기계학습 기술을 활용한 화학분야 특허문서의 조성/물성 정보 자동추출 방법 연구

김홍기O, 이하영*, 박진우*
한국특허정보원 R&D센터O
한국특허정보원 R&D센터*
e-mail: {khkcapO, youngsll*, znu808*}@kipi.or.kr

A Study on the Automatic Extraction of Fomulation and Properties in Chemical Field Patent Document by Using Machine Learning Technology

Hongki KimO, Hayoung Lee*, Jinwoo Park*
R&D Center, Korea Institute of Patent InformationO
R&D Center, Korea Institute of Patent Information*

● 요 약 ●

본 논문에서는 화학분야 특허 문서에 존재하는 도표(TABLE) 데이터를 인공지능 기술을 활용하여 자동으로 추출하고 정형화된 형태로 가공하는 방법을 제안한다. 특허 문서에서 도표 데이터는 실시예에서 실험결과나 비교결과를 간결하고 가시적으로 표현하기 위하여 주로 사용되나, 셀의 속성을 정의하는 헤더부분과 수치가 표현되는 값 부분의 경계가 모호하여 구조화하는데 어려움이 있다. 본 논문에서 제안하는 방법은 소량의 학습데이터를 구축하고 기계학습을 통해 도표에 존재하는 셀의 속성을 예측하고, 예측된 속성을 토대로 조성과 물성 정보를 자동으로 구분하여 추출하는 방법을 제시한다. 제시된 방법을 활용하여 화학 분야 조성물 특허의 도표데이터에 시뮬레이션 결과 각 항목별 98.17%의 속성 예측 정확도를 나타내었으며 기존 규칙기반 연구보다 작업난이도, 예측정확도에서 우수한 성과를 보인다.

키워드: 인공지능(AI), 문서(HTML, XML), 도표(table), 자동추출(automatic extraction), 기계학습(Machine Learning), 고분자(polymer)

I. Introduction

최근 4차 산업혁명에 대한 관심이 고조됨에 따라 관련 핵심기술인 인공지능에 대한 연구가 활발히 진행되고 있다. 하드웨어의 발전으로 컴퓨팅 파워가 커지면서 여러 분야에서 수많은 알고리즘이 개발되고 수없이 많은 데이터를 활용하여 다양한 의미를 찾아내는 연구가 시도되고 있다. 본 논문에서는 고분자 복합수지 관련 특허 문헌 중 도표에 존재하는 조성에 따른 물성정보를 추출하고자 기계학습 알고리즘을 활용하였다. 화학분야 특허 문헌에서 도표는 해당 조성물을 구성하는 실험결과나 물질적 특성을 일목요연하게 나타내기 위해 주로 사용한다. 그러나 국내 특허 문서 내 도표데이터는 행과 열에 대한 순번 외에 각 셀이 조성인지 물성인지에 대한 속성 정보는 포함되어 있지 않다. 따라서 도표의 데이터를 자동으로 추출하고 도표의 구조와 내용을 정형화된 형태로 관리하는 것은 매우 중요하다. 속성은 도표를 구성하고 있는 텍스트 데이터가 가진 의미를 말하며 관련지식이 있는 사람은 도표를 보면 바로 이해할 수 있으나 기계는 이해할 수 없다는 한계가 있다. 이를 해결하기 위하여 테이블에 존재하는 모든 셀에 대해 속성이 가지는 의미를 부여하고, 부여된 속성 기준으로 기계학습을 통해 도표 정보를 자동추출하는 방법을 제안한다.

II. Preliminaries

1. 기계 학습의 정의

기계학습이라는 용어는 1959년 A. L. Samuel의 논문 “Some Studies in Machine Learning Using the Game of Checkers”[1]에서 처음 발표되었다. 논문에서 기계학습은 “기계가 일일이 코드로 명시하지 않은 동작을 데이터로부터 학습하여 실행할 수 있도록 하는 알고리즘을 연구하는 분야” 라고 정의하고 있다. 기계학습은 정보의 양이 방대해짐에도 불구하고 정확하면서 저렴하고 빠른 정보처리와 계산이 가능하여 많은 연구 분야에서 각광을 받고 있다[2].

