

# 지역 꽃 축제 개선사항 도출을 위한 뉴스 데이터 분석 연구

이정원<sup>1</sup>, 이충호<sup>2\*</sup>

<sup>1</sup>한밭대학교 정보통신공학과 박사과정, <sup>2</sup>한밭대학교 정보통신공학과 교수

## A Study on the Analysis of News Data for the Improvement of Local Flower Festival

Jeongwon Lee<sup>1</sup>, Choong Ho Lee<sup>2\*</sup>

<sup>1</sup>Doctoral Student, Dept. of Information and Communication Engineering, Hanbat National University

<sup>2</sup>Professor, Dept. of Information and Communication Engineering, Hanbat National University

**요약** 지역의 관광산업은 지역경제 활성화 및 지역 이미지 개선의 효율적인 수단이다. 이것을 활성화 시키기 위하여 지역적으로 특화된 관광 상품을 만들고, 고유문화와 전통을 살리기 위한 노력이 필요하다. 이 중 관광객들에 대한 정보 수집 및 관광 콘텐츠 내용의 질적 경쟁력 확보는 문화관광축제의 잠재력을 키우는데 매우 중요하다. 본 논문은 방문객의 관광 욕구 및 만족도를 제고하기 위하여 특정 지역의 축제 관련 데이터를 수집, 정제, 처리하여 이 데이터와 관광산업과의 연관성을 시각적으로 표현하고자 한다. 특히 지역에서는 축제에 대한 분석 의지는 높으나 데이터 수집을 위한 별도의 인프라가 없는 상태에서 외부 통신사 및 카드사의 데이터에 의존 할 수밖에 없었다. 이로 인한 행정 예산의 부담이 가중되고 있었으며, 이를 조금이나마 해소하고자 공공데이터 및 인터넷에 존재하는 데이터들을 수집하여 분석하고자 하였다. 본 연구에서는 분석하고자 하는 지역의 공공데이터 및 주변에서 쉽게 접할 수 있는 지역 뉴스 데이터를 활용하여 지역 축제에서 이슈가 되었던 부분을 키워드로 도출할 수 있는지에 목적을 두고 연구하였다.

**키워드** : 문화, 관광, 축제, 문제점, 의사결정

**Abstract** Regional tourism is an effective means of revitalizing the local economy and improving the image of the region. In order to revitalize this, efforts should be made to create regionally specialized tourism products and to preserve the unique culture and traditions. Among them, gathering information about visitors and securing the quality competitiveness of the contents of tourism contents are very important to increase the potential of cultural tourism festival. This paper collects, refines, and processes the festival-related data in a specific area in order to enhance the visitor's tourism needs and satisfaction. In particular, negative words and positive words raised during the festival were analyzed through big data visualization using word cloud.

**Key Words** : Culture, Tourism, Festivals, Issues, Decision Making

### 1. 서론

#### 1.1 연구의 배경과 내용

지역축제는 무형의 관광자원으로서 지역경제 활성화

및 지역 이미지 개선의 효율적인 수단 중 하나이기 때문에 이에 대한 관심이 증대되고 있다. 이것을 발전시키기 위하여 특화된 관광 상품을 개발하는 것, 지역 고

유의 전통을 살리는 일, 축제 개최의 지원 및 육성정책을 시행함으로써 지역 관광산업이 발전되고 있다. 그러나 현재의 문화관광축제는 양적인 성장에도 불구하고 전문 축제 기획, 운영 인력 부족, 관람객에 대한 정보 수집 및 분석 부족으로 질적인 성장이 충분히 이루어지지 않고 있으며 나아가 세계적인 축제로 발전되지 못하고 있다[1,2].

각 지역별로 축제에 대한 관심이 매우 높은 가운데 축제 이후 어떠한 문제점들이 있었는지, 개선할 점은 무엇인지 등에 대하여 데이터분석을 통하여 도출하고자 많은 노력을 기울이고 있다. 그러나 이러한 데이터를 분석하는 중요한 데이터 수집 인프라가 없기에, 대부분의 지역에서는 각 축제 지역 내 통신사 및 카드사 데이터를 높은 비용을 들여서 데이터를 구매하여 분석을 하고 있는 실정이다.

