775).

다는 한계점을 가진다. 그 때문에 새로운 개체 또는 관계를 추출하기 위해서는 추가적인 데이터의 처리나 새로운 관계 추출을 위한 다른 정보가 필요하다. 여기서 다른 정보란 문헌상에서의 출현 빈도나 동시 출현 빈도와 같이 문헌을 분석하여 얻을 수 있는 데이터를 제외하고 새로운 관계 추출을 위해 사용 가능한 정보를 말한다. 본 원고에서는 구글(Google) 데이터를 새로운 생물학적 개체와 개체들의 관계를 추출할 수 있는 데이터로 소개하고, 구글 데이터를 활용한 질병 연구에 대한 전망에 대해 논의해 보고자 한다.

2. 구글 데이터

구글은 페이지 순위라는 독자적인 알고리즘(Algorithm)을 개발해 검색 시장을 장악한 세계 최대의 인터넷 검색 서비스 회사다. 130개 이상의 언어로 검색 인터페이스(Interface)를 제공하고 있어 전 세계적으로 많은 사용자를 보유하고 있다.

그림에서 볼 수 있듯이 구글은 전 세계 검색 엔진 시장에서 약 65.7% 점유율을 보인다. 다른 검색 엔진들과 비교했을 때 월등한 수치다. 가장 많은 사용자를 보유하고 있는 검색 엔진인 만큼 방대하고 다양한 데이터를 생산한다. 그 때문에 구글은 정보를 가장 많이 포함하는 데이터베이스라고 할 수 있으며, 구글 데이터는 유용한 정보를 찾기 위한 자료로 활용될 수 있다.



출처 : comScore(2011년 3월) [13]

그림 2 세계 검색 엔진 점유율

2.1 구글 스칼라와 PubMed

구글 스칼라(Google Scholar)는 구글에서 제공하는 검색 서비스 중의 하나이다. 주로 학술 검색 서비스를 제공하며, 논문, 학술지, 간행물 등의 검색을 수행한다. PubMed와 같이 구글 스칼라를 통해서도 생물학적 문헌 데이터들을 수집할 수 있다. 검색에 따른 문헌 데이터를 제공한다는 점에서 목적은 같지만, 두 검색엔진 사이에는 차이점이 존재한다. PubMed는 특정한 생물학 관련 논문지로부터 문헌 데이터를 추출한다. 또 색인으로 MeSH Term이라는 것을 사용하여 구체적이며 정확한 검색을 수행하고, 사용자가 원하는 정확한 데이터 제공을 가능하게 한다. 반면에 구글 스칼라는 논문지(Journal articles), 학회 발표지(conference proceedings), 기관 저장소(institutional repositories) 등 다양한 텍스트로부터 문헌 데이터를 추출한다. 이 때문에 검색을 통해 PubMed 보다 더 다양하고 많은 데이터를 얻을 수 있다. 하지만 MeSH Term이라는 색인을 사용하지 않는다는 점에서, PubMed에 비해 검색의 정확도가 떨어진다는 단점을 가진다. 정확도가 떨어진다는 단점이 존재하지만, 검색 과정에서의 전처리(Pre-processing)와 검색 결과에 대한 데이터의 후처리(Post-processing)를 적절히 적용한 후에 정제된 데이터를 사용한다면 구글 스칼라로부터 충분히 유용한 생물학 문헌 데이터를 추출할 수 있을 것이다. 생물학 주제 검색에 있어서 PubMed와 Google Scholar를 비교 분석하는 연구[14] 들도 진행되고 있으며, 향후 두 데이터베이스를 활용하여 생물학 문헌 데이터를 추출하는 방법론에 대한 많은 연구가 진행될 것이다.

2.2 구글 데이터의 이용 사례

실제로 구글 데이터를 활용해 질병과 관련한 유용한 정보들을 추출하는 연구들이 진행되고 있다. 2009년에는 구글 플루 트렌드(Google Flu Trends)라는 구글 데이터를 활용하여 인플루엔자(Influenza)의 지역별 발생(Outbreak)을 예측한 연구 결과가 있었다. 이 연구 결과에서는 구글 데이터를 사용하는 예측방법이 기존의 질병 관리 센터의 예측 시스템보다 7-10일 이전에 인플루엔자의 발생을 감지했다고 보고 했다.

