J Korean Med Assoc 2012 August; 55(8): 741-747
http://dx.doi.org/10.5124/jkma.2012.55.8.741
pISSN: 1975-8456 eISSN: 2093-5951
http://jkma.org

임상근거지식 추출을 위한 메디컬 인포매틱스 기법

송 미 화[1] · 박 동 균[2] · 이 영 호[3*] | [1]가천대학교 유헬스케어연구소, [2]가천대 길병원 유헬스케어센터, [3]가천대학교 IT대학

Medical informatics methods for the clinical evidence extraction

Mi Hwa Song, PhD[1] · Dong Kyun Park, MD[2] · Young Ho Lee, PhD[3*]

[1]U-Healthcare Institute, Gachon University, [2]U-Healthcare Center, Gachon University Gil Hospital, [3]IT Department, Gachon University, Incheon, Korea

***Corresponding author:** Young Ho Lee, E-mail: lyh@gachon.ac.kr

Received June 29, 2012 · Accepted July 12, 2012

Clinical professionals gain new information to assist in patient care when they read the medical literature. Similarly, in clinical preventive medicine, medical science documents that have previously published can be searched and evaluated in order to confirm the scientific support for the clinical preventive medical service offered in order to prevent chronic disease. This paper introduces the medical informatics techniques for knowledge extraction that can become the basis for clinical practice. Particularly, it discusses the clinical document retrieval and knowledge discovery tools that can search for extracting the knowledge which the medical expert desires with data mining techniques. For example, Clinical medical personnel and medical researchers can locate the information from the latest literature rapidly or find and evaluate the scientific basis for the treatment and prevention of infection. This study can be used when they analyze the correlation between accumulated and different type of data and contributes to the detection of new knowledge. Recently, the concern about the visualization of massive data and information is high as the importance of big data has received greater attention. Contributions to this technique and decision support tools will increase gradually due to the way support for decision-making through scientific evidence for the pattern changing disease is evaluated or as one of the clinical practice guidelines is accepted.

Keywords: Evidence-based medicine; Medical informatics computing; Decision support techniques; Information storage and retrieval; Artificial intelligence

서 론

임상의가 의학문헌을 읽을 때는 환자진료에 도움이 되는 새로운 정보를 얻음과 동시에 그 타당성에 대한 평가를 하게 된다. 마찬가지로 임상예방의학에서도 만성질환을 예방하기 위하여 제공하는 임상예방의료서비스의 과학적 증거를 확인하기 위하여 기존에 발표된 의학문헌을 검색하고 평가하게 된다. 한편 임상 실무 가이드라인(clinical practice guideline)은 의료현장에서 적절한 질병관리 방법을 결정하는 데 있어 효과적인 도구이다. 의료진의 판단을 돕기 위해

체계적인 과정을 제시하고, 의사의 진료 판단과 과학적 근거의 간격을 최소화한다[1]. 임상 가이드라인 모델링 서비스는 임상 진료 프로세스(알고리즘)를 시각화 툴을 통해 저작, 추론 엔진에 의해 실행이 가능한 형식으로 저장한 것이다[2]. 이는 현재 환자의 건강상태나 행동변화와 같은 상황정보(dynamic context)에 최적화 하거나 새로운 의학적 연구에 따라 업데이트하는 것을 의미한다. 지식 모델링 및 수정 과정에서는 수많은 근거 자료에 대한 참조 작업이 발생한다. 그러므로 근거 기반의 지식 수집, 추출 및 관리 기능이 필수적이다. 이 글에서는 임상적 행위의 근거가 되는 지식 추출을 위한 메디컬 인포매틱스 기법들에 대해 소개한다. 특히 데이터 마이닝 기법을 통해 임상문헌으로부터 의료 현장의 전문가가 원하는 지식을 추출하는 방법을 살펴보고 의료인에게 도움을 줄 수 있는 임상 문헌 검색과 지식 발견 도구를 소개하고자 한다.

