Journal of KIIT. Vol. 17, No. 7, pp. 23-29, Jul. 31, 2019. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2019.17.7.23

직무분석을 위한 머신러닝 기반 전자문서 자동 분류 모델

강은숙*, 고대식**

Automatic Classification Model of Electronic Documents Based on Machine Learning for Job Analysis

Eun-Sook Kang*, Dae-Sik Ko**

요 약

정부에서는 4차 산업혁명시대를 대비하고, 공정하고 투명한 정부 인사관리 실현을 위해 직위에 적합한 맞춤형 인사추천 서비스를 도입하였다. 또한 시범부처운영을 통해 추천 서비스 모델의 정확도 및 품질을 높이기 위한 노력을 기울이고 있다. 본 연구에서는 부서에서 생산되는 전자문서를 머신러닝 기술의 지도학습으로 업무 성격별로 자동 분류하여 직무분석에 활용하는 방법을 제안하였다. 기존의 직무분석 시 부서의 업무성격에 대한 비중은 담당자의 경험기반으로 수작업으로 제시하였으나, 본 연구에서는 지도학습의 대표적인 유형인 베이시안 확률 기반의 나이브 베이즈(Naive Bayes)모델링으로 전자문서를 분류하여 비중을 산출하였다. 실험 결과 담당자가 사전에 업무의 성격별로 라벨링 한 전자문서와 자동 분류된 전자문서와 비교한 결과 80%이상의 정확도를 보였다. 따라서 업무성격별로 전자문서를 자동 분류하는 결과가 직무분석을 위한 정보로 활용 가능함을 확인하였다.

Abstract

The goverrnment has introduced customized personnel recommendation services suitable for positions in order to prepare for the fourth industrial revolution era and realize fair and transparent personnel management. In addition, the government has made efforts to improve the accuracy and quality of the recommendation service model through the operation of test departments. This study proposes a method to automatically classify by job characteristics electronic documents produced by departments through supervised learning in the machine learning technology, and apply them to job analysis. Contrary to the existing job analysis manually presenting weights of job characteristics at departments on the basis of the experience of the person in charge, the study assorts the electronic documents with Naive Bayes modeling based on Bayesian probability as a representative type of supervised learning and calculates weights accordingly. Compared with the electronic documents that the person in charge have previously labeled according to the nature of the work, 80% or more. Therefore, it is acknowledged that the result of automatic sorting of electronic documents by job characteristics can be utilized as information for job analysis.

Keywords
machine learning, supervised learning, automatic classification, job analysis

* 목원대학교 지능정보융합과 박사과정
- ORCID: https://orcid.org/0000-0002-4495-4135
** 목원대학교 전자공학과·지능정보융합과 교수 (교신저자)
- ORCID: https://orcid.org/0000-0002-6232-476X

· Received: May 20, 2019, Revised: Jul. 07, 2019, Accepted: Jul. 10, 2019
· Corresponding Author: Dae-Sik Ko
Dept. of Electronic Engineering, Mokwon University, 88 Doanbuk-ro, Seo-gu, Daejeon, Korea.
Tel.: +82-42-829-7652, Email: kds@mokwon.ac.kr

Ⅰ. 서 론

최근 정부에서는 지능정보사회 구현을 위한 국가 정보화 패러다임 전환을 추진하고 있다. 또한 인사 및 성과에 활용하기 위한 정부 주요 직위 직무분석의 중요성과 적재적소의 맞춤 인재추천의 필요성이 부각되고 있다. 인재추천을 위해서는 기본적으로 직위에 대한 직무분석이 최적화 되어야 하고, 지속적인 관리가 요구된다. 직무분석은 정부뿐만 아니라 사회구조가 복잡해짐에 따라 더욱 중요하지만, 현재까지 분석에 따른 시간 및 비용이 많이 소요되며 변경에 따른 현행 화는 더욱 힘든 실정이다. 지능정보 기술과 데이터를 활용하고 업무성격 자동 분류 모델을 적용하여 직무분석시스템을 설계한다면 적재적소 인사를 추천하기 위한 직무요건 정의에 큰 도움을 줄 것으로 판단하였다.