다음 그림은 구글 데이터를 활용하여 시간에 따른 인플루엔자 감염 환자를 예측한 결과와 실제로 그 기간에 인플루엔자에 감염된 사람의 수를 비교 분석하여 나타낸 그래프이다. 그림에서 확인할 수 있듯이 예측데이터와 실제 데이터의 시간에 따른 감염자의 수에 대한 흐름이 상당 부분 일치하는 것을 볼 수 있다. 위 연구에서는 인플루엔자와 관련된 단어들을 검색하는

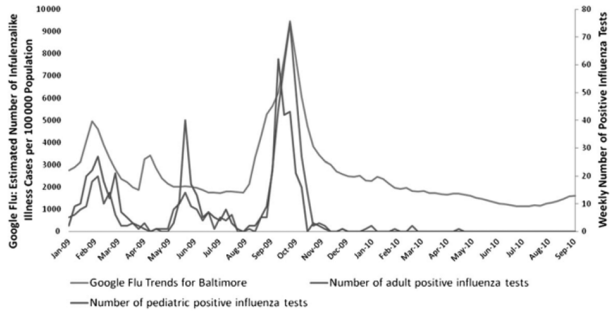


그림 3 구글 플루 트렌드 예측 결과[15]

사람과 실제로 인플루엔자의 증상을 보이는 사람 사이의 관계를 분석하여 인플루엔자의 발생을 감지하였다. 위 연구의 연구자들은 세상이 빠르게 변화하고, 많은 사람이 인터넷에 의존하기 때문에 기존의 전통적인 인플루엔자 예측 시스템보다 구글 데이터를 활용한 방법이 더 좋은 결과를 보였으며, 향후 전염성 질병의 예후를 예측하는 데 있어서 구글 데이터가 유용하게 사용될 것이라고 전망했다. 인플루엔자 이외에도 모기로부터 감염되는 뇌염의 일종인 웨스트 나일 바이러스(West Nile virus), 기관지의 염증을 일으키는 RS 바이러스(Respiratory syncytial virus), 조류독감(Avian influenza) 등의 전염성이 있는 질병에 대해서 연구를 진행하고 있다. 질병의 발병에 대한 예측은 질병을 예방하는 데 있어서 효과적으로 사용될 수 있으므로, 질병의 치료 연구만큼이나 중요하다. 또 위의 연구와 같이 구글 데이터를 활용하여 전염성이 있는 질병들의 향후 예측을 성공적으로 수행한 선례가 있는 만큼 앞으로 구글 데이터를 활용한 질병 예후 예측 방법에 대한 연구가 많이 이루어질 것으로 보인다.

2.3 구글 데이터를 활용한 질병 관련 유전자 식별

최근 구글 데이터를 활용해 질병 관련 유전자를 추출하는 LGscore라는 방법론이 소개되었다.

그림 4는 문헌 데이터와 구글 데이터를 활용한 질병 관련 유전자 식별 방법인 LGscore의 전체적인 프로세스를 나타낸다. LGscore는 PubMed로부터 특정 질병과 관련된 문헌 데이터를 얻고, 그 문헌 속에서 유전자들의 동시 출현 빈도를 기반으로 유전자 사이의 관계를 추출했다. 추출한 관계의 가중치를 계산하기 위해서, 추가로 구글 검색 결과(Google Search Results)라는 구글 데이터를 사용하였다. 구글 검색 창에 두 유전자의 이름을 검색한 후에, 결과로써 보이는 구글 검색 결과 수를 두 유전자 사이의 유사도(Similarity)로 계산하였다. 즉, PubMed로부터 얻은 문헌 데이터에서의 동시 출현 횟수와 구글의 문헌 데이터

