J. Soc. Korea Ind. Syst. Eng Vol. 41, No. 4 : 189-195, December 2018
https://doi.org/10.11627/jkise.2018.41.4.189
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

Extracting Specific Information in Web Pages Using Machine Learning

Joung-Yun Lee · Jae-Gon Kim†

Industrial and Management Engineering, Incheon National University

머신러닝을 이용한 웹페이지 내의 특정 정보 추출

이정윤 · 김재곤†

인천대학교 산업경영공학과

With the advent of the digital age, production and distribution of web pages has been exploding. Internet users frequently need to extract specific information they want from these vast web pages. However, it takes lots of time and effort for users to find a specific information in many web pages. While search engines that are commonly used provide users with web pages containing the information they are looking for on the Internet, additional time and efforts are required to find the specific information among extensive search results. Therefore, it is necessary to develop algorithms that can automatically extract specific information in web pages. Every year, thousands of international conference are held all over the world. Each international conference has a website and provides general information for the conference such as the date of the event, the venue, greeting, the abstract submission deadline for a paper, the date of the registration, etc. It is not easy for researchers to catch the abstract submission deadline quickly because it is displayed in various formats from conference to conference and frequently updated. This study focuses on the issue of extracting abstract submission deadlines from International conference websites. In this study, we use three machine learning models such as SVM, decision trees, and artificial neural network to develop algorithms to extract an abstract submission deadline in an international conference website. Performances of the suggested algorithms are evaluated using 2,200 conference websites.

Keywords : Data Extraction, Machine Learning, SVM, Decision Tree, Neural Network

1. 서 론

디지털 시대의 도래로 전 세계적으로 웹페이지의 생산과 유통이 폭발적으로 증가하고 있다. 인터넷을 이용하는 사용자들을 중심으로 이러한 방대한 웹페이지들 속에서 자신이 원하는 특정한 정보를 신속하게 찾고 싶어 하는 요구가 날이 갈수록 증가하고 있다. 하지만 탐색 대상 정보량이 많아짐에 따라 사용자가 웹서핑을 통해 직접 인터넷 웹페이지에서 원하는 정보를 찾기 위해서는 많은 시간과 노력이 필요하다. 원하는 정보를 얻고자 보편적으로 사용하는 검색엔진은 사용자가 찾고자 하는 정보가 포함된 웹페이지를 인터넷에서 검색하여 사용자에게 제공해 주기는 하나, 검색결과가 방대할 경우 결과 내에서 사용자의 원하는 정보를 찾기 위한 시간과 노력이 추가적으로 많이 요구된다. 따라서 주어진 웹사이트에서 사용자가 필요로 하는 정보를 자동적으로 추출할 수 있는 특정 정보 자동 추출 알고리즘 개발이 필요하다.

문서 속의 특정한 정보를 추출하기 위한 알고리즘에

Received 26 November 2018; Finally Revised 12 December 2018; Accepted 13 December 2018
† Corresponding Author : jaegkim@inu.ac.kr