본 연구에서는 정부 주요직위 인사에 활용 중인 직무기술서 가이드에서 제시하는 부서별 9가지 업무성격을 분류기준으로 정하였다[1]. 직무분석을 위한 내용 중 업무의 성격 항목은 구분이 명확하고, 관련 내용도 표준화되어 있다. 그러나 현재는 담당자가 직무요건의 업무 성격 비중을 전문적인 경험 기반으로 수작업으로 작성하고 있다. 이런 문제점을 해결하기 위해 직무분석의 핵심인 업무성격 비중을 도출하기 위한 방안으로 전자문서를 자동 분류하는 모델이 필요하다. 이렇게 분류기준이 명확히 있는 경우 머신러닝의 지도학습으로 자동분류를 하는 것이 정확도가 높은 것으로 평가되고 있다. 그래서 머신러닝의 지도학습의 대표적인 나이브 베이즈 모형(Naive bayesian model)을 적용하였고[2][3], 다양한 변수와 데이터를 활용하여 학습하고 최종적인 정확도를 검증하였다.

Ⅱ. 직무분석을 위한 전자문서 자동분류 모델

2.1 관련연구

본 연구는 현재 사람들에게 주목을 받고 있는 인공지능 기술 중 머신러닝을 활용하여 전자문서 자동 분류를 수행하고자 한다. 머신러닝이란 프로그래밍 되지 않은 데이터를 컴퓨터가 학습할 수 있도록 돕는 컴퓨터 분야이다[4]. 특히, 명시적인 프로그래밍으로 좋은 성능을 기대하기 어려운 부분에 주로 사용된다. 학습 데이터 의존도에 따라 지도 학습(Supervised learning)과 비지도 학습(Unsupervised learning)으로 분류된다[5]. 지도학습은 머신러닝(Machine learning)의 한 종류로서 입력 값에 대한 명시적인 정답(Label)을 사전에 학습 시킨 후 새로운 입력 값이 주어졌을 때 미리 학습된 데이터를 통하여 알맞은 명시적 정답(Label)을 예측하는 방법이다. 학습 데이터를 입력으로 받아 가중치를 부여하고, 활성화 함수(Activation function)를 통과하여, 손실 함수를 이용해 예측 값 및 레이블 사이의 손실 오차를 구하고, 오차를 갱신하여 학습하는 일련의 과정을 갖는다[6]. 이러한 지도학습에서 분류를 수행하는 대표적인 알고리즘들은 의사결정 트리, K-최근접 이웃 알고리즘(K-Nearest Neighbors, K-NN), 서포트 벡터 머신(Support Vector Machine, SVM), 나이브 베이즈 모형등이 있다[7].

공공분야에서 지도학습의 대표적인 사례로 서울시의 전자문서 분류 테스트 사례가 있다. 서울시의 경우 업무시스템의 결재문서를 반드시 기능분류체계인 BRM(Business Reference Model)을 이용하여 분류하도록 되어 있다. 이에 한국전자통신연구원은 엑소 브레인(Exobrain, http://exobrain.kr)을 이용하여 서울시 결재문서를 BRM에 맞춰 분류해 보는 실험을 하였다. 이 실험은 총 8개월에 걸쳐 진행되었으며, 실험결과 최종으로 엑소 브레인이 만든 모델이 97.81%의 분류 정확성을 보였다. 엑소 브레인의 형태소 분석, 개체명인식, 분류 등 3종 기술을 활용하였으며, 이 기술은 Sequence Labeling, Structural SVMs알고리즘에 기반하고 있다[8]. 서울시 사례의 경우 '서울시 정보소통광장'에 공개된 결재문서를 활용한 것이며, 자동분류에 사용된 툴 또한 업무협약에 의해 제한적으로 제공받은 것이다. 따라서 이 연구 결과를 본 연구에 적용하기 쉽지 않다. 다만 문서 자동분류의 도입 가능성을 서울시 사례를 통해서도 사전에 확인할 수 있었다. 또한 본 연구의 경우 방대한 데이터 안에서 인공지능 에이전트가 특징이나 패턴을 추출하여 스스로 학습하는 방식인 비지도학습의 군집(Clustering)모델이 아닌 사전에 데이터 라벨링을 통한 지도학습이 적합하다는 사실을 알 수 있었다.

2.2 직무분석의 현행 전자문서 분류방식

정부의 직무분석에 필요한 요건은 다양하게 존재하지만, 본 연구에서는 인사추천 및 채용 등을 위한 주요직위에 대한 직무를 분석하는 것으로 범위를 한정하였다.