기계학습 기법

의료 분야에서 활용되는 데이터 마이닝 기법은 maximum entropy, 베이지안 네트워크, 지지벡터기계(support vector machine, SVM), 신경망, 의사결정나무, 회귀분석 등 매우 다양하다. 다음에서 이 중 일부를 간략히 살펴본다. Maximum entropy 모델은 불충분한 정보에 기초하여 추론을 할 경우, 현재 알려진 사실에 어긋나지 않는 것 중 entropy가 가장 높은 확률 분포를 사용한다[3]. 최대 entropy 모델을 유도하기 위해서 데이터에서 특정 컨텍스트와 함께 발생하는 인스턴스 클래스의 확률을 추정하기 위한 파편적 증거(evidence)들을 결합한다. 이 때 조건부 확률 p를 사용하는데, 이를 획득하기 위해서 데이터로부터 증거들을 수집하는 특징 함수(feature function)가 필요하다. 최대 entropy에서 특징 추출 함수는 일반적으로 표시함수(indicator function)로서 Boolean 값(0,1)을 출력한다.

베이지안 네트워크는 변수 집합 사이의 확률적인 관계를 네트워크 형태로 표현하는 방법이다[4]. 일반적으로 다른 노드들의 확률값들을 기초로 특정 노드가 가질 값에 대한 조건부 확률을 계산하는데 이용할 수 있다. 심장병과 연관된 위험인자를 갖고 있는 환자를 모델링하기 위해 베이지안 네트워크를 이용하여 질병 분류를 수행한 연구는 잘 알려져 있다[4,5]. 또한 text analysis와 decision support systems에서 지식을 모델링 하는데 사용된다[6]. 다층 퍼셉트론 신경망의 모델의 하나로, 비선형 활성화 함수(activation function)의 집합으로 구성된 은닉 노드 계층을 갖고 있다. 신경망은 입력 벡터에 국부적으로 존재하는 특징 추출 상의 에러, 은닉변수, 이상치(outlier)로 인한 노이즈에 강건(robust)하다. 그렇기 때문에 문장 범주 분류와 같이 불확실한 요소가 많은 도메인 문제에 많이 응용되고 있다. 한편 Cho 등[7]은 암을 분류하기 위한 기계학습 모델로 다층퍼셉트론을 이용한 연구를 수행하였다. 원형 기반 함수 네트워크(radial basis function network)도 다층 퍼셉트론과 마찬가지로 신경망 모델의 하나이다. 원형 기반 함수는 n차원 가중치 공간(weight space)에 존재하면서 입력 벡터와의 유클리드 거리에 따른 최적 가중치 벡터를 출력 노드에 선형 결합(linear combination)하기 위한 은닉 노드 계층 요소이다. 또한 입력 시그널의 값에 따른 가중치를 출력하게 되어 보다 비선형적인 패턴들 간의 유사성이 출력 뉴런에서 반영되도록 한다[8].

나이브 베이지안은 텍스트 인스턴스가 특정 클래스 에 속할 확률을 베이즈 법칙(Bayes' theorem)을 근거로 추정한다. 즉, 문장를 표현하는 각 단어들 간에 조건부 독립성(conditional independence)이 있다고 가정하는 것이다. 나이브 베이지안은 노이즈에 강하고 문서의 양이 많은 경우에도 매우 높은 정확률을 갖고 있는 것으로 보고되고 있다[9].

SVM에는 서로 다른 클래스에 속하는 n차원 데이터 포인트 집합을 분리하는 결정경계(hyperplane)가 존재한다. 이 결정경계간 마진(margin, hyperplane 사이의 거리) 폭이 최대화될 때 이 결정경계에 최-근접한 데이터 포인트를 지지 벡터(support vector)라고 한다. 지지벡터기계(SVM)는 이 결정 경계에 대한 최적 가설을 뜻한다[3,10]. SVM에서는 기본적으로 훈련 데이터가 2개의 클래스에 속하는 것으로 간주된다. 이 때문에 3개 이상의 클래스 분류에서는 다중의 지지 벡터에 대한 최적 가설을 선택해야 하는 단점이 있다.

임상문헌의 문장단위 분류

임상문헌으로부터 의료전문가가 원하는 정보를 추출하는 방법에 대한 다수의 연구가 진행되고 있다. 여기서 정보의 추출이란 비구조화(unstructured)된 문서로부터 구조화(structured)된 정보를 추출, 재구성하는 문제로 정의되는데, 문제의 특성이나 접근하는 방법에 따라 정보 검출 및 정보 분류로 구분된다[11]. 이 글에서는 추출 대상 정보에 대한 객체(instance)를 정의하고, 각 객체 정보의 유형을 분류하는 것을 목적으로 하는 분류 문제를 중심으로 소개하기로 한다.

