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.

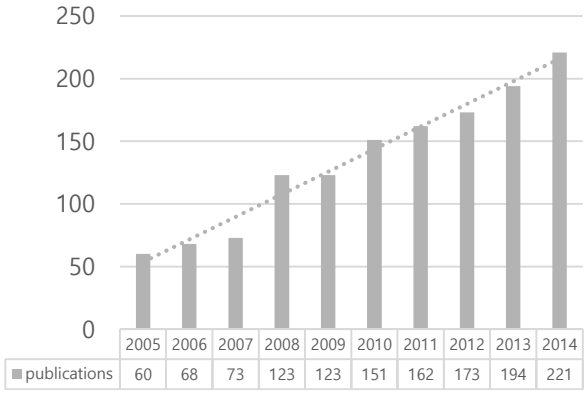

그림 1 최근 10 년간 생물학 텍스트 마이닝 관련 문헌 수의 변동 추이

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).

다. 끝으로 4 장에서는 본 연구의 결론과 발전방향에 대해서 기술한다.

2. 관련연구

본 연구에서 사용할 기계학습(Machine Learning) 방법론인 SVM(Support Vector Machine)은 V. N. Vapnik 등이 1963 년에 처음으로 그 개념을 소개하였으며, 1992 년에 B. E. Boser, I. M. Guyon 그리고 V. N. Vapnik 에 의해 첫 논문이 발표되었다[5]. 이후 1995 년에 과적합(overfitting)을 피할 수 있어 현재 널리 사용되는 소프트 마진(soft margin) 개념이 소개되었다[6]. SVM 은 기존의 알고리즘들과 달리 일반화를 고려한 분류기를 생성하여 높은 정확도를 보여준다. 이에 따라 다양한 연구에 활용되기 시작했으며 텍스트 마이닝 분야에도 성공적으로 적용되었다[7][8][9].

SVM 은 기계학습 알고리즘 중 지도학습(supervised learning) 에 속하는 알고리즘으로, MMH(maximal marginal hyperplane)를 구하는 점에서 다른 알고리즘들과 차별화된다. 다른 기계학습 알고리즘들이 단순히 트레이닝 데이터에서 분류 경계(decision boundary)를 구하는 것에 그치는 반면 SVM 은 가장 일반화된 분류 경계를 의미하는 MMH 를 구한다. 즉, 더욱 일반화된 분류기(classifier)의 모델링이 가능하며 따라서 더욱 좋은 분류(classification) 결과를 보여준다.

문헌데이터들은 생물학적으로 중요한 정보를 추출하는데 활용될 수 있다. Iacucci[17]는 생물학 텍스트 데이터를 분석하여 수용체(receptor) - 리간드(ligand) 쌍에 우선순위를 설정하는 방법론을 제안하였고, Adamic[19]은 텍스트 분석을 통해 질병과 유전자 사이의 관계를 추출하는 방법론을 제안하였다. 이외에도 생물학 문헌을 정확하고 유용하게 활용할 수 있도록, 생물학 문헌 데이터에 AZ(Argumentative Zoning) 주석을 다는 연구[18]등 생물학 문헌 데이터를 활용하는 다양한 연구들이 진행되고 있다.

3. 방법

3.1. 데이터 확보

데이터의 확보는 데이터 마이닝(data mining)의 수행에 있어서 가장 중요한 선결과제 중 하나다. 본 연구에 필요한 데이터는 생물학 분야에서 특정 주제에 대해, 그 주제에 대한 정보를 담고 있는 문헌과 그렇지 않은 문헌들이다. 해당 정보를 담고 있는 문헌은, 우리가 찾고자 하는 대상으로서 긍정 데이터(positive data)에 속한다. 반면, 정보를 담고 있지 않은 문헌은 우리가 걸러내고자 하는 대상으로서 부정 데이터(negative data)에 속한다. 분류기(classifier)의 학습을 위해, 두 종류의 데이터가 모두 필요하다. 다음 세 가지 PubMed 검색 결과를 사용하여 이 데이터를 얻는다.

