ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 28, No. 6, pp. 327-331, 2022. 6
https://doi.org/10.5626/KTCP.2022.28.6.327

등장인물 감성 네트워크에 기반한 소설 작품 군집화

(Character Sentiment Network based on Clustering for Novels)

박 명 건 † 박 성 홍 † 신 현 정 ††
(Myeonggeon Park) (Sunghong Park) (Hyunjung Shin)

요 약 본 연구에서는 기계학습을 이용하여 소설을 그룹핑하는 방법을 제안한다. 제안하는 방법에서는 소설에 등장하는 인물들을 네트워크로 구성하고 주인공과 동조 또는 대립하는 인물들 간의 감성점수의 흐름을 파악하여 소설들을 군집화 한다. 인물 네트워크의 노드인 인물들은 name entity recognizer을 통하여 추출하였고 엣지는 소설에 기술된 감성단어를 기반으로 감성분석에 의하여 가중치를 부여하였다. 인물 네트워크 상에서의 동조그룹과 대립 그룹의 분류는 signed graph clustering으로 수행하였다. 이를 바탕으로 소설 한 작품은 기승전결 각 단락별 한 쌍의 감성점수 벡터로 구성되어 총 여덟 튜플로 표현된다. 제안하는 방법은 22편의 소설을 군집화하는 데 적용하였으며 각 군집 별 특성을 프로파일링 한다.

키워드: 인물 네트워크, 벡터표현, 감성분석

Abstract In this study, we propose a method of clustering novels using machine learning. The proposed method is to organizes the characters in the novel into a network, and to clusters the novel by grasping the flow of emotional scores between the characters who agree with or oppose the main character. Characters, which are nodes in the character network, were extracted through a name entity recognizer, and edges were weighted through sentiment analysis based on emotional words depicted in the novel. The classification of the protagonist and antagonist groups in the character network was performed by signed graph clustering. Based on this, the novel was composed of a pair of emotional score vectors for each paragraph and is expressed as a total of 8 tuples. The proposed method was applied to 22 novels, and the characteristics of each cluster were analyzed.

Keywords: sentiment analysis, vector presentation, character network

· 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2021R1A2C2003474)
· 이 논문은 2020 한국소프트웨어종합학술대회에서 '등장인물 감성 네트워크에 기반한 소설 작품 군집화'의 제목으로 발표된 논문을 확장한 것임

† 비 회 원 : 아주대학교 인공지능학과 학생
kack7090@gmail.com
pshong513@ajou.ac.kr
†† 종신회원 : 아주대학교 산업공학과, 인공지능학과 교수(Ajou Univ.)
shin@ajou.ac.kr
(Corresponding author임)

논문접수 : 2021년 4월 22일
(Received 22 April 2021)
논문수정 : 2022년 3월 10일
(Revised 10 March 2022)
심사완료 : 2022년 3월 25일
(Accepted 25 March 2022)

정보과학회 컴퓨팅의 실제 논문지 제28권 제6호(2022. 6)

1. 서 론

소설(novel, narrative)은 사건을 미적으로 질서화 하여 통일적인 의미가 구현될 수 있도록 산문으로 서술한 서사 문예로, 인물·사건·배경을 구조적 기본 요소로 하는 허구적인 서사 문예이다. 그동안 소설에 대한 요약본이나 로드맵 혹은 비평은 주로 전문가들의 영역이었다. 그러나 근래 인터넷 소설을 비롯한 신간소설의 수는 끊임없이 증가하고 있으며 기존 작품들과 더불어 그 양이 방대하다. 이러한 상황에서 전문가들을 보조하기 위한 수단으로 기계학습 모델들을 활용하면 효율적일 것이다. 실제로, 기존 연구를 살펴보면, 소설의 장르 분류를 위해서 특성 추출 모델로 document-term matrix(DTM)[1,2], term frequency - inverse document frequency(TF-IDF)[3-5], word2vec [6,7] 등을 사용하였다. 소설을 분석하는 것은 문장(단어) 속에 함축되어 있는 등장인물의 감정 변화와 인물들 간의 관계 변화를 분석하는 것이다. 기존 연구에서는, 감정변화를 따라가기 위해서 문장에 등장하는 단어들의 감정 점수를 하나의 대푯값으로 표현하기도 한다. 예를 들어, 한 문장의 감정을 표현하기 위해서 문장 내 단어들의 긍정/부정 감성점수들을 평균내거나 빈도수로 결정한다. 그러나, 대부분의 기존 연구에서는 인물 간의 관계 분석 및 변화를 간과하고 있다. 소설에는 항상 주인공의 갈등구조가 있는데 이 때 주인공에 대하여 우호적인 인물들로 구성된 동조그룹과 적대적인 인물들로 구성된 대립그룹이 등장한다. 두 그룹을 기술하는 단어들의 감정을 평균 내거나 빈도수만을 고려한다면 소설의 갈등 흐름이 퇴색한다.