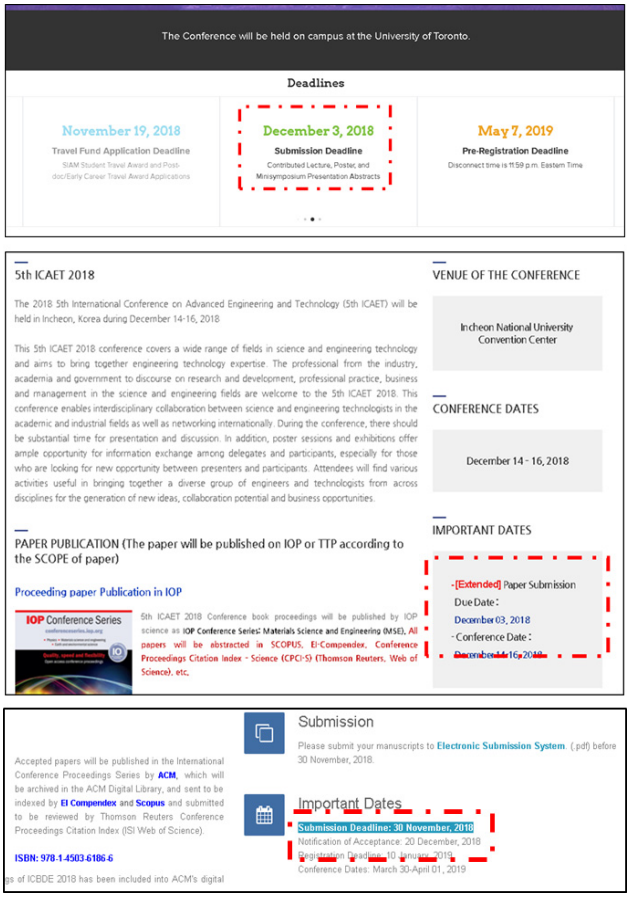
관한 연구들은 현재까지 통계적 추출 알고리즘 개발 위주로 진행되어 왔다. 즉, 웹페이지의 HTML 태그를 분석하여 정보의 특징을 매칭하거나[2, 4], 문서 속 단어의 상관관계와 형태소를 이용하여 제목, 주제, 이름, 주소 등을 정보를 추출하였다[3, 8, 9, 12, 14, 15, 16]. 이러한 방법들의 공통점은 추출하고자 하는 정보의 표현 또는 등장 규칙을 알아내어 추출하는 방법이다. 찾아낸 규칙은 사람이 이해할 수 있을 정도로 명확한 규칙이며 이러한 규칙이 명확히 존재하지 않는 정보들을 추출하는 것은 어려운 일이다. 또 특정 규칙을 갖는 정보라 할지라도 각 정보가 가지는 의미를 파악하기는 쉽지 않다. 예를 들어 어떤 텍스트에서 날짜를 추출하더라도 그 날짜가 의미하는 바가 무엇인지를 알아내기 위해서는 전체 문맥을 분석해야만 한다.

본 연구에서는 이런 문제를 해결하고자 머신러닝 기반의 정보 추출 방법을 개발하였다. 머신러닝 기반의 추출 기법은 학습 데이터를 사용하여 정보추출 학습 모델을 개발한 뒤 해당 모델을 이용하여 실험 데이터로부터 원하는 정보를 추출하는 방법이다. 대표적인 머신러닝 알고리즘으로는 인공신경망(Artificial Neural Network), 의사결정나무(Decision Tree), SVM(Support Vector Machine)[11, 13, 16] 등이 있다.

본 연구에서는 웹페이지 내의 특정 정보를 추출하는 머신러닝 기법을 이용한 추출 알고리즘을 개발하기 위하여 SVM, 의사결정나무, 그리고 인공신경망 기법을 사용하였고 세가지 알고리즘들의 성능을 실험을 통해 평가하였다.

본 연구에서는 국제학술대회 웹페이지에서 초록 투고 마감 날짜를 추출하는 문제를 다루도록 한다. 매년 수천 개 이상의 국제학술대회가 세계 곳곳에서 개최된다. 각 국제학술대회는 소개 웹 사이트를 가지고 있으며 개최 날짜, 개최 장소, 인사말, 초록 투고 마감일, 등록 마감일 등과 같은 학술대회 개최 및 참가에 필요한 전반적인 정보를 제공한다. 이중 초록 투고 마감일은 학술대회 웹 사이트마다 서로 다른 페이지에 다양한 포맷으로 표시되고 해당 정보도 자주 업데이트 되기 때문에 연구자가 정확한 정보를 빨리 파악하기 쉽지 않다. <Figure 1>은 국제학술대회 웹페이지 마다 다양한 형태로 표시되는 논문 초록 마감일을 나타내고 있다.

따라서 본 연구에서는 컨퍼런스 웹사이트에서 논문 초록 마감일을 사람의 개입 없이 자동으로 추출하는 알고리즘을 개발하고자 한다. 컨퍼런스 마다 논문 초록 마감일이 존재하는 경우도 있고 그렇지 않은 경우도 있기 때문에, 해당 웹사이트가 논문 초록 마감일을 가지고 있는 경우에는 정확한 정보를 추출하고 그렇지 않은 경우에는 원하는 정보가 존재하지 않음을 알려주는 알고리즘을 개발하도록 한다.