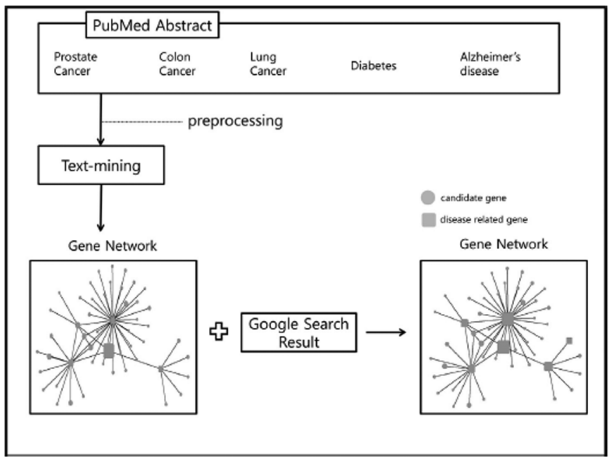


그림 4 LGscore의 프로세스[16]

에서의 동시 출현 횟수를 함께 사용하여 유전자 사이의 유사도를 계산하는 방법론을 제안한 것이다. 이렇게 추출된 유전자 관계들을 기반으로 유전자 네트워크(Gene-network)를 구축하였고, 네트워크를 분석하여 질병 관련 유전자를 식별하였다. 위의 연구에서는 다섯 가지 질병에 대해 실험을 진행하였으며, 기존의 질병 관련 유전자를 찾는 방법론들보다 구글 검색 결과를 사용하여 유전자를 추출하는 방법론이 우수한 결과를 보임을 입증하였다.

실제로 구글 데이터를 활용하는 생물학 분야의 다양한 연구들이 진행되고 있음을 확인하였다. 본 원고에서 소개하고 있는 구글 데이터들 외에도 웹 문서나 이미지, 기사, 연관 검색어, 사전 등 구글에서부터 얻을 수 있는 데이터는 무궁무진하다. 그렇기 때문에 기존의 생물학 연구와 구글 데이터의 융합은 향후 비전 있는 연구 분야로 전망이 있을 것으로 생각한다.

3. 문헌 데이터와 구글 데이터를 활용한 질병 연구에 대한 방향

이번 장에서는 생물학 분야의 연구에 있어서 문헌 데이터와 구글 데이터를 함께 활용 가능한 방법론들에 대한 생각을 기술한다.

3.1 구글 스칼라의 활용

구글 스칼라는 PubMed와 같이 검색 질의(Query)에 따라 문헌 데이터를 제공해준다. 하지만 같은 질의에 대해서도 서로 다른 결과를 보여준다.

그림 5는 PubMed와 Google Scholar를 비교 연구한 논문에서 발췌한 그림이다. 그림에서 확인할 수 있듯이, 같은 질의에 대해서도 PubMed와 Google Scholar의 검색 결과가 차이가 나는 것을 볼 수 있다. 검색마다

Table 2
Number of retrieved items

Search #	PubMed results	PubMed unique items	Google Scholar results	Google Scholar unique items
1	25	21	26	20
2	10	8	11	8
3	4	2	27	24
4	10	3	52	38
5	6	0	20	10
6	11	0	20	8
7	2	1	10	8
8	13	0	18	4
9	51	7	49	4
10	15	4	14	1

그림 5 PubMed와 구글 스칼라 검색 결과[17]

조금씩 차이는 있지만 모든 검색 결과에서 구글 스칼라를 이용해야만 얻을 수 있는 문헌 데이터들이 존재하는 것을 확인할 수 있다. 물론 아직은 구글 스칼라로부터 얻어진 문헌 데이터들의 정확도가 떨어지는 것은 사실이지만 이러한 점을 보완한다면 충분히 가치 있는 데이터가 될 것이다. 그렇기 때문에 기존의 PubMed 문헌 데이터에 구글 스칼라로부터 얻어지는 문헌 데이터를 추가로 사용한다면 더 많은 생물학 문헌 데이터를 활용한 연구가 가능 할 것으로 생각한다.