따라서, 본 연구에서는 소설에 등장하는 인물들을 네트워크로 구성하고 주인공과 동조 또는 대립하는 인물들 간의 감성점수의 흐름을 파악한다. 인물 네트워크의 노드인 인물들은 개체명 인식기(name entity recognizer)을 통하여 추출하고 엣지는 소설에 기술된 감정단어를 기반으로 감성분석에 의하여 가중치를 부여한다. 인물 네트워크 상에서의 동조그룹과 대립그룹의 분류는 부호 그래프 군집화(signed graph clustering)방법으로 수행한다. 이를 바탕으로 한 작품은 기승전결 각 단락별 한 쌍의 감성점수 벡터로 구성되어 총 여덟 튜플로 표현된다. 비정형 텍스트인 소설을 정형 벡터로 표현하면 기계학습을 활용할 수 있는 분야는 무궁무진하다. 소설의 장르를 분류할 수도 있고 표절 소설들을 파악할 수도 있으며, 새로운 소설을 창작하는 데도 도움이 될 수 있다. 본 연구에서는 제안한 방법을 적용하여 유사한 작품들을 군집화하는 사례를 제시한다.

2. 감성 네트워크 기반한 소설 특성 벡터 추출

인물 네트워크 기반 소설 군집화를 위해서 다음 단계를 순차적으로 진행한다: (1) 소설의 단락(Act) 구분 및 단락별 인물 네트워크 구성, (2) 인물 네트워크에서 주인공, 대립인물 그룹 생성, (3) 소설 별 벡터 정의 및 소설 군집화.

2.1 단락 별 인물 네트워크 구성

인물 네트워크 구성을 위해 문장 단위로 연산, 불용어 제거, 소문자 변환 등의 전처리를 진행한다. 먼저, 소설 전개에 따른 등장인물의 감정 변화를 구별하기 위해서 한 편의 소설을 기승전결 단락으로 나눈다[8]. 소설의 전체 문장을 기준으로 균등하게 네 단락으로 나눈다. 그리고 단락 별로 인물 네트워크를 구성한다. 인물 네트워크는 등장인물을 노드로, 등장인물 간의 관계를 엣지로 정의한다[9]. 등장인물을 추출하기 위해서 개체명 인식기(https://spacy.io/)를 적용한 후 후처리를 함께 진행한다. 각 등장인물은 단락 별로 하나의 감정 곡선을 가진다. 등장인물의 이름이 언급된 각 문장에서 감정 단어의 유무를 파악하고 해당하는 감정 점수를 감성 단어 사전(hedonometer)으로부터 추출한다[10]. 한 문장은 감정 곡선의 한 점으로 표현된다. 소설을 갈등의 흐름으로 볼 때, 갈등을 가장 잘 나타내 주는 값은 긍정 혹은 부정 점수가 높은 값이다. 따라서 긍정 혹은 부정의 값(-4~+4)들 중 절대값이 가장 큰 점수를 대푯값으로 정의한다. 엣지는 두 명 이상의 인물이 한 문장에서 함께 등장하면 관계가 있는 것이므로 정의하고 두 노드를 엣지로 연결한다[11]. 엣지는 두 인물이 우호적 또는 적대적 관계임에 따라 '+' 혹은 '-'부호를 갖는다. '+'엣지를 가지는 관계를 긍정, '-'엣지는 부정으로 본다. 문장 내 감정단어들 중 절대값이 큰 점수를 대푯값으로 정하고 그 부호를 따른다.