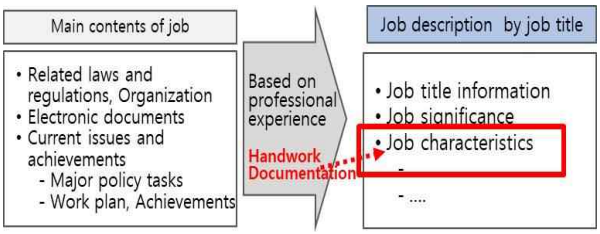

그림 1. 현행분석방식
Fig. 1. Current analysis method

그림 1과 같이 현재는 업무담당자가 직무와 관련된 전자문서, 성과보고자료 등 관련 수작업을 활용하여 경험을 기반으로 한 분석을 통해 직무 기술서를 작성한다. 고위직의 경우「고위직 직무기술서 작성 안내서」와 관련 법령, 직제 및 현안과제, 전자문서 등 수작업 자료를 참고하여 판단한 뒤 직무기술서를 작성하고 있다.

직무기술서 내용 중 업무의 성격은 표 1과 같이 총 9가지 분류로 나누어진다. 업무의 성격은 직무담당자가 수행해야 하는 업무의 기능상의 특징을 분석하여, 직무의 합리적인 조정, 개방형 직위의 운영 등에 필요한 기초정보를 얻는 것이다[1][9]. 또한 직무분석 시 업무성격에 대한 비중은 현행 직무와 관련해 수행하고 있는 활동을 종합적으로 고려하여 9개 분류항목에 값을 작성하고 총 합이 100%가 되도록 해야 한다.

2.3 업무성격별 전자문서 자동분류 모델 설계

그림 2는 직무분석을 위한 업무성격별 비중 값을 산출하기 위해 전자문서를 자동 분류하는 모델 흐름을 표현하였다. 모델 흐름을 세부적으로 살펴보면, 먼저 부처별 운영 중인 전자문서관리시스템에서 14~18년 전자문서를 수집하였다. 수집된 데이터를 분석하기 이전에 개인정보 및 형태소 분석을 위한 전처리 작업을 하였다. 다음으로 지도학습의 알고리즘 중 나이브 베이지 알고리즘을 적용하여 데이터 분석을 하였다. 그 결과 업무성격별로 분류한 값을 생성하여 직무기술서 가이드에서 제시하는 바와 같이 총합을 100%가 되게 하였다.

최종적으로 그림 2에서와 같이 실제적으로 자동분석된 값을 자동분류 항목에 표시하고, 직무분석에 바로 적용하지 않았다. 비중(변경) 항목을 두어 담당자가 자동 분류된 값을 확인한 뒤 필요하면 수정 및 보정을 할 수 있도록 하였다. 직무분석에 필요한 전자문서 자료 외에 내용을 감안할 수 있는 부분을 살려두어 실제적으로 활용할 수 있는 모델을 제안하였다.

표 1. 직무기술서 가이드 업무의 성격
Table 1. Job characteristics in job description guidance

Division	Content
1) Strategy	Direction setting and decision of specific policy and business
2) Execution management	Business execution and management by decision and instruction of upper department
3) Securement and allocation of resources	Securing and allocating necessary resources externally and internally
4) Regulation and supervision	Generally managing regulation and supervision jobs on various groups
5) Judgement and examination	Judgement and examination on a specific matter
6) Research and analysis	Research and analysis on policy and technology
7) Adjustment of disputes and conflicts	Resolution and arbitration of disputes and conflicts of interest
8) Internal support and management	Generally managing internal management jobs within the authorities
9) Others	