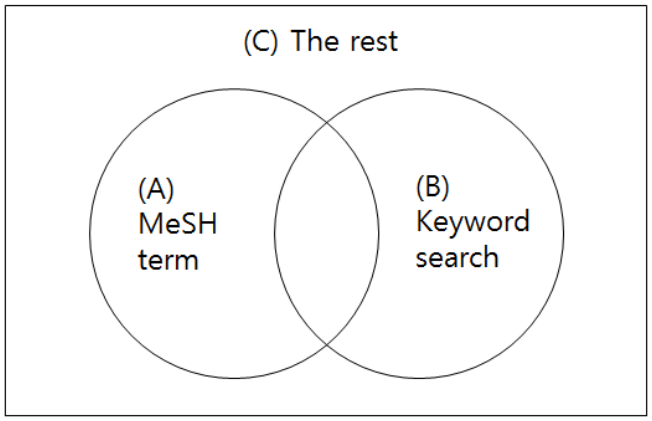

그림 2 PubMed 검색 결과 분류

(A) MeSH term 검색 결과. MeSH 는 미국 국립의학도서관(NLM[11])이 정하는 주제명표목으로, 각 문헌마다 문헌의 내용을 나타내는 적절한 10~15 개의 용어가 부여된다. 즉, MeSH term 검색으로 얻은 데이터는 미국 국립의학도서관에서 검증한 관련문헌이라고 할 수 있어, 긍정 데이터에 적합하다.

(C) MeSH term 검색과 키워드 검색 결과를 제외한 나머지. MeSH term 에도 포함되지 않으면서, 키워드 검색 결과에도 포함되지 않는 즉 전혀 키워드와 관련없는 문헌들에 해당하므로 부정 데이터에 적합하다.

(B)-(A) MeSH term 검색 결과를 제외한 키워드 검색 결과. 관련 문헌일 수도, 아닐 수도 있는 나머지 검색 결과에 해당한다. 우리가 분류기를 적용하여 분류해내고자 하는 데이터다.

3.2. 특성 선택

데이터 마이닝의 두 번째 단계는 데이터 전처리(data preprocessing)이고, 그 중에서 가장 중요한 작업이 특성 선택(feature selection)이다. 이 특성 선택은 특성이 한없이 많아질 수 있는 텍스트 마이닝에서 더더욱 중요하다.

먼저 각 문헌을 The C&C Tool[12]과 GENIA corpus[13]를 사용해서 품사 분석(POS tagging & parsing)을 수행한다. 이렇게 구분한 품사들 중 명사만을 사용하여, TF-IDF(Term Frequency – Inverse Document Frequency) 분석을 수행한다. TF-IDF 분석은 각 문헌에서 등장하는 단어들의 빈도와, 전체 문헌에서 해당 단어가 등장하는 문헌의 빈도를 분석하여 각 문헌 고유의 특징적인 단어들을 추출할 수 있는 방법론이다. 즉, 이렇게 추출한 단어들이 각 문헌을 나타내는 고유값 이므로 본 연구의 특성으로 사용한다.

이후 각 문헌들의 특성 단어를 바탕으로, SVM 의 학습을 수행한다. 학습 데이터는 긍정 데이터의 특성 단어와 부정 데이터의 특성 단어로 구성된다. 본 논문에서는 다양한 SVM 구현체 중 LIBSVM[14]을 사용한다.

3.3. 평가

데이터 마이닝의 최종 단계는 분류기의 평가다. 즉, 우리가 학습시킨 분류기를 테스트하여, 어느 정도의 성

능을 보이는지 평가한다.

평가를 위해 정답 데이터(gold standard)가 필요하다. 해당 주제에 대한 전문가에게 의뢰하여 이 데이터를 확보한다. 의뢰하는 데이터는 분류기로 분류하고자 하는 문헌들로, 그림 2 에서 (B)-(A)에 해당한다. 전문가는 각 문헌을 읽고, 이 문헌이 해당 주제와 관련되어 있는 문헌인지 아닌지를 판별하여 기록한다. 이 데이터를 기반으로 분류기의 성능을 평가한다.

텍스트 데이터를 확보함에 있어, 가장 중요한 것은 정확한 데이터만을 확보하는 것이다. 즉, 관련 없는 문헌을 관련 없다고 분류하는 것 보다, 관련 있는 문헌을 관련 있다고 분류하는 것이 더욱 중요하다. 이를 제대로 평가하기 위해, F-score[15]를 사용한다. F-score 는 이진 분류(binary classification)에 대한 통계적인 분석법으로, precision 과 recall[16]을 사용하여 테스트의 정확도를 측정한다.

4. 결론 및 발전방향

본 연구를 통해 충분히 가치 있음에도 불구하고 사용하기 어려웠던 텍스트 데이터들을 찾아낼 수 있다. 텍스트 데이터는 텍스트 마이닝의 핵심 리소스로, 이후 생물학 분야의 텍스트 마이닝 연구에 널리 활용할 수 있을 것으로 기대한다.

