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.

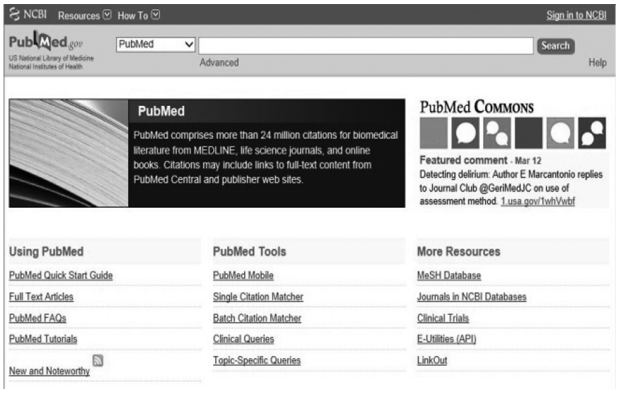

그림 1 생물학 문헌 데이터베이스 PubMed

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).