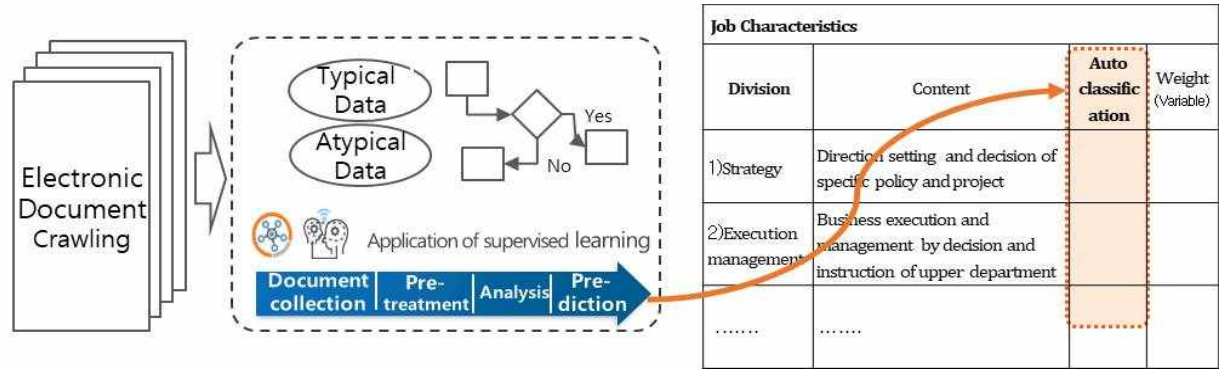

그림 2. 업무성격별 전자문서 자동분류 모델 흐름도
Fig. 2. Flowchart of automatic classification model of electronic documents by job characteristics

III. 실험 및 고찰

3.1 실험 조건

2014~2017년간 4개 부서에서 생산된 전자문서를 대상으로 학습을 준비하고, 2018년도 전자문서는 분류모델을 검증 데이터로만 활용하였다. 수집된 전자문서는 전처리 후 분석 DB로 구축하였고, 표 2와 같이 변수를 선별하였다.

표 2. 데이터 변수 종류
Table 2. Type of data variables

Input variables	Target variable
Document title (RECORD_TITLE)	Job characteristics
Summary of the document	
Body of the document	
Name of attached file	
Name of unit task card	
Department	

모델은 지도학습방법의 대표적인 모델인 나이브 베이즈 분류기를 이용하였고, 과 적합(Overfitting) 방지를 위해 학습과 검증을 위한 데이터를 분할하여 학습 80%, 검증 20%로 결정하였다. 지도학습은 모델링 전 예측해야 되는 목표변수를 미리 달아줘야 하는 어려움이 있었다. 또한 담당자의 일관성 있는 판단이 정확도에 미치는 영향이 높았다. 모델에 대한 평가는 모델 성능을 정확도로 측정하고 정확도 개선에 고려된 사항을 살펴보기로 하였다.

3.2 실험 방법 및 결과

실험방법은 표 3과 같이 4차례 순서대로 진행하였다. 담당자가 사전에 수작업으로 라벨링 한 데이터를 중심으로 1번에서 3번까지 변수 값을 다르게 적용하는 학습을 한 뒤 가장 결과가 좋은 경우의 모델을 적용하여 4번으로 검증을 실시하였다.

표 3. 실험 순서 및 데이터
Table. 3. Test sequence and data

Order of experiment	Electronic document
1. Learning by document title only	2014 ~2017
2. Learning by document title and name of unit task card	
3. Learning by document title, body of the document and name of attached file	
4. Learning by model classified by document title	2018

첫번째 실험방법은 A부서의 문서제목(RECORD_TITLE)을 입력변수 값으로 학습하였다.

실험결과 표 4와 같이 2014~2017년 전자문서 9757건을 대상으로 학습 7806건, 검증 1951건에 대해 업무성격에 대한 오답이 287개, 정확도는 85%가 나타났다. 업무담당자가 수작업으로 진행한 라벨링 데이터의 정확도를 높이기 위한 정비를 사전에 수행한 뒤 실험한 결과이다.

A부서 실험을 마친 후 3개 부서 데이터를 추가 학습하였다. 다른 3개 부서에 대한 종합분석 결과 (학습데이터 68,286건) 81%로 정확도가 나타났다. 각 부서별 정확도는 A부서 85%, B부서 88%, C부서 85%, D부서 83%로 부서별로 정확도가 상이했으며 부서별 업무성격이 다르다는 것을 확인하였다. 비슷

한 업무 유형이지만 부서별로 전자문서 제목이 다른 경우도 발생하였다.