<Figure 1> Indication of Submission Deadlines in Conference Web Sites

2. 데이터

2.1 데이터 수집

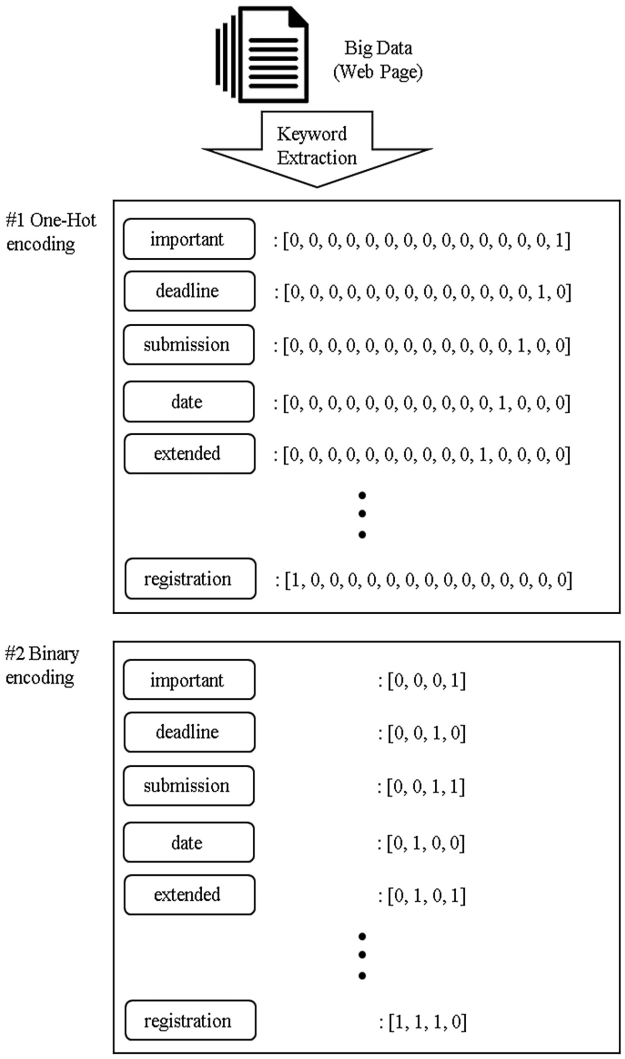
본 연구에서는 국제컨퍼런스 정보를 제공하고 있는 A사의 웹사이트 존재하는 총 2,200개의 국제컨퍼런스 웹사이트들을 실험 데이터로 사용하였다. 앞에서 살펴 본 바와 같이 학술대회 웹페이지의 비정형적 텍스트에서 특정 정보를 추출을 하기 위해선 비정형적 데이터를 정형적인 데이터로 변환하고 이를 분석하는 단계를 거쳐야한다. 이는 텍스트 마이닝(Text Mining)으로 많은 연구가 진행되어 왔지만 텍스트에서 감정을 추출하거나 문서의 종류별로 분류하는 분야에 한정되어 있다. 그러나 본 연구에서는 텍스트에 내포되어 있는 의미보다 텍스트 속에 있는 정보가 가지는 패턴을 이용하여 추출하고자 하므로 키워드 기반 연구 모형(Keyword-based Research Model)[7]을 이용하여 해당하는 정보와 관계되어 있는 키워드를 바탕으로 정형화하는 모형을 세운다. 우선, 웹페이지마다

논문 초록 마감일을 나타내는 날짜와 앞과 뒤로 인접한 텍스트 중에서 논문 초록 마감일에 대한 의미가 명확하고 빈도가 많은 “important”, “deadline”, “submission”, “dates” or “date”, “extended”, “abstract”, “paper”, “due”, “notification”, “acceptance”, “final”, “full”, “post”, “registration” 총 14개의 단서 단어들을 선정하였다.