2.2 주인공 vs 대립인물 그룹 생성

소설의 등장인물 간 관계 변화를 파악하기 위해서 단락별로 구성된 인물네트워크를 주인공, 대립인물 중심으로 그룹을 나눈다. 그룹을 나누기 위하여 부호가 있는 그래프 군집화 방법을 사용한다[12]. 주인공은 가장 많은 문장에서 등장하는 인물로 정의하고, 대립인물은 주인공과 가장 많은 부정관계를 가지고 있는 인물로 정의한다. 마지막으로 부호가 있는 그래프 군집화는 같은 그룹 내 인물들의 긍정관계의 최대화와 그룹 간 인물들의 부정관계를 최대화하는 방향으로 최적화하는 방법이다.

2.3 소설 별 대표 벡터 정의

소설의 군집 별 특성을 파악하기 위해서 소설 별로 벡터를 추출하고 이를 사용하여 군집화한다. 하나의 소설에서 각 단락 별로 주인공 그룹과 대립인물 그룹이 추출되면, 그룹 내 등장인물들의 개별 감정 곡선을 사용하여 두 그룹의 감정 곡선을 정의한다. 이 두 감정 곡선은 벡터로 표현될 수 있고, 단락 별로 추출되기 때문에 한 작품마다 총 8개의 벡터가 추출하게 된다. 이 8개의 벡터를 이어서 붙여 소설의 대표 벡터로 정의한다. 그런데,

작품마다 문장 수가 다르기 때문에 각각 길이가 다른 벡터가 생성된다. 이렇게 길이가 다른 소설의 대표 감정곡선을 비교하기 위해서 dynamic time warping (DTW) 방법론을 사용한다[13]. 일반적으로 DTW는 2개의 길이가 다른 시퀀스 사이의 최적의 연결을 계산하는 방법이다. 마지막으로, K-평균 알고리즘(K-means clustering) 기법을 사용하여 수집된 22편의 소설을 군집화하였다.

3. 실 험

3.1 데이터

본 실험에서 소설의 군집화를 위하여 사용된 작품은 총 22편이다. 실험에 사용된 텍스트는 모두 Project Gutenberg에서 이용했다. 22편의 작품의 평균 문장은 110문장이고, 평균 단어의 수는 970단어이다. 작품 분석에 사용한 감성 단어 데이터는 총 10,222개의 단어이고 긍정(27%), 부정(29%)와 나머지 단어는 중립이다. 작품에 평균적으로 등장하는 인물의 수는 10명이다.

3.2 단락 별 네트워크에서 주인공 vs 대립인물 그룹화 결과

그림 1은 백설공주와 일곱 난쟁이를 부호가 있는 그래프 군집화를 통한 단락별 인물 네트워크의 그룹을 나눈 결과이다. 이 작품에는 총 17명의 등장인물이 있고, 소설 228 문장을 네 단락으로 나누면 각 단락은 58문장을 포함한다. 네트워크에서 파란색 노드는 주인공이 속한 그룹의 노드, 노란색 노드는 대립인물이 속한 그룹의 노드, 회색은 단락에 등장하는 인물이지만, 주인공 혹은 대립인물 그룹에 속하지 않는 인물을 나타냈다. 각 노드를 연결하는 엣지들 중 긍정관계는 파란색, 부정관계는 빨간색으로 나타냈다. 각 네트워크 밑에 있는 그래프는 주인공 그룹(파랑)과 대립인물 그룹(빨강)의 감정곡선을 그린 것이다. 그래프의 x축은 단락 별로 진행 순서에 맞는 문장 번호이다. y축에 인물이 등장하는 경우는 그 문장의 감정 점수를 나타내고, 인물이 등장하지 않는 문장의 값은 0으로 표기했다. 그림 1(a)의 네트워크는 소설의 가장 처음 단락이다. 사냥꾼과 왕비가 백설공주와 부정관계인 이유는 공주를 암살하려고 하기 때문이다. 그림 1(b)에서 사냥꾼이 백설공주를 살해하지 않고 놓아줌으로 두 인물의 관계는 긍정으로 바뀌었다. 하지만 사냥꾼은 여전히 여왕과 같은 군집에 속하였다. 이는 백설공주에게 선의를 베풀지만 사냥꾼은 여왕과 연관관계가 높기 때문이다. 그림 1(c)에서 난쟁이들과 백설공주의 부정관계에는 여왕으로 계략으로 인한 공주의 죽음에 슬퍼하는 난쟁이들의 슬픔이 반영되었다. 그림 1(d)에서는 소설이 해피앤딩으로 끝나기 때문에 인물들이 대립관계를 나타내지 않는다. 따라서 대립 인물 그룹의 감정 곡선은 0이 되었다.