2. 기계 학습의 활용 및 구분

현재 기계학습은 인터넷 정보검색, 텍스트 마이닝, 생물정보학, 바이오메트릭스, 자연언어처리,음성인식, 컴퓨터비전, 컴퓨터그래픽, 로보틱스, HCI, 통신사업, 서비스업, 제조업 등 거의 모든 융합연구 분야에서 핵심 기반 기술로 활용되고 있다[3].

기계학습은 학습 데이터를 획득하는 방법에 따라 지도학습, 비지도학습으로 구분된다. 지도 학습은 입출력 데이터 쌍을 기반으로 입력을 출력에 매핑하는 기능을 학습하는 방법[4]이며 대표적으로 SVM(Support Vector Machine), 은닉 마르코프 모델(Hidden Markov Model), 나이브 베이즈(Naive Bayes), 로지스틱회귀(Logistic Regression), 신경망(Neural Network) 등이 존재한다. 반면, 비지도 학습은 입력 값만을 데이터로 갖는다. 따라서 입력 데이터에 내재되어 있는 의미를 분석하는 것을 목적으로 하는 방법[5]이다.

본 논문에서는 속성을 예측하기 위하여 지도학습 알고리즘에 속한 SVM(Support Vector Machine), 나이브 베이즈(Naive Bayes), 로지스틱회귀(Logistic Regression) 알고리즘을 활용한 연구를 진행한다.

III. The Proposed Scheme

1. 적용 방안

본 논문에서는 기계학습 관련 기술을 활용하여 도표 데이터의 셀 정보를 학습하고 학습된 결과를 토대로 학습하지 않는 셀에 속성을 부여한다. 이를 위하여 관련 특허 문헌의 도표정보를 추출하고 학습데이터를 구축하여 기계학습 알고리즘으로 학습한다. 학습이 완료된 모델에 대하여 테스트를 수행하여 학습에 따른 예측 정확도를 측정한다. 측정결과에 따라 가장 정확도가 높은 알고리즘을 선정하고 선정된 알고리즘을 기반으로 도표데이터에 속성을 예측하고 조성/물성에 의한 실험정보를 추출한다.

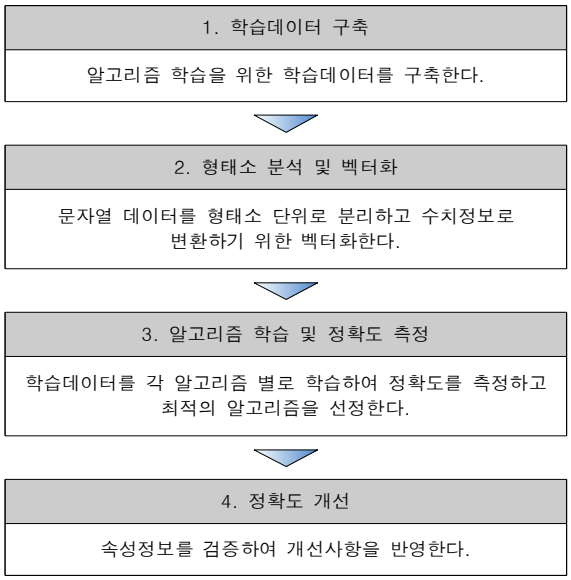

Fig. 1. 기계학습을 위한 알고리즘 적용 방안

2. 실험 및 결과

2.1 학습데이터 구축

열 및 전기전도도 특성을 가지는 고분자 화합물에 관련된 특허문서 280건의 도표 정보 내에서 수작업으로 추출한 셀 항목을 정된 속성(조성, 물성, 실험번호, 수치)으로 Labeling하여 정답 데이터 56,190개을 CSV형태의 파일로 구축하였다.