2.2 데이터 전처리

입력 데이터가 범주형 특징이거나 제한되는 수의단어들로 이루어질 경우 특수한 행렬을 생성하고자 누메릭인코딩(Numeric Encoding), 바이너리인코딩(Binary Encoding), 원핫인코딩(One-Hot Encoding)을 주로 사용하여 데이터를 변환한 후 실험에 사용한다. 누메릭인코딩은 단어의 리스트 순서대로 숫자를 부여하여 데이터를 변환하는 것이며, 이 경우 머신러닝 학습 시 생성되는 가중치에 정도가 각 단어 리스트 사이의 거리에 따라 영향을 많이 받게 되어 입력 데이터의 전처리 과정이 달리 되면 실험을 통해 얻게 되는 결과값이 큰 폭의 차이를 나타낸다. 바이너리인코딩은 이진법의 행렬을 이용하여 데이터를 변환한 것으로 행렬의 크기는 log(N+1)/log(2)(N = 단어의 수)로 정해진다. 바이너리인코딩은 누메릭인코딩보다 단어 리스트 사이 거리의 영향을 작게 받지만 여전히 상당한 영향을 끼친다. 하지만 단어의 개수가 많을 시 발생할 수 있는 과적합(Overfitting)의 우려가 있을 경우 사용한다. 원핫인코딩은 단어의 개수만큼의 크기의 행렬을 만들어 데이터를 처리하는 것으로 단어 리스트 사이의 거리에 큰 영향을 받지 않지만 과적합의 우려가 발생한다. 본 연구에서는 의사결정나무 모델에서는 바이너리인코딩을 사용하였고, SVM과 인공신경망 모델에서는 원핫인코딩을 사용하였다. <Figure 2>는 원핫인코딩과 바이너리인코딩이 사용된 예를 보여준다.

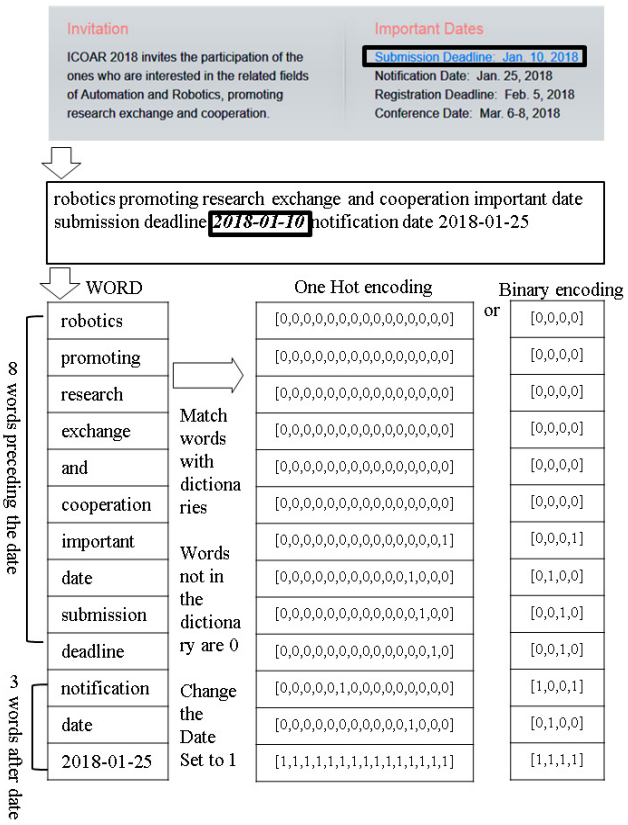
본 연구에 사용하는 문제의 특성상 국제학술대회 웹 페이지의 논문 초록 마감일은 날짜의 형태를 가지고 있다. 웹 페이지의 HTML 문서를 크롤링(Crawling)하여 텍스트를 변화한 뒤 날짜를 나타내는 모든 정보를 규칙 기반의 알고리즘을 이용하여 추출한다. 연도를 나타내는 텍스트의 바로 앞 혹은 뒤에 텍스트가 월을 의미한다면 그 텍스트와 바로 인접한 텍스트가 일을 의미하는지를 확인하고 날짜를 나타낸다고 가정하였다. 날짜를 나타내는 모든 텍스트를 묶어 하나의 텍스트로 간주하고 하나의 값을 주었다. 원핫인코딩에서는 “[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]”이고, 바이너리인코딩에서는 “[1, 1, 1, 1]”으로 표현하였다. 날짜를 표현하는 텍스트들의 묶음을 날짜셋이라 하겠다. 웹 페이지를 크롤링 하였을 때 첫 페이지에 날짜셋이 전혀 존재하지 않을 경우,



<Figure 2> Keyword Dictionary Building

HTML 태그 중 링크를 포함하는 태그 a가 포함하는 text에서 “date”라는 단어가 포함한 링크 하나를 선택하여 해당 페이지를 크롤링하여 위와 동일한 절차로 날짜셋을 찾았다.