본 논문은 높은 비용을 들이지 않고 일반적인 인터넷에 존재하는 데이터 및 공공데이터들을 융통성 있게 사용함으로써 방문객의 관광 욕구 및 만족도를 도출해 보고자 본 연구에서는 지역 뉴스를 활용 하였다.. 특정 지역에 관한 축제 관련 뉴스 데이터를 수집, 정제, 처리하여 이 데이터와 관광산업과의 연관성을 분석한다. 특히 축제 중 제기된 부정적인 단어 및 긍정적인 단어들을 워드클라우드를 이용한 빅데이터 시각화를 통하여 도출 가능한지에 대하여 연구 하였다.

### 1.2 연구 목적

본 연구의 목적은 상용 데이터의 활용에서 벗어나 지역 공공데이터 및 지역 축제를 보도하는 온라인 매체 데이터의 텍스트를 빅데이터 분석 기법으로 분석하여 관광산업에 부정적 또는 긍정적 연관성을 가지는 이슈들을 시각적으로 도출하여 직관적인 의사결정이 될 수 있도록 하는데 있다.

## 2. 이론적 배경

타 지역 관광객들을 유치하고 이를 통해서 각 지방의 관광 활성화를 통한 균형적인 발전을 도모하기 위하여 정부는 지역 문화 축제를 집중적으로 육성하고 있다[3, 4]. 이런 정책을 위한 데이터 분석을 위하여 기존의 부족한 내·외국인 관광객 통계 집계 방식을 벗어나, 공공·민간 데이터 기반의 관광객 현황을 객관적으로 파악하여

지역별 관광 활성화를 위한 맞춤형 관광정책 수립을 지원하고 있다[5]. 또한 축제 관련 주요 민간데이터를 종합적으로 분석하여, 축제 방문객 현황 및 경제 효과를 객관적으로 파악하고, 축제 활성화를 위한 기초 자료를 수집하고 분석하는데 노력을 기울이고 있다[6]. 이것들은 축제 운영 측면에서 방문객 관광 편의 시설과 지역 고유의 문화를 방문객들이 정확히 인식할 수 있도록 프로그램화 하는데 매우 중요하다[7,8].

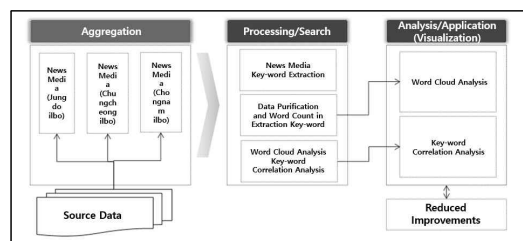
양적인 성장만이 아닌 경쟁력 있고 실질적으로 외국 관광객 유치를 통한 지역경제 활성화와 지역문화의 전통을 계승 발전시킬 수 있는 문화관광축제의 내실화에 정진해야 할 시점에서 시대 흐름에 맞는 데이터 기반 행정서비스 개선을 위한 노력에 집중하기 위하여 정형 또는 비정형 데이터 분석을 위한 빅데이터 분석 방법 적용이 필요하며[9, 10], 이를 위한 분석 프로그램의 활용이 요구된다.

## 3. 분석 방법

본 연구의 분석을 위하여 부여군의 서용연꽃축제를 선정하여 관련 온라인 매체 비정형 공개데이터를 수집하여, 아래와 같이 분석을 수행하였다.

### 3.1 분석 절차

축제 관련 분석에 사용된 데이터는 Fig. 1과 같이 중도일보, 충청일보, 충남일보에 나타난 텍스트이다. 원데이터는 3개의 지역 신문 매체로부터 수집한다. 수집된 데이터는 처리/탐색 단계를 거친다. 이 단계에서는 축제 관련 기사에서 주요한 키워드의 빈도수를 계산하고 키워드 간에 주요 연관단어를 수집한다. 최종적으로 분석/응용(시각화)단계에서 워드클라우드로 표현하고 키워드 간 연관성을 분석한다. 또한 이 분석을 통하여 개선사항을 도출한다.