	폴리카보네이트 중량(%)	탄소나노튜브 중량(%)	유리섬유 중량(%)	전기 전도성 (Ω/sq)	인장 강도 (MPa)
실시예 1	89.5	1.0	15	10^{5}	92.8
비교예 1	89.5	1.0	15	10^{7-8}	88.4
실시예 2	80	5	15	10^{3}	82.3
비교예 2	80	5	15	10^{4-5}	70.3

Fig. 2. 화학 분야 도표 데이터 예시

예를 들어 <Fig. 1>의 도표에서 테이블에 존재하는 화학 조성물에 대한 속성을 조성(폴리카보네이트, 탄소나노튜브, 유리섬유), 물성(전기전도성, 인장강도), 실험번호(실시예1, 비교예1 등)의 헤더정보와 수치정보에 해당하는 값의 4가지로 지정한다.

Table 1. 학습데이터 구성 요소

구분	갯수
조성	15,737
물성	12,365
실험번호	15,723
값(수치)	12,365
합계	56,190

2.2 형태소 분석 및 벡터화

기계학습의 학습데이터로 활용하기 위해서 문자열 데이터를 최소한의 의미를 가지는 형태소 단위로 분리하고 수치화된 배열형태로 벡터화하는 작업을 수행한다. 벡터화 과정에서의 학습 효율성을 판단하기 위해 전체 학습데이터에서 빈도수가 높은 형태소를 2,000개와 4,000개를 선정하였으며 각 셀의 내용을 분리된 형태소의 존재 여부(0 또는 1)를 값으로 가지는 배열(1*2000, 1*4000)로 변환하였다. 벡터화 된 전체 정답데이터를 분리하여 40,000개은 학습데이터로 활용하였으며 학습되지 않은 16,190개을 테스트 데이터로 활용하였다.

Table 2. 학습데이터 및 벡터화 구성 환경

구분	내용
학습용 서버	Dell(Linux)
GPU	Tesla P100
개발언어	Python 3.6
활용 알고리즘 구현 API	konlpy, keras
학습 데이터 수	40,000개
테스트 데이터 수	16,190개

2.3 정확도 측정

벡터화 된 학습 데이터 40,000개을 학습데이터로 활용하여 각 기계학습 알고리즘과 빈도수에 따른 벡터사이즈(1*2000, 1*4000)으로 학습하였고 테스트 데이터 16,190개의 정답률을 측정하였다. Python에서 제공하는 SkleamClassfier API를 활용하여 각 알고리즘을 구현하였다.

Table 3. 알고리즘별 속성 예측 정확도

알고리즘		정확도	
		1*2000	1*4000
SVM	**LinearSVC**	**97.22%**	**98.17%**
	SVC	65.67%	67.59%
Naive Bayes	Original	94.82%	94.82%
	BernoulliNB	94.10%	94.47%
	MNB	90.90%	91.95%
LogisticRegression		96.03%	96.42%

측정결과 SVM(Support Vector Machine) 알고리즘에 해당하는 LinearSVC 알고리즘으로 1*4000형태로 벡터화 데이터를 학습하였을 때 98.17%로 가장 높은 정확도를 보이고 있음을 확인하였으며 이는 기존에 학습되지 않은 10*10을 형태(100셀)를 가지는 도표데이터에서 2개 미만의 셀을 제외하고 정확하게 분류할 수 있는 수준이다. 학습된 모델을 활용하여 <Fig. 1>의 예시 데이터에 대하여 개별 속성정보를 예측하였으며 예측된 속성을 반영하여 웹페이지로 재구성하였다.

	폴리카보네이트 중량(%)	탄소나노튜브 중량(%)	유리섬유 중량(%)	전기 전도성 (O/sq)	인장 강도 (MPa)
실시예 1	89.5	1.0	15	10^{5}	92.8
비교예 1	89.5	1.0	15	10^{7-8}	88.4
실시예 2	80	5	15	10^{3}	82.3
비교예 2	80	5	15	10^{4-5}	70.3

null	폴리카보네이트 중량(%)	탄소나노튜브 중량(%)	유리섬유 중량(%)	전기 전도성 (O/sq)	인장 강도 (MPa)
실시예 1	89.5	1.0	15	10^{5}	92.8
비교예 1	89.5	1.0	15	10^{7-8}	88.4
실시예 2	80	5	15	10^{3}	82.3
비교예 2	80	5	15	10^{4-5}	70.3

Fig. 3. 예측 결과를 반영한 도표 데이터

<Fig. 3>에서 폴리카보네이트 및 탄소나노튜브, 유리섬유가 기재되어있는 셀을 조성으로 예측(노란색)하였으며 전기전도도, 인장강도 정보가 기재된 셀을 물성(녹색)으로 예측하였다. 또한 실험정보(붉은색)와 값(파란색) 셀을 정확하게 예측하였음을 확인하였다.