여기에서는 PubMed 및 Google을 포함한 검색엔진에서 제공하는 검색 결과에 포함된 문서가 문장요소의 집합으로 간주될 수 있다. Kim 등[12]은 근거기반의학(evidence based medicine)을 위한 자동 문장 분류 연구를 진행하였다. 이 연구에서는 의학 논문의 초록을 대상으로 background, population, intervention, outcome, study design, other의 다양한 의미 태그를 정의하여 사용하였다. 논문의 초록은 배경, 본문, 결과 등의 순서로 문장이 나열되어 있다는 특징이 있다. 이러한 순차적인 데이터를 학습하는 데 유용한 conditional random fields를 이용해 분류를 수행하였는데, 문장 데이터 내의 어휘 정보, 의미 정보, 구조 정보, 순차적 정보를 기반으로 한 다양한 특징(feature)을 사용하여 성능평가를 한 것이 특징이다. 또한 Nawaz 등[13]은 bioevent를 분류하기 위해 다중 태깅 연구를 진행하였다. 이 연구에서는 knowledge type, manner, certainty level, logical type, source, lexical polarity로 태그를 나누었다. 그리고 그 중 지식의 유형을 investigation, observation, analysis, general로 나누었다. 지식의 유형을 나누는 방법은 lexical clue라고 불리는 지식의 유형을 대표할 만한 단어들을 통해 전문가와 협의하여 태깅하게 된다. Pan [14]은 단일 문장에 대한 다중 태깅(multi-label tagging)을 통한 문장 분류 연구를 진행하였다. 이 연구에서 의학/생명 분야 논문들을 구성하는 문장 또는 절(clause)을 focus, polarity, certainty, evidence, direction or trend와 같은 의미 클래스가 복합적으로 결합하여 생성된 텍스트 인스턴스로 간주한다. 개별 훈련 알고리즘간의 성능을 비교하여 알고리즘을 평가하였다. Song 등[15]은 문장 인스턴스 특징 공간(feature space)의 차원을 줄이기 위해 변형함수(transformation)를 사용하였는데, 변형함수란 차원의 저주(curse of dimensionality) [16]에 대한 차원 감소(dimensionality reduction)를 목적으로 사용되는 특징 추출(feature extraction) 함수이다. 이 특징 추출 함수를 통해 텍스트의 어휘적, 통사론적, 동시 발생 이벤트(co-occurrence event)와 같은 특징을 정량화하며, 그 결과를 특징 벡터로 표현하는 기능을 포함한다.

특징 선택의 개념과 적용

특징 선택(feature selection)이란 기계학습 분야에서 매우 중요한 주제 중 하나로, 원본 데이터로부터 분류의 정확도를 높일 수 있는 데이터의 부분 집합(subset)을 선택하는 과정이다[17]. 분류의 목적에 가장 연관되어 있는 특징들만을 추출함으로써 원본 데이터에 비해 줄어든 데이터를 얻을 수 있으며, 분류의 기준이 되기에 기여도가 떨어지는 중복(redundant) 데이터, 잡음(noise)데이터를 제거할 수 있다. 이러한 과정을 통해 보다 빠른 연산 시간과 더 정확한 분류 성능을 기대할 수 있다. 특징 선택 방법은 특징 부분집합의 생성 방법에 따라 크게 두 가지 유형으로 나뉜다[18]. 첫 번째로, filter는 특징에서 부분집합을 먼저 선택하여 이를 분류 알고리즘을 실행하는데 사용하는 것이다. 선택된 부분집합이 얼마나 좋은가에 대한 평가는 부분집합에 속한 특징과 분류 기준 사이의 고유 속성들을 이용하여 평가하게 된다. Filter의 예로 정보이득(information gain)을 들 수 있다. 이는 기계학습 분야에서 용어의 적합도 기준으로 자주 사용되며, 문서에서의 출현 빈도뿐 아니라 출현하지 않은 빈도까지 고려하여 각 범주에서의 용어 정보량을 계산한다[19]. 두 번째로, wrapper는 최적의 특징을 찾기 위해 분류 알고리즘을 dataset에 적용시키는 것이다. Filter 방식과 달리, 직접 분류기를 사용하여 해당하는 특징 부분집합의 우수성을 평가하며, 사용할 분류기를 미리 결정한 후 매번 특징 부분집합이 생성될 때 마다 분류하고 그 분류 성능이