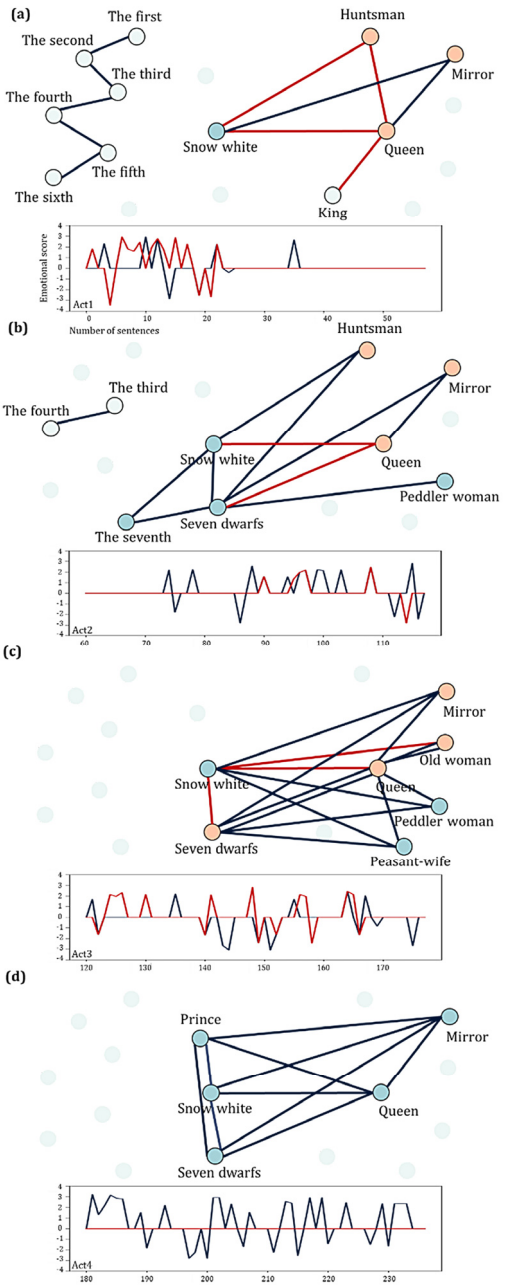


그림 1 백설공주와 일곱 난쟁이의 기승전결(a-d) 인물 관계 그래프

Fig. 1 The relationship between Snow White and the Seven Dwarfs (a-d)

3.3 소설 군집화 결과

그림 2는 주성분분석을 사용한 군집화 결과를 시각화하였고, 표 1에는 군집 별 작품을 표기하였다. 주성분분

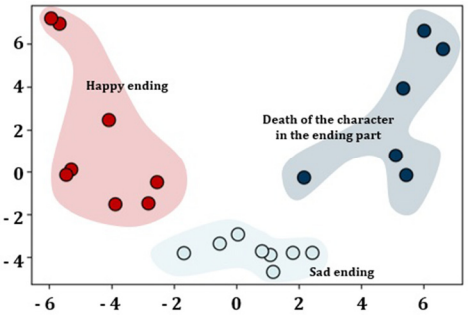


그림 2 소설 군집 결과: 군집1(navy), 군집2(red), 군집3(light-blue)

Fig. 2 Novel cluster results: cluster1(navy), cluster2(red), and cluster3(light-blue)

표 1 군집 별 작품

Table 1 Works by cluster

Clusters	Titles
1	Snow White and the seven dwarfs, Hansel and Gretel, The remarkable rocket, The devoted friend, The valiant little tailor, The fir tree
2	The fisherman and his wife, The happy prince, The swineherd, The old house, The elderbush, The emperor's new clothes, Ashputtel, The bell
3	The nightingale and the rose, The golden bird, Snow-White and Rose-Red, The juniper-tree, Iron Hans, The story of a mother, Tom Thumb, The selfish giant