날짜셋이 수식되는 정보는 날짜셋의 앞에 오는 단어이거나 뒤에 오는 단어들로 판별이 되는데, 경험적인 판단으로 대부분의 수식되는 정보는 날짜 셋의 앞에 있기에 입력 데이터를 만들기 위하여 판별하고자 하는 날짜셋보다 앞선 단어는 8개를 뒤의 단어는 3개를 가져와 총 11개의 단어를 가져온다. 가져온 단어 중 앞서 정의한 사전과 비교하여 일치하는 단어를 매치 하였고, 일치하지 않는 단어는 모두 같은 값을 부여 하였는데 원핫인코딩에서는 “[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]”, 바이너리인코딩에서는 “[0, 0, 0, 0]”의 모습인 입력 데이터로 사용하였다. <Figure 3>은 데이터 전처리의 과정을 나타낸다.



<Figure 3> Data Pre-Processing

3. 머신러닝

머신러닝(Machine Learning)은 데이터를 이용하여 명시적으로 정의되지 않은 패턴을 컴퓨터로 분석하고, 분석을 통해 학습하며, 학습한 내용을 기반으로 결과를 예측하는 학문 분야로 크게 지도학습(Supervised Learning), 비지도학습(Unsupervised Learning) 그리고 강화학습(Reinforcement Learning)으로 나뉜다. 지도학습에는 회귀(Regression), 분류(Classification) 모형이 있고, 관측값과 목표값을 연결시켜주는 예측모델이 나무의 형태인 의사결정나무, 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단하는 서포트 벡터 머신, 머신러닝과 인지과학에서 생물학의 신경망에서 영감을 얻은 통계학적 학습 알고리즘인인 인공신경망 등의 기법이 분류 모형에 속한다[10, 14].

3.1 SVM 모델

SVM은 1995년 Vapnik[1]에 의해 개발된 이진분류를 위한 학습기법이다. 두 개의 범주로 구성된 N개의 점이 하나의 분리경계면(Hyperplane)으로 구분이 될 때, 두 범주를 구분하는 분리경계면은 무수히 많을 수 있으나 SVM은 지지벡터(Support Vector)를 이용해 두 그룹을 마진(Margin)이 최대가 되도록 구분하고 지지벡터를 이용해 입력값에 의한 결과값이 속한 범주를 분류하는 모델이다[5].

SVM은 명백한 이론적 근거에 기반하므로 결과 해석이 용이하고 높은 정확도와 적은 데이터만으로 분별을 수행할 수 있기에 이와 같은 이유로 본 논문에서는 SVM의 도구로서 가장 단순한 선형 회귀 방식의 SVM-Light를 사용하였다[16]. 학습을 통해 얻어진 선형함수 $f(X) = W \cdot X+b$로 나뉜 평면에서 새로운 실험 데이터가 어느 범주에 속하는지를 알아내고 선형함수와의 거리로 나눈 범주에 속하는 확률을 알아낸다.

입력데이터의 선정을 위해서 수차례를 실험을 통해 높은 정확도를 나타내는 방식인 원핫인코딩의 데이터를 사용하였다.

3.2 의사결정나무 모델

의사결정나무는 나무구조의 모형으로 도식화된 노드 속에 적절한 의사결정규칙에 따라 도식화하여 분류를 수행하는 방법이다. 의사결정나무분석에서 목표변수에 영향을 미치는 요인을 찾아내고 각 요인의 중요도에 대한 가시적인 관계가 보이며, 각 과정에서 영향을 미치는 중요한 요인에 따라 발생비율이 어떻게 달라지는 지를 도식화된 흐름도로 표현할 수 있다[6].

본 연구에서는 각각의 단어들의 변화로 결과의 확률을 계산하기에 학습 데이터를 몇 개의 소집단으로 분류하고 예측하는 기법인 CART(Classification and Regression Tree) 알고리즘을 채택하였다. 입력 데이터는 원핫인코딩을 사용할 때에 과적합의 문제로 데이터의 크기인 종속변수가 보다 작은 바이너리인코딩을 사용하였다. 분리기준(split criterion)은 0.5수준, 정지규칙(stopping rule)은 4수준으로 지정하였다. 입력 데이터는 바이너리인코딩으로 전처리한 데이터로 하였다[6].