[Fig. 1] Flower Festival Data Analysis Procedure

### 3.2 분석 도구

#### 3.2.1 수집 데이터

본 연구를 위하여 Fig. 2와 같이 부여서동연꽃축제 온라인 매체인 지역 주요 신문사 뉴스인 중도일보, 충청일보, 충남일보 등 비정형 뉴스 데이터를 수집하였다.



[Fig. 2] Online Collection Data

#### 3.2.2 분석 환경

분석을 위한 운영체제는 윈도우(Windows), 분석 프로그램은 R 프로그램, 활용 데이터는 비정형 수집데이터를 활용 하였다.

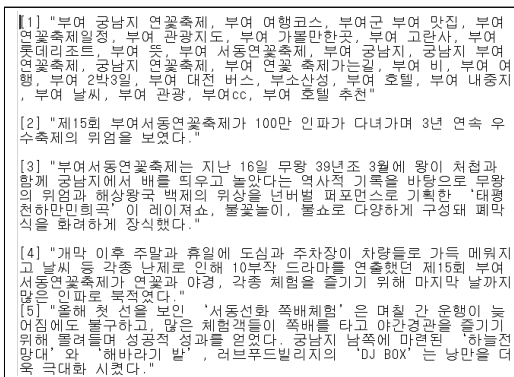
### 3.3 분석 방법

분석 내용으로는 부여서동연꽃축제 온라인 매체 뉴스 데이터 등 비정형 데이터를 수집하여 워드클라우드 분석을 통한 이슈 키워드 도출을 위한 연관성을 분석을 수행하였다.

## 4. 분석 결과

### 4.1 비정형 수집 데이터 전처리

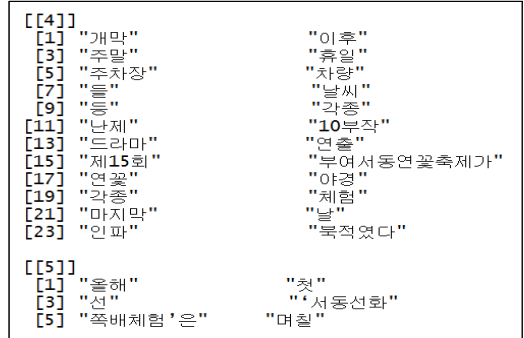
Fig. 3과 같이 중도일보, 충청일보, 충남일보 등 지역 중심 온라인 뉴스 데이터를 수집하여 문단별 데이터를 분리 한다.



[Fig. 3] Atypical Collection Data Preprocessing

### 4.1.1 데이터 키워드 추출

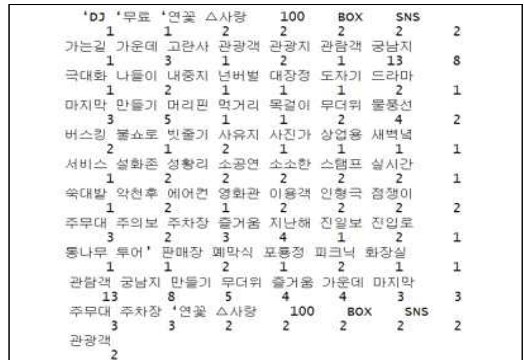
Fig. 4는 수집된 원천 데이터로부터 키워드를 추출한다.



[Fig. 4] Keyword Extraction From Collection Data

### 4.1.2 데이터 처리

Fig. 5과 같이 추출된 키워드들은 1차적으로 무의미한 단어를 정제하고 단어의 빈도수를 계산(wordcount)하여 빈도수 순위대로 키워드 순위를 결정한다.