석은 추출된 벡터를 사용하여 구한 거리/유사도 행렬을 사용하였다. 군집 형성에 가장 영향을 미치는 요인은 서술적인 요인과 기술적인 요인으로 나눠서 설명할 수 있다. 따라서 군집을 세 개로 나눈 이유도 소설의 결말에 따라서 정석으로 분석하여 나눈 것이다. 군집1 소설의 서술적인 특징은 주인공이 대립인물에 의해서 겪게 되는 어려움을 극복하고 행복한 결말을 맞는 것이다. Hansel and Gretel을 예로 들면, Hansel과 Gretel은 계모와 마녀에 의해서 겪게 되는 위협을 극복하고 아버지와 행복한 결말을 맞이한다. 군집2의 작품들은 주인공이 과한 욕심에 때문에 맞게 되는 비극적인 결말이다.

The fisherman and his wife에서 어부와 부인은 구해준 물고기에 의해서 부와 명예를 모두 얻게 되지만, 끝없는 욕심으로 모든 것을 잃게 된다. 군집3은 결말 부분에서 주인공 혹은 대립인물이 죽는다. Snow-White and Rose-Red는 결말 부분에 욕심쟁이 난쟁이를 죽이면서 마법에 걸렸던 왕자가 원래 모습으로 돌아온다. 22편의 소설은 대부분이 아이들을 위한 이야기들이기 때

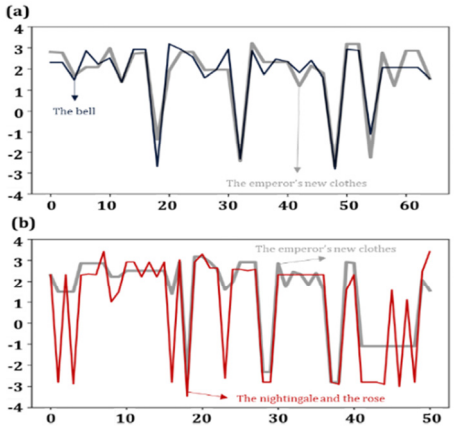


그림 3 군집화를 위한 소설간 유사도 비교 예시

Fig. 3 Example of comparison of similarity between novels for clustering

문에 대부분 긍정적인 단어로 표현되었다. 따라서 기술적인 특징은 부정단어의 평균 등장 횟수가 된다. 군집1은 부정단어가 평균 30회 등장, 군집2는 11회, 마지막으로 군집3은 20회 등장한다. 그림 3은 소설 간의 유사도를 비교하기 위해, 군집2에서 The emperor's new clothes(회색)과 The bell(파란색)을 군집3에서 The nightingale and the rose(빨간색)의 벡터를 비교하였다. 실험 결과, 부정적인 단어의 등장 패턴이나 횟수가 군집화에 중요한 영향을 미친다는 것을 알았다. 그림 2(a)는 부정 단어가 비슷한 패턴으로 발생하며 횟수도 일치하는 패턴을 볼 수 있다. 그림 2(b)는 두 작품의 부정어의 출현 횟수와 패턴 또한 다른 것을 보인다.

3.4 비교 실험

제안하는 방법의 성능을 검증하기 위하여 기존의 텍스트 분류 방법들과 비교실험을 진행하였다. 실험에 사용된 20편의 소설들은 감정의 흐름에 따라 6가지로 분류된 레이블을 가지고 있다[14]. DTM, TF-IDF, word2vec, proposed method를 사용하여 텍스트를 벡터로 추출하고, 이를 사용하여 군집화를 실시하였다. 레이블이 있는 텍스트를 사용하였기 때문에 군집의 지니 불순도(gini impurity)를 계산할 수 있다. 지니 불순도의 측정은 기존

표 2 군집 개수 변화에 따른 지니 불순도(Gini impurity)

Table 2 Gini impurity according to the change in the number of clusters

군집의 개수	2	3	4	5
Proposed method	0.13	0.06	0.15	0.13
TF-IDF	0.21	0.21	0.21	0.14
Word2vec	0.24	0.26	0.21	0.19
DTM	0.15	0.13	0.17	0.07

의 분류되어진 레이블과 같게 분류하는 것이 아닌, 추출된 벡터를 사용하여 같은 특성을 가진 데이터들이 군집에 모여 있는 정도를 평가한다. 지니 불순도는 0에 가까울수록 분류가 잘 되었다고 판단할 수 있다. 표 2를 근거로 제안하는 방법이 문학 텍스트를 기존의 방법론보다 유의미하게 분류하는 것으로 볼 수 있다.