두 번째 실험은 제목과 요약, 단위 과제 카드명을 포함하여 실험하였다. 그러나 데이터를 확인한 결과 전자문서에 요약을 기록하지 않는 경우가 생각보다 많아 변수에서 제외했고, 단위과제카드명만 포함하여 학습하였다. A부서의 정확도는 80%로 감소하여 단위과제카드가 업무성격과 연관성이 없는 변수로 실험 결과, 확인되었다. 세 번째 실험은 제목과 본문, 첨부 파일 명을 포함하여 실험하였다. 정확도가 첫 번째 제목만으로 입력 변수 값을 설정한 실험에 비해 17% 감소, 정확도 68%로, 비 식별화 기호, 수·발신자 등 비 식별화 된 단어가 많아 오류를 발생시키는 걸로 확인하였다. 이런 결과를 확인하고 본문의 전처리를 한 뒤 다시 학습하여 전처리 전보다 정확도가 증가한 74%로 학습 결과가 나타났다. 또한 직무에 특화된 형태소 분석사전(불용어, 동의어 등)을 구축할 경우 정확도 개선 향상에 대한 가능성을 발견하였다. 그리고 본문전처리 중요 첨부 파일 명만 파싱하여 분석하려고 했으나 첨부 파일 명만을 추출하기 어려웠다. 첫 번째부터 세 번째 실험까지 한 결과 첫 번째 실험인 문서제목만을 입력변수로 처리하는 분류모델이 가장 정확도가 높았다. 이 학습모델에 대한 검증을 위해서 4번째 실험을 진행하였다. 네 번째 실험은 문서 제목만으로 학습한 분류모델을 검증하기 위해 18년도 데이터를 이용하였다. 의미 없는 것으로 확인된 1건을 제외한 2014~2017년 학습데이터 7805건으로 학습 사전에 담당자가 18년도 데이터를 정답 셋으로 라벨링을 하여 별도로 보관하고, 실험에는 라벨링 없는 데이터(3724건)를 모두 검증데이터로 입력하였다.

표 4 문서제목으로 학습한 결과
Table 4. Learning results by record title

Job characteristics	Learning	Varification			Accuracy
	Learning data	Varification data	Learning model		
			Correct answer	Wrong answer	
Strategy	232	69	24	45	35%
Execution management	2425	753	660	93	88%
Securement and allocation of resources	48	13	2	11	15%
Regulation and supervision	0		0		
Judgement and examination	45	15	3	12	20%
Research and analysis	96	203	170	33	84%
Adjustment of disputes and conflicts	0		0		
Internal support and management	4960	898	805	93	90%
Total	7806	1951	1664	287	85%

표 5. 문서제목 분류모델로 18년도데이터 검증결과
Table 5. Learning results in 2018 by record title

Job characteristics	Learning	Varification			Accuracy
	Learning data	Varification data	Learning model		
			Correct answer	Wrong answer	
Strategy	234	139	49	90	35%
Execution management	3064	1636	1356	280	83%
Securement and allocation of resources	44	25	4	21	16%
Regulation and supervision	0		0		
Judgement and examination	45	17	0	17	0%
Research and analysis	892	354	315	39	89%
Adjustment of disputes and conflicts	0		0		
Internal support and management	3526	1563	1286	277	82%
Total	7805	3734	3010	724	81%

표 5와 같이 학습데이터는 7,805건이고 정확도는 81%이었다. 정확도 분석 결과 2018년에만 수행한 사업명 등 해당년도에만 관련된 특성 키워드가 있을 경우, 학습데이터(2014~2017)모델에서는 오 분류 가능성이 높아 새로운 키워드에 대한 정답 셋 모델이 추가로 필요함을 알 수 있었다.

표 6. 실험 결과
Table 6. Experimental results

Test method	Type	Data	Accuracy
1. Documnt title (+3 departments)	Learning	2014 ~ 2017	85% (81%)
2. Document title + name of unit task card			80%
3. Document title + body of the document + name of attached file			68%
4. Document title	Varification	2018	81%

실험결과를 정리하면 표 6과 같다. 첫 번째 실험에서 문서제목으로 학습하는 경우 86%로 가장 정확도가 높았으나, 3개 부서를 학습 하였을 때 부서별 편차가 일부 나타나는 결과를 알 수 있었다. 두 번째와 세 번째의 경우 다른 변수를 추가로 적용했으나, 문제제목만으로 할 경우에 비해 정확도가 떨어지고, 전처리 등 학습에 어려움이 발생되었다. 첫 번째 학습모델에 대한 검증을 위해 18년도 데이터를 검증데이터로 활용하여 실험한 결과 81%의 정확도가 나타났다.