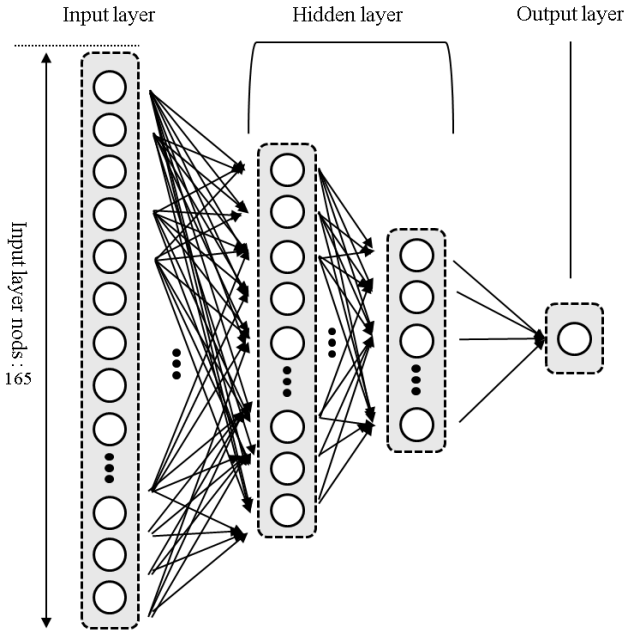
3.3 인공신경망 모델

인공신경망은 생물학적 신경망을 모티브로 하는 알고리즘으로 인공 뉴런을 노드로 적용하여 결합한 네트워크를 일컫는다. 퍼셉트론(Perceptron)은 입력과 출력의 층으로 구성하여 각 계층에서 뉴런간의 가중치(Weight)를 학습하여 분류하는 모델이다. 가중치 조정, 비선형 분류의 불가능 등의 문제를 해결하고자 입력과 출력의 층 사이에 은닉계층(Hidden layer)를 추가한 다중 퍼셉트론(Multilayer Perceptron)이 개발되었다[14].

인공신경망 모델에 입력 데이터를 원핫인코딩 데이터

를 사용하였고 입력하는 단어 개수인 11개에 원핫인코딩으로 늘어난 15개 차원의 곱으로 165개의 입력 노드를 설정하였고 수차례의 실험과 피드백으로 은닉층과 노드 개수, 가중치 조절 및 활성 함수를 선택하였다. <Figure 4>는 사용한 인공신경망 모델을 도식화 한 것이다.

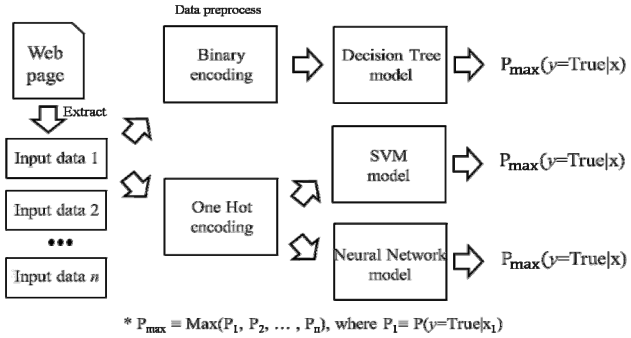
본 연구에서는 입력층에 165개의 노드와 은닉층 2개, 1개의 출력층으로 구성하고, ReLU 활성함수을 사용하여 0.1의 오차에 도달하기 위해 5,000번의 반복, 0.01의 학습률을 설정하였다.



<Figure 4> Artificial Neural Network Model

4. 실험결과

본 연구에서는 2,200개의 실험 데이터(컨퍼런스 웹사이트)에서 2,000개는 학습 데이터를 사용하여 3가지 머신러닝 모델을 학습시키고, 이를 이용하여 나머지 200개 웹사이트에서 논문 초록 마감일을 추출하였다. <Figure 5>는 웹페이지에 머신러닝 모델을 적용하여 논문 초록 마감일을 추출하는 과정을 보여준다. 웹페이지에 n개의 입력 데이터가 존재할 때 각 입력 데이터(x)를 인코딩하여 머신러닝 모델에 넣으면 결과값으로 나오는 확률값 P(x)를 구하고, 이렇게 구해진 n개의 P(x) 중 가장 큰 값인 P_{max} 값을 최종 결과값으로 취한다. 만약 P_{max} 값이 0.7 이상이면 해당 웹 페이지는 논문 초록 마감일이 존재한다고 판단하고, 해당 P_{max} 값을 주는 입력 데이터의 날짜를 논문 초록 마감일로 돌려주게 된다. <Figure 5>는 정보 추출 과정을 도식화하여 보여주고 있다.