4. 결 론

본 연구에서는 소설 속 인물들의 감정 변화와 인물들 간의 관계 변화를 반영한 벡터표현 방법을 제시하였다. 소설에 등장하는 인물들을 네트워크로 구성하고, 주인공과 동조 또는 대립하는 인물들 간의 감성 점수의 흐름을 파악하였다. 제안하는 방법론을 군집화에 적용한 결과 22편의 소설을 효과적으로 세 개의 군집으로 나눌 수 있었다. 군집 별로 공통된 서술적, 기술적 특징을 지닌다. 추출된 벡터는 표절 확인, 장르 분류, 작품 창작 등에 활용할 수 있을 것으로 기대된다.

References

[1] Silge, Julia, and David Robinson, *Text mining with R: A tidy approach*," O'Reilly Media, Inc.," 2017.

[2] Afzali, Maedeh, and Suresh Kumar, "Text Document Clustering: Issues and Challenges," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, IEEE, 2019.

[3] Aizawa, Akiko, "An information-theoretic perspective of tf-idf measures," *Information Processing & Management* 39.1 (2003): 45-65.

[4] Bengfort, Benjamin, Rebecca Bilbro, and Tony Ojeda, *Applied text analysis with python: Enabling language-aware data products with machine learning*," O'Reilly Media, Inc.," 2018.

[5] Qaiser, Shahzad, and Ramsha Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents," *International Journal of Computer Applications,* 181, 1, (2018): 25-29.

[6] Budiman, Irwan, et al., *"A Study on Effect of Generated Features From Word2Vec Vectors For Text Classification"*

[7] Jang, Beakcheol, Inhwan Kim, and Jong Wook Kim, "Word2vec convolutional neural networks for classification of news articles and tweets," *PLoS ONE*, 14, 8 (2019): e0220976.

[8] Bryan Thomas Schmidt, *How to write a novel: The fundamentals of fiction*, "Independently published," 2019.

[9] Elson, David, Nicholas Dames, and Kathleen McKeown, *"Extracting social networks from literary fiction,"* 2010.

[10] Dodds, Peter Sheridan, et al., "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter," *PloS one*, 6, 12, (2011): e26752.

[11] Park, Gyeong-Mi, Sung-Hwan Kim, and Hwan-Gue Cho, "Structural analysis on social network constructed from characters in literature texts," *Journal of Computers*, 8, 9, (2013): 2442-2447.

[12] Traag, Vincent A., and Jeroen Bruggeman, "Community detection in networks with positive and negative links," *Physical Review E*, 80, 3, (2009): 036115.

[13] Sakoe, Hiroaki, and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, 26, 1, (1978): 43-49.

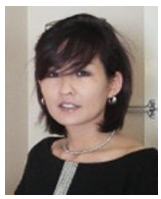
[14] Reagan, Andrew J., et al., "The emotional arcs of stories are dominated by six basic shapes," *EPJ Data Science,* 5, 1, (2016): 1-12.

박 명 건

2016년 연세대학교 영어영문과 졸업(학사). 2022년 아주대학교 인공지능학과 졸업(석사). 관심분야는 자연어 처리, 데이터 마이닝, 텍스트 마이닝, 기계 학습

박 성 홍

2016년 아주대학교 산업공학과 졸업(학사). 관심분야는 준지도학습, 도메인 적응, 그래프 기반 알고리즘, 바이오메디컬 인포매틱스



신 현 정

2005년 서울대학교 산업공학과 졸업(박사) 2004년~2005년 Max-Planck-Institute for Biological Cybernetics 독일(연구원) 2005년~2006년 Friedrich-Miescher-Laboratory, Max-Planck-Institute 독일(수석 연구원). 2006년 서울대학교 의과대학 연구교수. 2006년~현재 아주대학교 산업공학과, 데이터사이언스학과, 융합시스템공학과 교수. 관심분야는 머신러닝, 데이터 마이닝, 바이오메디컬 인포매틱스