SVM 과 PubMed 를 이용한 추가적인 생물학 텍스트 데이터 확보 방법

차준범, 김정우, 박상현*
연세대학교 컴퓨터과학과
e-mail : khanrc@yonsei.ac.kr

A method to extract additional biomedical text data using Support Vector Machine and PubMed

JunBeom Cha, JungWoo Kim, Sanghyun Park*
Dept. of Computer Science, Yonsei University

요 약

생물학 분야에서의 텍스트 마이닝(text mining) 분야가 급격하게 성장하면서, 텍스트 마이닝의 핵심 리소스인 텍스트 데이터(text data)의 중요성도 함께 증가하고 있다. 대부분의 텍스트 데이터는 PubMed 의 MeSH(Medical Subjects Headings) term 검색 결과를 사용해왔는데, 이 과정에서 MeSH 분류에는 포함되지 않지만 충분히 가치있는 데이터들을 놓치게 된다. 본 논문에서는 풍부한 텍스트 데이터를 확보하기 위해, 기존의 MeSH term 검색을 사용한 텍스트 데이터 외에 추가적인 텍스트 데이터를 확보할 수 있는 방법을 제안한다.

1. 서론

텍스트 마이닝은 자연언어로 된 문서를 분석하여 사용자가 원하는 정보를 선별하고, 그 결과를 정제되고 가공된 형태로 제시하는 것이다. 1980 년대에 처음 소개되어, 1990 년대에 접어들며 급격하게 발전하기 시작했다[1][2]. 텍스트 마이닝의 발전에 따라 생물학적 문헌에 대한 연구도 같이 진행되었다[3]. 이뿐만 아니라 1990 년부터 진행된 인체 유전연구 프로젝트(Human Genome Project)는 유전자에 대한 다양한 연구를 가능케 했으며, 이 프로젝트의 가시적인 결과가 나타나기 시작한 1995 년 경부터 방대한 자료를 다루는 분자생물학과 전산학의 결합인 생물정보학(Bioinformatics)이라는 분야가 성장하기 시작하였다[4].

이러한 연구의 발전에 따라, 생물학 분야에서의 텍스트 마이닝은 매년 급격하게 성장하고 있다. 그림 1 에서 생명과학(life science) 과 생의학(biomedicine)에 대한 데이터베이스 검색 엔진인 PubMed[10]에서 “text mining” 또는 “literature mining”으로 검색한 결과가 매년 상승하는 것을 볼 수 있다.

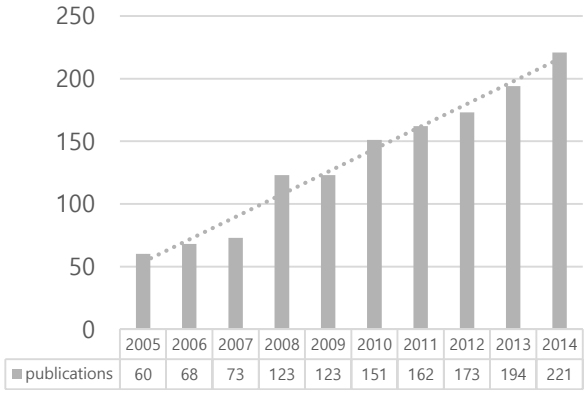

그림 1 최근 10 년간 생물학 텍스트 마이닝 관련 문헌 수의 변동 추이

이와 같이 바이오 분야에서 텍스트 마이닝이 성장함에 따라, 그 핵심 리소스인 바이오 텍스트 데이터의 중요성도 증가하고 있다. 일반적으로 바이오 텍스트 데이터의 확보는 PubMed 의 MEDLINE(Medical Literature Analysis and Retrieval System Online) 데이터베이스의 MeSH term 검색을 사용한다. 하지만 이러한 방법은, MeSH term 에는 포함되지 않지만 충분히 가치 있는 문헌들을 놓치게 되는 문제가 있다. 이러한 문제점을 해결하기 위해, 본 논문에서는 키워드 검색(keyword search) 결과를 텍스트 마이닝을 통해 가치 있는 데이터만을 찾아, 기존의 방법보다 더 많은 데이터를 확보할 수 있는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2 장에서는 바이오 텍스트 마이닝 및 SVM(Support Vector machine)과 관련한 기존의 연구들을 살펴본다. 3 장에서는 키워드 검색 결과로부터 가치 있는 문헌들을 추출하는 방법론을 제안한

* : 교신저자, e-mail: sanghyun@cs.yonsei.ac.kr
※ 이 논문은 2015 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2015R1A2A1A05001845).