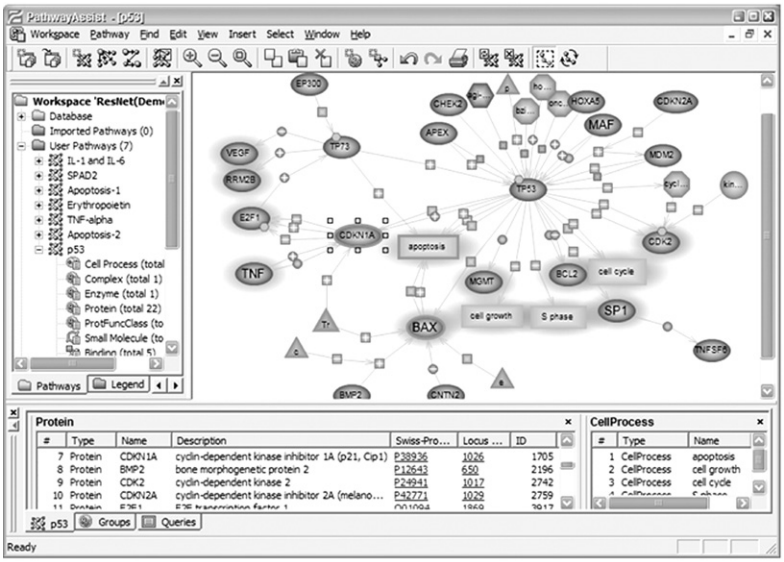

Figure 1. PathwayAssist graphical interface. Pathway assist enables researchers to create their own pathways and produce publication quality pathway diagrams. PathwayAssist uses a proprietary graph visualization engine to allow for the visual features (From Nikitin A, et al. Bioinformatics 2003;19:2155-2157, with permission from Oxford University Press) [22].

우수함의 척도가 된다. Wrapper 방법은 feature의 수가 많을 때 시간이 많이 걸린다. Wrapper의 대표적인 예는 유전자 알고리즘(genetic algorithm, GA)이다. 이는 최적화 문제에서 후보가 되는 해들 중 개체군이 더 좋은 해를 찾아서 진화하는 것으로, Silla 등[20]의 연구 등에서 텍스트 자동 요약을 위한 특징 선택을 하기 위해 GA를 채택하였다. 또한 Anbarasi 등[21]은 심장병을 예측 하기 위한 변수를 선정하는 과정에서 GA를 활용하였고, 서로 다른 세 개의 기계학습 이용하여 성능을 비교, 평가하였다.

의학문헌 검색 및 지식 발견 도구의 활용

현재 National Center for Biotechnology Information의 PubMed에는 18,000,000건 이상의 문헌들이 수록되어 있고 하루에도 수십 편에서 수천 건의 논문들이 업데이트 되고 있다. 최근 의료정보학, 생물정보학 분야의 이슈는 현재까지 공개 데이터베이스에 축적된 데이터와 자신이 보유하고 있는 데이터를 활용하여 새로운 정보의 의미를 발견하는 부분에 있다. 따라서 다양한 분야의 실험데이터, 문헌데이터, 공개된 DB 등을 서로 연결하여 새로운 지식을 발굴할 수 있는 시스템이 주목 받고 있다. 특히 이러한 다양한 데이터와 인터넷 문서 등 비정형 데이터로부터 자연언어처리 및 문서처리기술을 적용하여 유용한 정보를 추출하고 가공하는 기술을 텍스트 마이닝이라고 한다.