Ⅳ. 결 론

이번 연구는 4개의 샘플부서를 대상으로 정부 주요직위 인사에 활용 중인 직무기술서 가이드에서 제시하는 부서별 9가지 업무성격을 분류기준으로 정하였다. 해당부서에서 14~17년 생산된 전자문서를 담당자가 사전에 라벨링을 하였고, 학습/검증 데이터 분할(학습 : 80%, 검증 : 20%)으로 지도학습의 대표적인 모델인 지도학습의 유형인 베이시안 확률기반의 나이브 베이즈 모델링으로 분류하였다. 실험결과 전자문서의 제목으로 학습했을 경우 가장 정확도가 높았음을 확인했다. 최종적으로 라벨링 된 18년 문서를 기 학습된 분류기에 적용하여 최종적인 정확도를 검증하였다. 담당자의 수작업으로 라벨링 한 2018년 결과와 비교해서 81% 정확도가 나타나 전자문서를 기반으로 한 직무분석에 대한 자동분류가 긍정적임을 알 수 있었다. 또한 정확도를 높이기 위한 노력의 일환으로 전자문서 작성 시 담당자의 정확한 업무 표현에 대한 안내와 정교한 라벨링 작업이 지속적으로 필요 할 것이다. 이러한 연구결과는 기존 방식에서 과감히 탈피, 직위·직무 성과중심의 인사관리체계로 전환하고, 적재적소 인사원칙을 구현할 수 있는 기반이 될 것으로 기대된다.

References

[1] Job description for ranking government official, Ministry of Personnel Management, 2017.

[2] Sunggwan Kang, Bongkyung Kwon, Cheolwoo Kwon, Sangmin Park, and Ilsoo Yun, "Development of Incident Detection Algorithm Using Naive Bayes Classification", The Korea Institute of Intelligent Transport Systems. Vol. 17, No. 6, 25-39, Dec. 2018.

[3] Ling Xiao Li and Siti Soraya Abdul Rahman, "Students' learning style detection using tree augmented naive Bayes", Royal society open science, 5: 172108.

[4] Nelishia Pillay, Rong Qu, Dipti Srinivasan, Barbara Hammer, and Kenneth Sörensen, "Automated Design of Machine Learning and Search Algorithms", IEEE Computational intelligence magazine, Vol. 13, No. 2, 16-17, May 2018.

[5] Jung-Been Yu, Min-Sik Shin, and Taekyoung Kwon, "Analysis of Research Trend on Machine Learning Based Malware Mutant Identification", Journal of The Korea Institute of Information Security and Cryptology, Vol. 27, No. 3, pp. 12-19. Jun. 2017.

[6] Seong Jun Bae and Hyung Seok Kim, "A Study on Performance Improvement by Answer Label Change in Supervised Learning", Korea Institute of

Communication Sciences, pp. 255-256, Jun. 2017.
[7] Sun-Iee Bang, Jae-Dong Yang, and Hyung-Jeong Yang, "Automatic Document Classification Based on k-NN Classifier and Object-Based Thesaurus", Journal of KISS : Software and Applications, Vol. 31, No. 9, 1204-1217, Sep. 2004.
[8] Hae Chan Sol Kim, Dae Jin An, Jin Hee Yim, and Hae-Young Rieh, "A Study on Automatic Classification of Record Text Using Machine Learning", Korean Society for Information Management, Vol. 34, No. 4, 321-344, Dec. 2017.
[9] Final Report on Consulting for Human Resources Policy Support Platform, Ministry of Personnel Management, 2018.

저자소개

강 은 숙 (Eun-Sook Kang)

2001년 8월 : 동국대학교 경영정보학과(경영정보석사)
2018년 3월 ~ 현재 : 목원대학교 지능정보융합과 박사과정
관심분야 : 지능정보기술, 머신러닝, 빅데이터, 텍스트마이닝, 자동분류

고 대 식 (Dae-Sik Ko)

1982년 2월 : 경희대학교 전자공학과 졸업(공학사)
1991년 2월 : 경희대학교 전자공학과(공학박사)
1994년 ~ 1995년 : UCSB Post-Doc
2011년 1월 ~ 2012년 12월 한국정보기술학회 회장
1989년 ~ 현재 : 목원대학교 전자공학과·지능정보융합과 교수
관심분야 : ICT융합, 사물인터넷, 신호처리