[Fig. 5] Collective Data Refinement and Wordcount

### 4.1.3 데이터 시각화

워드클라우드는 글자가 클수록 많이 언급된 단어이다. Fig. 5와 Fig. 6에서 관련 긍정적 키워드로는 '관람객', '궁남지', '연꽃', '축제', '체험', '즐거움' 등이 많은 부분을 차지하고 있으며, 부정적 키워드로는 많은 관람객이 참여한 축제인 만큼 '주차장', '푸드', '문제' 등의 단어가 일부 나타나고 있다. 또한, 여름철 계절의 영향으로 인하여, 장마의 영향에 따른 무더위 및 집중호우에 따른 '날씨', '무더위' 등의 키워드가 이슈화 된 것으로 보아, 축제 동안 '날씨'의 부정적 영향을 많이 받은 것으로 나타나고 있다. Fig. 7은 불필요한 단어를 정제하는 R프로그램 소스의 주요한 부분이다.



[Fig. 6] Collection Data Word Cloud Visualization

```
word_1 <- sapply(text_1, extractNoun,
                USE.NAMES = F)
vector_1 <- unlist(word_1)
search_length <- Filter(function(x)
                        {nchar(x) == 3}, vector_1)
head(search_length, 100)
text_table_1 <- table(text_new_1)
text_table_1
#head(sort(text_table_1, decreasing=T), 50)
```

[Fig. 7] Collected Data Refinement and Keyword Tracking Code

```
wordcloud(names(text_table_1),
          freq = text_table_1,
          scale = c(6, 1),
          rot.per = 0.5,
          min.freq = 3,
          random.order = F,
          random.color = T,
          colors = select_color)
```

[Fig. 8] Collection Data Word-Cloud Visualization Code

## 4.2 연관성 분석

### 4.2.1 이슈 키워드 도출

수집된 원천 데이터로부터 연관성 있는 키워드를 추출한다. Fig. 9에 결과의 일부를 나타내었다.

### 4.2.2 데이터 정제 및 가공

추출된 키워드들에서 1차로 무의미한 단어를 정제하고 연관성 워드카운트(wordcount)과정을 통하여 빈도수가 높은 순위대로 순위를 결정한다. Fig. 10에 그 결과를 일부 나타내었다.

[[13]] [1] "{년버벌}" "{물쇼로}"	[[23]] [1] "{마지막}" "{주차장}"
[[14]] [1] "{물쇼로}" "{년버벌}"	[[24]] [1] "{주차장}" "{마지막}"
[[15]] [1] "{년버벌}" "{폐막식}"	[[25]] [1] "{관람객}" "{공남지}"
[[16]] [1] "{폐막식}" "{년버벌}"	[[26]] [1] "{공남지}" "{관람객}"
[[17]] [1] "{물쇼로}" "{폐막식}"	[[27]] [1] "{관람객}" "{먹거리}"
[[18]] [1] "{폐막식}" "{물쇼로}"	[[28]] [1] "{먹거리}" "{관람객}"
[[19]] [1] "{드라마}" "{마지막}"	[[29]] [1] "{공남지}" "{먹거리}"
[[20]] [1] "{마지막}" "{드라마}"	[[30]] [1] "{먹거리}" "{공남지}"
[[21]] [1] "{드라마}" "{주차장}"	[[31]] [1] "{나들이}" "{점쟁이}"
[[22]] [1] "{주차장}" "{드라마}"	[[32]] [1] "{점쟁이}" "{나들이}"

[Fig. 9] Derivation of Issue Keywords for Analysis of Relation

관람객	1	1	0	0	0	0	1	0	0
공남지	1	1	0	0	0	0	1	0	0
나들이	0	0	1	0	0	0	0	0	1
년버벌	0	0	0	1	0	0	0	1	0
드라마	0	0	0	0	1	1	0	0	0
마지막	0	0	0	0	1	1	0	0	0
먹거리	1	1	0	0	0	0	0	1	0
물쇼로	0	0	0	1	0	0	0	1	0
점쟁이	0	0	1	0	0	0	0	0	1
주차장	0	0	0	0	1	1	0	0	0
즐거움	0	0	1	0	0	0	0	0	1
폐막식	0	0	0	1	0	0	0	1	0
관람객	0	0	0	0	0	0	0	0	0
공남지	0	0	0	0	0	0	0	0	0
나들이	0	1	0	0	0	0	0	0	0
년버벌	0	1	0	1	0	0	0	0	0
드라마	1	1	0	0	0	0	0	0	0
마지막	1	1	0	0	0	0	0	0	0
먹거리	0	0	0	0	1	0	0	0	0
물쇼로	0	0	0	1	0	0	0	0	0
점쟁이	0	1	0	0	0	0	0	0	0
주차장	1	0	0	0	0	0	0	0	0
즐거움	0	1	0	0	0	0	0	0	0
폐막식	0	0	1	0	0	0	0	1	0