<Figure 5> Information Extraction Process

본 실험에서는 사용하는 학습 데이터의 크기에 따라 머신러닝 모델의 정확도가 어떻게 달라지는지를 알아보기 위해 세 가지 종류의 학습 데이터 셋(TS1, TS2, TS3)을 사용하였다. 이를 위해 총 2,000개의 학습 데이터 중에서 랜덤하게 500개, 1,000개, 2,000개의 학습 데이터를 선택하여 세 가지 크기의 학습 데이터 셋을 생성하였다. 이를 이용해 SVM 모델, 의사결정나무 모델, 인공신경망 모델을 학습시킨 후, 새로운 200개의 컨퍼런스 웹사이트를 테스트 데이터로 사용하여 각 사이트에서 논문 초록 마감일을 추출하였다.

SVM 모델과 의사결정나무 모델은 R프로그래밍을 이용해 개발하였고 인공신경망 모델은 구글사가 개발한 TensorFlow 프레임워크와 Python 프로그래밍을 통해 구현하였다. 실험은 학습 데이터를 10번 랜덤하게 생성하면서 10번 반복 수행하였고, 정확도의 산술평균값으로 나타내었다. <Table 1>은 실험결과를 웹 사이트에서 논문 초 마감일이 있는 경우와 없는 경우를 구분하여 보여준다.

<Table 1> Decision Accuracies of the Three Machine Learning Models

Training Data Set	Conference Category	SVM	Decision Tree	Neural Network
TS1 Size of Training Data : 500	In cases that submission deadline exists	68.7%	67.1%	72.1%
	In cases that submission deadline does not exist	95.2%	90.3%	98.7%
TS2 Size of Training Data : 1,000	In cases that submission deadline exists	72.8%	69.4%	80.3%
	In cases that submission deadline does not exist	95.3%	92.7%	98.7%
TS3 Size of Training Data : 2,000	In cases that submission deadline exists	78.4%	70.0%	88.6%
	In cases that submission deadline does not exist	95.6%	96.9%	98.9%

학습 데이터의 크기가 가장 큰 TS3의 경우 세 개의 머신러닝 모델들 모두 70% 이상의 정확도를 보여주었으며, 그 중에서도 인공신경망 모델의 정확도가 가장 높았고 SVM과 의사결정나무모델이 그 뒤를 따랐다. 학술 대회 웹 사이트에 논문 초록 마감일이 포함되어 있지 않는 경우에는 세가지 모델 모두 95% 이상의 높은 정확도를 보여주었다.

학습 데이터의 크기의 증가에 따른 머신러닝 모델의 정확도 향상을 살펴보면, 의사결정나무의 경우 정확도 증가폭이 미미 하였지만 SVM과 인공신경망의 경우 학습 데이터의 크기가 커질수록 더 좋은 성능을 보여주었다. 특히 인공신경망 모델의 경우 학습 데이터의 크기가 증가할수록 모델의 정확도가 눈에 띄게 증가하였다.

5. 결 론

본 논문에서는 머신러닝을 이용하여 국제학술대회 웹 사이트에서 논문 초록 마감일을 자동 추출하는 알고리즘을 개발하였다. 세 가지 대표적인 머신러닝 모델 즉, SVM, 의사결정나무, 인공신경망을 사용하였으며 이를 위해 입력 데이터 전처리 및 인코딩 방법, 복수 개 입력 데이터 배치 처리, 결과값 판단 프로세스 등을 제시하였다. 실험 결과 세 가지 머신러닝 모델 중 인공신경망 모델이 가장 높은 정확도를 보여주었으며, 특히 학습 데이터의 크기가 커질수록 더욱 뛰어난 정확도를 보여주었다.

추후연구로 제시한 인공신경망에서 은닉층의 개수를 늘여 딥러닝모델로 발전시키고 훨씬 더 많은 학습 데이터를 사용하여 모델을 학습시킴으로써 모델의 정확도를 더욱 향상시킬 필요가 있다. 본 연구에서 제안한 머신러닝 모델은 학술대회 사이트에서 논문 초록 마감일을 추출하는 것 이외에도, 웹페이지 내에서 특정 정보를 자동 추출하는 인공지능 모델 개발에 활용될 수 있을 것으로 기대된다.