3.2 문헌 데이터와 구글 데이터를 활용한 전염성 질병 예측

앞에서 구글 데이터를 활용하여 인플루엔자의 발병을 예측한 연구에 대해서 소개했었다. 인플루엔자를 예측하기 위한 자료 중에 하나로 인플루엔자와 관련된 단어들을 활용하였는데, 이 부분에서 기존의 바이오 텍스트 마이닝을 활용할 수 있다. 문헌 속에서 특정 질병과 관련된 단어를 텍스트 마이닝을 통해 분석하고 질병과의 관련성을 순위화하여 질병 관련 단어들을 추출한다. 그리고 추출된 단어들의 검색 추이를 살펴본다면 선행 연구에서와같이 질병의 발생 예측에 활용될 수 있을 것이다.

3.3 Drug 검색 분석

약에는 두 가지의 효과가 존재한다. 목표(Target)로 하는 질병을 치료하는 본연의 효과와 원하지 하는 현상이 신체와 정신적으로 나타나는 부작용(Side effect)이 있다. 약에 대한 연구와 임상 시험(Clinical test) 과정을 통해 약의 효과와 부작용을 어느 정도 파악할 수 있지만, 많은 변수가 존재해서 완벽하게 이점을 파악 할 수는 없다. 구글 데이터를 활용하면 이러한 점을 어느 정도 보완 할 수 있을 것이다. 기존의 연구들에 대한 내용이 기록되어 있는 문헌 데이터를 분석하여, 연구로부터 밝혀진 기본적인 약의 효능과 부작용을 파악할 수 있고, 이외에 발생할 수 있는 효과들은 구글 데이터를 활용하여 추가로 분석할 수 있다. 예를 들어, 특정한 약을 처방받은 환자의 구글 검색 기록 중 고열이나 기침과 같은 증상(Symptom) 관련 단어들이 존재한다면 이러한 증상들은 해당 약으로부터 발생하는 추가적인 증상이라고 판단될 수 있다. 이렇게 축적된 데이터를 분석하면 연구와 임상 시험으로부터 알려진 효과 이외에 추가적으로 발생할 수 있는 효과까지 파악할 수 있을 것이다. 약을 처방받은 환자들의 구글 검색데이터는 또 다른 임상 시험의 데이터가 될 수 있기 때문이다.

4. 결 론

본문을 통하여 기존의 바이오 텍스트 마이닝에 대한 간략한 소개와 한계점을 살펴보았다. 그리고 본 원고에서는 한계점을 극복하기 위한 방법으로 구글 데이터를 제시하였다. 그러나 구글 데이터는 비정형화된(Unstructured) 데이터이고 데이터의 생성이 전문가와 비전문가를 포함하는 모든 사용자로부터 이루어진다는 점에 있어서 고려할 사항이 많다. 하지만 구글 데이터는 방대한 양의 정보를 포함하고 있으며, 그 종류도 다양해 생물학 분야에서의 활용 가능성도 충분하다고 생각한다. 구글 데이터를 활용한 인플루엔자 발병 예측, 구글 검색 결과를 활용한 질병 관련 유전자 추출에 대한 연구, 구글 스칼라와 PubMed의 비교에 대한 연구들을 살펴보면서 실제로 생물학 분야에 있어서 구글 데이터의 활용 가능성을 확인하였다. 추가로 3장에서는 기존의 문헌 데이터를 활용하는 바이오 텍스트 마이닝과 구글 데이터를 함께 사용하여 연구할 수 있는 주제에 대해서 간략하게 기술하였다. 구글 스칼라와 PubMed와의 결합을 통해 더 많은 생물학 문헌을 수집하는 방안과 생물학 문헌 데이터와 구글 검색의 기록을 활용한 전염성 발병 예측 방법, 약을 처방받은 환자의 증상 관련 단어 검색을 분석하여 알려지지 않은 약의 효과 및 부작용에 대한 연구의 가능성을 간략하게 살펴보았다. 이 외에 기존의 데이터들과 구글 데이터의 조합 혹은 구글 데이터만으로도 생물학 분야와 관련된 다양한 연구들이 가능할 것으로 전망한다. 기존의 정형화된(Structured) 데이터인 microRNA, RNA-Seq, microarray와 같은 데이터를 통해서도 많은 생물학 정보들을 추출할 수 있지만, 문헌 데이터나 구글 데이터와 같은 비정형 데이터를 활용한다면 정형화된 데이터를 기반으로 하는 연구에서 확인하지 못했던 중요한 정보들을 추출할 수 있을 것

이다. 특히 기존의 방법들과 구글 데이터의 융합은 향후 질병 연구에서 있어서 주요한 주제가 될 것이다.