[Fig. 10] Word Count Keyword Ranking for Correlation Analysis

### 4.2.3 데이터 연관성 분석

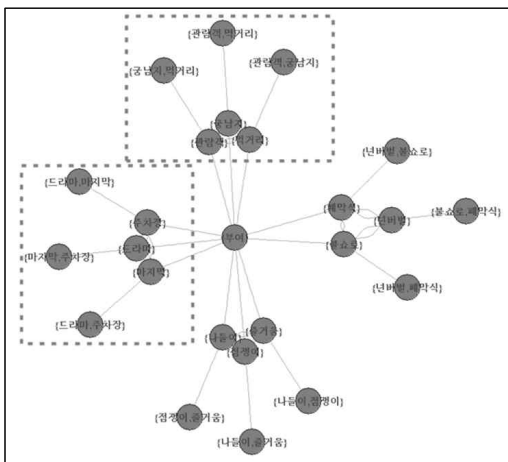
데이터 간의 연관성을 분석하기 위하여 지지도(0.1) 및 신뢰도(0.2) 기준의 연관성 데이터 분석을 수행하여 키워드간 연관성을 Fig. 11과 같이 도출하였다.

[1] {부여} --{물쇼로} {부여} --{폐막식} {부여} --{드라마}
[4] {부여} --{마지막} {부여} --{주차장} {부여} --{관람객}
[7] {부여} --{공남지} {부여} --{먹거리} {부여} --{나들이}
[10] {부여} --{점쟁이} {부여} --{즐거움} {물쇼로}--{년버벌}
[13] {물쇼로}--{년버벌} {폐막식}--{년버벌} {폐막식}--{년버벌}
[16] {물쇼로}--{폐막식} {물쇼로}--{폐막식} {드라마}--{마지막}
[19] {드라마}--{마지막} {드라마}--{주차장} {드라마}--{주차장}
[22] {마지막}--{주차장} {마지막}--{주차장} {관람객}--{공남지}

[Fig. 11] Correlation Analysis between Keywords

## 4.3 데이터 시각화

연관성 분석한 결과를 시각화하기 위한 네트워크 차트의 경우 부여 축제를 통해 이슈화된 키워드 간의 연관성을 분석하여 시각화하였고 이것을 Fig. 12에 보였다. 워드클라우드에서 이슈가 되었던 특정 키워드에 대하여 연관성 있는 단어들을 시각적으로 표현할 수 있다. Fig. 13은 연관성을 분석하기 위한 주요한 소스코드의 일부이다.



[Fig. 12] Data Visualization Network Char

```
ares <- apriori(wordtran,
  parameter = list(supp=0.1, conf=0.2))
inspect(ares)
rules <- labels(ares, ruleSep=" ")
rules <- sapply(rules, strsplit, " ",
  USE.NAMES = F)
rulemat <- do.call("rbind", rules)
```

[Fig. 13] Data Association Rule Analysis Code

```
ruleg <- graph.edgelist(
  rulemat[-c(1:1),], directed = F)
plot.igraph(ruleg, vertex.label=V(ruleg)$name,
  vertex.label.cex=1.0,
  vertex.size=13,
  layout=layout.fruchterman.reingold.grid,
  vertex.label.font=7,
  edge.color="darkgrey",
  vertex.color="yellow")
```