2.4 규칙기반 정의를 통한 정확도 개선

대상으로 선정된 열 및 전기전도도 특성을 가지는 고분자 화합물에 관련된 특허 문서의 도표 정보의 예측결과를 확인하여 예측치를 벗어난 셀에 대하여 규칙기반으로 정의하였다.

null	비교예 1	비교예 2	비교예 3	비교예 4
LCP 1(부피부)	100	100	100	100
과립 1 (부피부)	0	0	0	0
과립 2 (부피부)	0	0	0	0
과립 3 (부피부)	0	0	0	0
과립화되지 않은 알루미나 섬유 (부피부)	42.9	0	0	0
알루미나 입자 (부피부)	0	42.9	66.7	100
열전도성 (MD) (W/mK)	성형 불가능	1.6	1.8	2.1
열전도성(TD) (W/mK)	성형 불가능	0.7	0.9	1.4
인장 강도 (MPa)	성형 불가능	97	75	63

Fig. 4. 육안검증에 의한 예측 오류데이터 예시

<Fig. 4>에서 "성형 불가능" 의 정보가 기재된 셀은 열전도성 및 인장강도에 대한 값 정보임을 육안으로 확인할 수 있으나 모델 예측 결과 물성 속성으로 예측하였다. 이와 같은 오류 건에 대하여 규칙을 추가하여 정확도를 개선하였다.

IV. Conclusions

본 논문에서는 SVM(Support Vector Machine), 나이브 베이즈(Naive Bayes), 로지스틱회귀(Logistic Regression) 알고리즘을 통하여 문서 내 존재하는 테이블의 속성을 예측하여 조성/물성 정보를 추출하는 효과적인 방법을 제안하였다. 제안된 방법은 기계학습을 위한 학습데이터 구축 방법, 학습을 위한 문자열 데이터의 벡터화 방안, 알고리즘 별 학습 정확도 측정, 정확도 개선을 위한 규칙 정의를 포함하고 있다. 결론적으로 테이블에 존재하는 셀의 속성을 최고 98.17%의 높은 수준으로 예측하였다. 향후 연구에서는 추가적인 학습데이터 확보하고 알고리즘에 대한 개선을 통하여 예측정확도를 개선할 예정이다. 기계학습에 대한 연구는 특정 업무 영역에서 기존 사람의 육안으로 규칙을 발견하고 정의하여 구현하는 방식을 벗어나 데이터를 학습하고 학습한 결과를 토대로 예측하는 부분에서 높은 가능성을 보이고 있다.

ACKNOWLEDGEMENT

이 논문은 2019년도 정부(산업통상자원부)의 재원으로 한국산업기술평가관리원의 지원을 받아 수행된 연구임
(No.20000391, 열 및 전기특성 플라스틱 복합수지의 빅데이터 구축과 인공지능 기술을 활용한 정확도 90% 이상의 조성/물성 예측 및 용도 추천 플랫폼 개발)

REFERENCES

[1] 2.A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development", Vol. 3, No. 3, 1959.

[2] SAS Institute, “Machine Learning: What it is & why it matters”, http://www.sas.com/en_us/insights/ analytics/ machine-learning.html (December 1, 2015)

[3] B.-T. Zhang, “Next-Generation Machine Learning Technologies. Journal of Computing Science and Engineering”, Vol. 25, No. 3, pp. 96-107, 2007.

[4] Stuart J. Russell, Peter Norvig (2010) Artificial Intelligence: A Modern Approach, Third Edition, Prentice Hall ISBN 9780136042594.

[5] Hinton, Jeffrey; Sejnowski, Terrence (1999). Unsupervised Learning: Foundations of Neural Computation. MIT Press. ISBN 978-0262581684.