1. 바이오 텍스트 마이닝 도구

Ariadne사의 MedScan과 Pathway Studio는 텍스트 마이닝을 기법을 이용하여 주어진 문헌정보에서 유전자와 질병, 화학물질, 세포 내 프로세스, 대사회로와 같은 엔티티(entity)들의 관계를 자동으로 추출하여 테이블과 다양한 그래프로 관계들을 보여주는 프로그램이다(Figure 1) [22]. Pathway Studio는 척추동물, 식물 연구의 생물학적 연관관계, ontology와 pathway들의 정보를 포함하고 있는 ResNet 데이터베이스와 자연언어처리기술을 이용하여 과학문헌을 자동으로 읽고 생물학적인 관계를 추출하는 기능을 가진 MedScan으로 구성되어 있다.

MedScan의 경우에는 약 1천 개의 논문 초록을 대상으로 생물학적인 관계를 추출하는데 2-3분밖에 걸리지 않으므로, 대량의 수집된 논문에서 특정한 바이오 마커를 발굴하거나 특정 단백질 또는 질병과 관련된 네트워크 정보를 검토하기에는 상당히 유용하다고 할 수 있다. 보통 하나의 유전자와 관계하는 다양한 정보를 찾아보기 위해서는 수많은 데이터베이스와 문헌, 웹사이트를 검색하여 그 연관관계를 하나씩 도출해야 되지만, Pathway Studio와 같은 프로그램은 그와 같은 일련의 시간과 노동력이 상당히 투자되어야 하는 업무를 효율적으로 지원함으로써 연구자의 보다 빠르고 충실한 결과물을 얻을 수 있도록 지원한다.

한편 생물, 의학 분야의 지식이나 임상데이터는 임상 실험 데이터나 처방전 데이터와 같이 코드화된 형태로 존재하

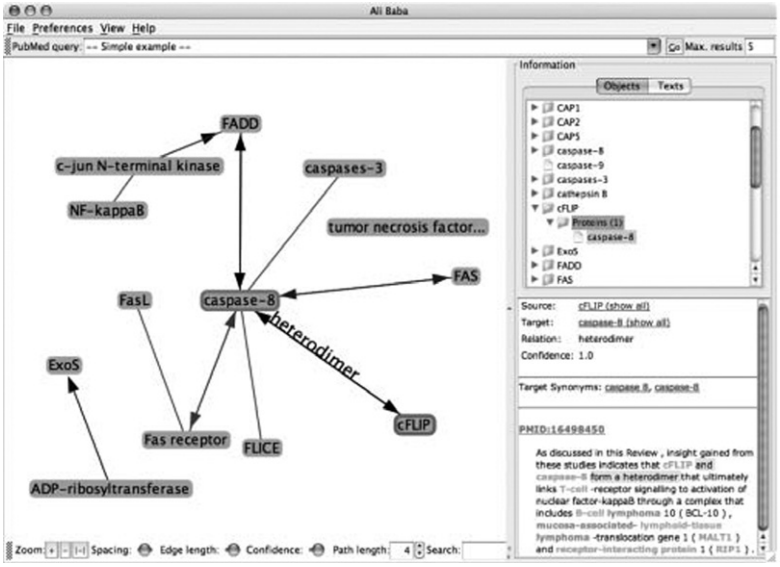

Figure 2. AliBaba Graphical Interface. (Parts of the) graph resulting from five PubMed abstracts for the query 'FADD'. Information on the selected protein 'caspase-8' is given in the right panel, for instance, association partners and evidence texts (From Plake C, et al. Bioinformatics 2006;22:2444-2445, with permission from Oxford University Press) [25].

거나, 퇴원요약지, 검사 소견과 같은 자유 텍스트(free text) 형태로 존재하는데, 대부분이 구조화되지 않은 텍스트 형식으로 되어있어, 컴퓨터를 사용해 논리적이고 일관된 방법으로 처리하기가 어렵다. 이러한 문제를 처리하기 위하여, 텍스트로부터 정보를 얻는 많은 시스템들은 문서에 있는 단어나 구의 개념을 식별해 주는 툴을 사용한다. 이러한 의학 개념 식별(medical concept identification)의 과정은 비구조화된 텍스트를 구조화된 표현으로 변환하는데 있어서 중요한 과정이며, 정보 검색을 용이하게 하기 위해 필요한 과정이다[23]. National Library of Medicine에서 개발한 MetaMap은 비구조화된 텍스트에서 통합개념체계인 Unified Medical Language System 메타시소러스 개념를 찾아주는 자연어 처리 기반의 툴로서 비구조화된 텍스트에 있는 용어를 표준화된 개념으로 매핑하는데 사용되고 있는 대표적인 프로그램이다[23].