Acknowledgement

This work was supported by the Incheon National University (International Cooperative) Research Grant in 2018.

References

[1] Coptes, C. and Vapnik, V., Support-Vector Networks, *Machine Learning,* 1995, Vol. 20, No. 3, pp. 273-297.

[2] Emilio, F., Rasquale, D.M., Giacomo, F., and Robert, B., Web data extraction, applications and techniques, *Knowledge-Based Systems,* 2018, Vol. 70, No. 1, pp. 301-323.

[3] Hwang, M.G., Choi, D.J., and Kim, P.K., A Context Information Extraction Method according to Subject for Semantic Text Processing, *Journal of Advanced Information Technology and Convergence,* 2010, Vol. 11, No. 8, pp. 197-204.

[4] Jimenez, P. and Corchuelo, R., On learning web information extraction rules with TANGO, *Journal Information Systems*, 2018, Vol. 62, No. C, pp. 74-103.

[5] Jo, S.R., Sung, H.N., and Ahn. B.H., A Comparative Study on the Performance of SVM and an Artificial Neural Network in Intrusion Detection, *Journal of the Korea Academia-Industrial cooperation Society*, 2016, Vol. 17, No. 2, pp. 703-712.

[6] Kim, G.S. and Park, J.A., Development of a Soil Moisture Estimation Model Using Artificial Neural Networks and Classification and Regression Tree(CART), *Korean Society of Civil Engineers Journal of Civil Engineering*, 2011, Vol. 31, No. 2, pp. 155-163.

[7] Kim, H.S. and Kim, C.S., An Analysis of IT Proposal Evaluation Results using Big Data-based Opinion Mining, *Journal of Society of Korea Industrial and Systems Engineering*, 2018, Vol. 41, No. 1, pp. 1-10.

[8] Kim, P.J., An Analytical Study on Automatic Classification of Domestic Journal articles Based on Machine Learning, *Journal of the Korean Society for Information Management,* 2018, Vol. 35, No. 2, pp. 37-62.

[9] Lee, J.Y., Moon, J.Y., and Kim, H.J., Examining the Intellectual Structure of Records Management and Archival Science in Korea with Text Mining, *Journal of the Korean Society for Library and Information Science*, 2017, Vol. 41, No. 1, pp. 345-372.

[10] Lee, Y.J., Sim, M.K., Lee, S.S., and Lee, C.K., Study of the Operation of Actuated signal control Based on Vehicle Queue Length estimated by Deep Learning, *The Journal of the Korea Institute of Intelligent Transport Systems*, 2018, Vol. 17, No. 4, pp. 54-62.

[11] Li, Y., Bontcheva, K., and Cunningham, H., Using Uneven Margins SVM and Perceptron for Information Extraction, *Proceedings of the Ninth Conference on Computational Natural Language Learning*, 2005, Catalonia, Spain, pp. 72-79.

[12] Noh, T.H. and Lee, S.J., Extraction and Classification of Proper Nouns by Rule-based Machine Learning, *Journal of Korean Institute of Information Scientists*

and Engineers, 2000, Vol. 27, No. 2, pp. 170-172.

[13] Park, N.R., Design and Implementation of Criminal Identification System Based on Deep Learning, [dissertation], [Seongnam-si, Korea] : Gachon University, 2017.

[14] Schneider, K.M. and Textkernel, B.V., Information Extraction from Calls for Papers with Conditional Random Fields and Layout Features, *Artificial Intelligence Review*, 2006, Vol. 25, No. 1, pp. 67-77.

[15] Shin, H.S., Kim, J.H., Lee, H.Y., and Choi, K.S., A Method for Automatic Extraction of Term Definition from Text, *Annual Conference on Human and Cognitive Language Technology*, Chongju-si, Korea, 2002, pp. 292-299.

[16] Son, J.R., SVM Spam Mail Analysis using Feature Selection [dissertation], [Seoul, Korea] : Hankuk University of Foreign Studies, 2005.

ORCID

Joung-Yun Lee | http://orcid.org/0000-0002-4922-1395

Jae-Gon Kim | http://orcid.org/0000-0002-4821-4441