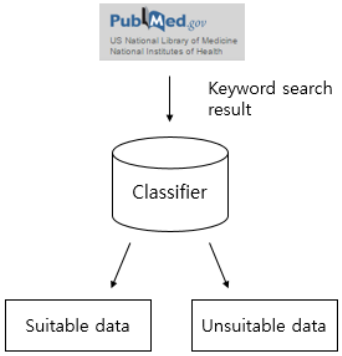

그림 3 최종 결과 모형

차후 연구에서는 본 논문의 제안을 실험 및 개선한다. 먼저 특정 질병에 대해 분류기를 학습하고, 전문가에게 의뢰하여 정답 데이터를 확보한다. 이후 이 데이터로 분류기를 평가하여 분류기의 학습 방법을 개선한다. 최종적으로는 어플리케이션을 만들어 배포하여, 바이오 텍스트 마이닝을 연구하는 연구자들이 기존보다 더 많은 텍스트 데이터를 활용할 수 있도록 제공하는 것을 목표로 한다. 그림 3 과 같이, 키워드 검색 결과를 분류기를 통해 적합한 데이터만을 추출하여 기존보다 더 풍부한 텍스트 데이터를 활용할 수 있다.

참고문헌

[1] MW Berry, ST Dumais, GW O'Brien, USING LINEAR ALGEBRA FOR INTELLIGENT INFORMATION RETRIEVAL. Vol. 37, No. 4, Society for Industrial and Applied Mathematics, pp. 573-595, 1995

[2] G Salton, C Buckley, TERM-WEIGHTING APPROACHES IN AUTOMATIC TEXT RETRIEVAL. Information Processing & Management, Vol. 24, No. 5, pp. 513-523, 1988

[3] G Salton, ANOTHER LOOK AT AUTOMATIC TEXT-RETRIEVAL SYSTEMS. Communications of the ACM, Vol 29, No. 7, pp. 648-656, 1986

[4] 김승목, Genome Project 와 Bioinformatics. The Korean Journal of Microbiology, Vol. 34, No. 1-2, pp. 1-5, 1998

[5] BE Boser, IM Guyon, VN Vapnik, A training algorithm for optimal margin classifiers, Computational learning theory, pp. 144-152, 1992

[6] C Cortes, V Vapnik, Support-Vector Networks. Machine Learning, Vol. 20, No. 3, pp. 273-297, 1995

[7] T. Joachims, Text categorization with Support Vector Machines: Learning with many relevant features. European Conf. on Machine Learning, Vol. 1398, pp. 137-142, 1998

[8] S Dumais, J Platt, D Heckerman, M Sahami, Inductive learning algorithms and representations for text categorization. Conference on Information and Knowledge Managment, pp. 148-155, 1998

[9] E Leopold, J Kindermann, Text categorization with support vector machines. How to represent texts in input space?. Machine Learning, Vol. 46, No. 1-3, 423-444, 2002

[10] PubMed, http://www.ncbi.nlm.nih.gov/pubmed

[11] NLM, http://www.nlm.nih.gov/

[12] JR Curran, S Clark, J Bos, Linguistically Motivated Large-Scale NLP with C&C and Boxer. ACL, pp. 33-36, 2007

[13] JD Kim, T Ohta, Y Tateisi, J Tsujii, GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics, Vol. 19, No. 1, pp. 180-182, 2003

[14] CC Chang, CJ Lin, LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, 2011

[15] M Sokolova, N Japkowicz, S Szpakowicz, Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. AI 2006: Advances in Artificial Intelligence(Book), Vol. 4304, pp. 1015-1021, 2006

[16] DM Powers, Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies, Vol. 2, No. 1, pp. 37-63, 2011

[17] E Iacucci, LC Tranchevent, D Popovic, GA Pavlopoulos, BD Moor, R Schneider, Y Moreau, ReLiance: a machine learning and literature-based prioritization of receptor—ligand pairings. Bioinformatics, Vol. 28, No. 18, pp. 569-574, 2012

[18] Y Guo, I Silins, U Stenius, A Korhonen, Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review. Bioinformatics, Vol. 29, No. 11, pp. 1440-1447, 2013

[19] LA Adamic, D Wilkinson, BA Huberman, E Adar, A literature based method for identifying gene-disease connections. IEEE Computer Society Bioinformatics Conference, pp. 109-107, 2002