2. PubMed 검색과 시각화

PubMed 검색결과를 그래프 또는 다이어그램으로 시각화하여 보여주는 다양한 PubMed third party tool이 존재한다. Pub anatomy는 미국 미시건 대 National Center for Integrative Bioinformatics에서 개발한 것으로 brain anatomy와 유전체 표현형의 correlation에 대한 문헌을 찾아 결과를 보여준다[24]. Ali Baba는 베를린의 홈볼트 대학에서 개발되었으며 relation-ship과 association에 대한 검색 결과를 시각적으로 요약하여 줌으로써 사용자에게 직관적인 판단의 근거를 제공한다(Figure 2) [25].

컬럼비아 대학의 Stoyanovich 등[26]은 PubMed를 통해 검색된 결과의 랭킹에 MeSH 어휘를 사용하여 의미(sementic) 정보를 활용하는 새로운 랭킹 알고리즘을 제안하였다. 또한 2차원의 skyline-based 시각화 방식을 사용하여 사용자 경험(user experi-ence)을 개선하는 방식으로 랭킹 알고리즘과 결합하는 방식을 제안하였다. 한편 국내 포항공대 연구팀에 의해 개발된 RefMed는 PubMed의 문헌 검색시 적합성 피드백(relevance feedback)을 지원하며, 초기 사용자의 질의에 의한 검색 결과를 보고 적합성을 판단하여 2차 검색을 할 수 있도록 설계, 개발되었다[27].

결 론

이 글에서는 임상에서 활용되는 여러 문헌으로부터 의미있는 정보를 추출하기 위한 다양한 기계학습 기법 및 그 적용 연구에 대해 소개하였다. 이러한 연구방법 및 결과는 임상 의료인과 의학 연구자가 신속히 최신 문헌의 정보를 확인하거나 진료 및 질병 예방을 위한 과학적 근거를 발견/평가하고, 축적되어 있는 서로 이질적인 데이터 간의 연관성 등을 분석하여 새로운 지식의 발견에 기여하는데 사용되고 있다. 특히 근래 들어 빅 데이터(big data)의 중요성이 강조

됨에 따라 대용량 데이터와 정보의 시각화에 대한 관심이 높다. 이러한 최근 연구동향과 배경기술, 지원도구를 숙지하여 적절하게 활용하는 것은 변화하는 질병의 패턴에 대한 과학적 근거를 평가하거나 진료행위에 도입하는 등 의사결정을 위한 지원방법으로 그 기여도가 점차 증가할 것이다.

Acknowledgement

This research was supported by grant no. 10037283 from the Industrial Strategic Technology Development Program funded by the Ministry of Knowledge Economy.