참고문헌

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

[2] Chen L, Liu H and Friedman C, “Gene name ambiguity of eukaryotic nomenclatures”, Bioinformatics, 21(2), 248-256, 2005

[3] Wermter L, Tomanek K, Hahn U, “High-performance gene name normalization with GENO”, Bioinformatics, 25(6), 815-821, 2009

[4] Leaman R, Dogan RI, Lu Z, “DNorm: disease name normalization with pairwise learning to rank”, Bioinformatics, 29(22), 2909-2917, 2013

[5] Adamic LA, Wilkinson D, Huberman BA and Adar E, “A literature based method for identifying gene-disease connections”, IEEE Computer Society Bioinformatics Conference, pp.109-117, 2002

[6] Al-Mubaid H. and Singh RK, “A new text mining approach for finding protein-to-disease associations”, Am J Biochem Biotechnol, 1, 145-152, 2005

[7] Chen JY, “Mining Alzheimer disease relevant proteins from integrated protein interactome data”, Pac. Symp. Biocomput, 11, 367-378, 2006

[8] Ozgur A, Vu T, Erkan G and Radev DR, “Identifying gene-disease associations using centrality on a literature mined gene-interaction network”, Bioinformatics, vol. 24, ISMB 2008, pp. i277-i285, 2008

[9] Chen H, Sharp BM, “Content-rich biological network constructed by mining pubmed abstracts”, BMC Bioinformatics, 5, 147-159, 2004

[10] Goh KI, “The human disease network”, Proc. Natl Acad. Sci. USA, 104, 8685-8690, 2007

[11] Yoon B, Park Y, “A text-mining-based patent network: Analytical tool for high-technology trend”, The Journal of High Technology Management Research, 15(1), 37-50, 2004

[12] Tseng YH, Lin CJ, Lin YI, “Text mining techniques for patent analysis”, Information Processing & Management, 43(5), 1216-1247, 2007

[13] comScore, http://www.comscore.com/

[14] Michael E Anders and Dennis P Evans, “Comparison of PubMed and Google Scholar Literature Searches”, RESPIRATORY CARE, vol. 55, no. 5, MAY 2010

[15] Herman Anthony Carneiro and Eleftherios Mylonakis, “Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks”, Clinical Infectious Diseases, vol. 49, no. 10, 2009

[16] Jeongwoo Kim, Hyunjin Kim, Youngmi Yoon and Sanghyun Park, “LGscore: A method to identify disease-related genes using biological literature and Google data”, Journal of Biomedical Informatics, 2015

[17] Mary Shultz, MS, AHIP, “Comparing test searches in PubMed and Google Scholar”, Journal of the Medical Library Association, vol. 95, no. 4, 2007

약 력

박 상 현

1989 서울대학교 컴퓨터공학과 졸업(학사)
1991 서울대학교 대학원 컴퓨터공학과(공학석사)
2001 UCLA 대학원 컴퓨터공학과(공학박사)
1991~1996 대우통신 연구원
2001~2002 IBM T. J. Watson Research Center Post-Doctoral Fellow
2002~2003 포항공과대학교 컴퓨터공학과 조교수
2003~2006 연세대학교 컴퓨터과학과 조교수
2006~20011 연세대학교 컴퓨터과학과 부교수
2011~현재 연세대학교 컴퓨터과학과 교수
관심분야: 데이터베이스, 데이터마이닝, 바이오인포매틱스, 적응적 저장장치 시스템, 플래쉬메모리, 인덱스, SSD
Email: sanghyun@cs.yonsei.ac.kr

김 정 우

2013 상명대학교 컴퓨터과학과 졸업(학사)
2013~현재 연세대학교 컴퓨터과학과 통합과정
관심분야: 바이오인포매틱스, 데이터마이닝, 텍스트마이닝, 데이터베이스
Email: jwkim2013@cs.yonsei.ac.kr