핵심용어: 근거기반의학; 의료정보 컴퓨팅; 의사결정지원 기술; 정보 검색; 인공지능

REFERENCES

1. Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J. Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. BMJ 1999;318:527-530.
2. Buchanan BG, Shortliffe EH. Rule based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Boston: Addison-Wesley Longman Publishing; 1984.
3. Vapnik VN. Statistical learning theory. New York: Wiley-Interscience; 1998.
4. Tan PN, Steinbach M, Kumar V. Introduction to data mining. Boston: Addison Wesley Longman; 2007.
5. Heckerman D. Bayesian networks for data mining. Data Min Knowl Discov 1997;1:79-119.
6. Lam W, Low KF, Ho CY. Using a Bayesian network induction approach for text categorization. In: Pollack M, editor. Proceedings of the 15th International Joint Conference on Artificial Intelligence; 1997 Aug 23-29; Nagoya, Japan. San Francisco: Morgan Kauffman; 1997. p. 745-750.
7. Cho SB, Won HH. Machine learning in DNA microarray analysis for cancer classification. In: Chen YP, editor. Proceedings of First Asia-Pacific Bioinformatics Conference; 2003 Feb 4-7; Adelaide, Australia. Darlinghurst: Australian Computer Society. p. 189-198.
8. Orr MJ. Introduction to radial basis function networks. Edinburgh: Centre for Cognitive Science; 1996.
9. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn 1997; 29:103-130.
10. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.
11. Grishman R, Sundheim B. Message understanding conference-6: a brief history. In: Tsujii J, editor. Proceedings of the 16th Conference on Computational Linguistics; 1996 Aug 5-7; Copenhagen, Denmark. Stroudsburg: Association for Computational Linguistics; 1996. p. 466-471.
12. Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support evidence based medicine. BMC Bioinformatics 2011;12 Suppl 2:S5.
13. Nawaz R, Thompson P, McNaught J, Ananiadou S. Meta-knowledge annotation of bio-events. In: Nicoletta C, Khalid C, Bente M, Joseph M, Jan O, Stelios P, Mike R, Daniel T, editors. Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010); 2010 May 17-23; Valletta, Malta. [place unknown]: European Language Resources Association; 2010. p. 2498-2505.
14. Pan F. Multi-dimensional fragment classification in biomedical text [dissertation]. Ontario: Queen's University; 2006.
15. Song MH, Kim SH, Park DK, Lee YH. A multi-classifier based guideline sentence classification system. Healthc Inform Res 2011;17:224-231.
16. Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York: Wiley-Interscience; 2000.
17. Huan Liu, Lei Yu. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 2005;17:491-502.
18. Baharudin B, Lee LH, Khan K. A review of machine learning algorithms for text-documents classification. J Adv Inf Technol 2010;1:4-20.
19. Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In: Fisher DH, editor. ICML 97: proceedings of the 14th International Conference on Machine Learning; 1997 Jul 8-12; Nashville, USA. San Francisco: Morgan Kaufmann; 1997;412-420.
20. Silla CN, Pappa GL, Freitas AA, Kaestner CA. Automatic text summarization with genetic algorithm-based attribute selection. Adv Artif Intell 2004;3315:305-314.
21. Anbarasi M, Anupriya E, Iyengar NC. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. Int J Eng Sci Technol 2010;2:5370-5376.
22. Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio: the analysis and navigation of molecular networks. Bioinformatics 2003;19:2155-2157.
23. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21.
24. Pub anatomy: integrated exploration of biomedical literature

and data [Internet]. Ann Arbor: Microarray Lab [cited 2012 Jul 23]. Available from: http://brainarray.mbni.med.umich.edu/Brainarray/prototype/PubAnatomy.

25. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. AliBaba: PubMed as a graph. Bioinformatics 2006;22:2444-2445.
26. Stoyanovich J, Mee W, Ross KA. Semantic ranking and result visualization for life sciences publications. In: Li F, Moro MM, Ghandeharizadeh S, Haritsa JR, Weikum G, Carey MJ, Casati F, Chang EY, Manolescu I, Mehrotra S, Dayal U, Tsotras VJ, editors. Proceedings of the 26th Interna-tional Conference on Data Engineering, ICDE 2010; 2010 Mar 1-6; Long Beach, USA. [place unknown]: IEEE; 2010. p. 860-871.
27. Yu H, Kim T, Oh J, Ko I, Kim S, Han WS. Enabling multi-level relevance feedback on PubMed by integrating rank learning into DBMS. BMC Bioinformatics 2010;11 Suppl 2:S6.

Peer Reviewers' Commentary

본 논문은 최근 빅 데이터의 중요성이 강조됨에 따라 '임상근거지식'의 자동 추출 기법들과 PubMed 등과 같은 임상문헌 검색 및 지식 발견 도구를 잘 정리하여 소개하고 있다. 또한 임상 의료인과 의학 연구자가 신속히 최신 문헌의 정보를 확인하거나 진료 및 질병 예방을 위한 과학적 근거를 발견/평가하고, 축적되어 있는 데이터 간의 연관성 등을 분석하여 기술하고 있다. 이를 통해 과학적 근거를 평가하거나 진료 행위에 도입하는 등의 의사결정을 위한 지원 방법으로 효율적이라 사료된다. 그러나 논문에서 소개된 다양한 기계학습 및 텍스트 마이닝 기법들은 각기 장단점(사용 용이성, 알고리즘 복잡도 등)을 가지고 있으므로 효과적인 지식 추출을 위해서는 상호 보완적으로 사용되는 것이 바람직하며 추출된 지식의 신뢰도를 측정.검증하기 위한 심도 깊은 연구가 필요할 것이다.

[정리: 편집위원회]