[Fig. 14] Correlation Analysis Visualization Code

### 5. 결론

본 연구는 지역의 높은 분석 의지와 데이터 수집을 위한 별도의 인프라가 없는 상태에서 외부 통신사 및 카드사의 데이터에 의존 할 수밖에 없었던 단점을 극복해 보고자, 인터넷에 존재하는 데이터들을 수집하여 주요 키워드를 도출하고자 하였다. 워드클라우드 분석을 통하여 부여 지역의 축제 정보를 중심으로 긍정적 또는 부정적 키워드의 빈도수와 연관성을 분석하였다. 분석 결과, ‘나들이’, ‘즐거움’ 등의 긍정적 키워드가 높게 나타나 축제에 대하여 호감을 느끼는 참여자가 많이 도출된 축제에 분석되고 있으나, 연관성 분석을 통한 워드클라우드에서 도출된 부정적 키워드들로 ‘주차장’, ‘먹거리(푸드)’, ‘무더

위’ 등이 나타나 부정적 이슈도 등장하는 것으로 분석되었다.

주요 이슈 중 ‘주차장’의 경우 매년 많은 인파가 몰리고 있어 이에 대한 주차장 부족에 따른 불편사항이 높게 나타난 것으로 해석될 수 있으며, ‘관람객이 많이 위치한 지역’ 및 ‘정확한 차량 현황’을 확인 할 수 있는 공공데이터가 부족하기 때문에 실질적 현장 데이터 수집을 위한 방안 마련이 필요할 것으로 분석되었다.

“먹거리” 이슈의 경우 먹거리 판매 부스의 위치가 관람객이 주로 이용하는 체험장 주변 또는 관람객 편의에 중점을 둔 위치 배정이 아님에 따른 불편함을 나타내고 있는 것으로 분석되었으며, 이에 대하여 [관람객의 이동 현황]을 확인 할 수 있는 공공데이터가 부족하기에 실질적 현장 데이터 수집을 위한 방안 마련이 필요할 것으로 분석되었다.

본 연구의 분석과정과 결과는 지역축제의 긍정적 또는 부정적 이슈들을 수집하고 분석하여, 행정서비스 개선의 시작을 위한 주요 키워드로 활용될 수 있을 것으로 기대된다. 보다 효과적이고 정확한 분석을 위해서는 축제관련 현장 데이터 수집을 위한 개선된 인프라 도입이 요구된다.

### REFERENCES

- [1] D. Hee. Choi. (2018). The Effects of Recognition on Cultural Tourism Festivals on Festival Satisfaction and Festival Effects. *Korean Journal of Korea Convergence Society*, 9(10). 339-346. DOI : 10.15207/JKCS.2018.9.10.339
- [2] W. S. Shim. (2015). Cultural Fusion Through the Linkage of Culture and Tourism. *Korean Journal of Korea Tourism Policy*, 62. 78-84.
- [3] H. S. Seo. (2000). The Impacts of Physical Environments on Iksan Jewelry Festivals' Satisfaction and Revisit, Word of Mouth. *Korean Journal of Korean Public Administration Review*, 34(1). 229-243.
- [4] Jenny (Jiyeon) Lee. (2014). Visitors' Emotional Responses to the Festival Environment. *Korean Journal of Travel & Tourism Marketing*, 31(1), 114-131.
- [5] J. H. Kim & H. G. Kim. (2010). The Impact of Culture Resources on City Brand Personality, Relationship Quality, and Loyalty in Tourism City. *Korean Journal of The Korean Society For Emotion & Sensibility*, 13(4). 741-752.

- [6] J. G. Park, T. Y. Cho & J. Y. Lee. (2011). Research on how Environmental Cues of and Sightseeing Experience are Affecting on Aftermath of Local Festival. *Korean Journal of The Academy of Korea Hospitality & Tourism*, 13(1), 22-35.
- [7] H. C. Uk & J. D. Kim. (2007). Enjoyment of Authenticity and Amusement from Local Residents and Tourists towards Cultural Tourism Festival. *Korean Journal of Korea Tourism Research Association*, 21(4), 85-99.
- [8] W. J. No & H. S. Oh. (2015). A study on the Local Governance organization of Cultural Tourism Festival: Focused on case of Ganggyeong Fermented Seafood Festival. *Korean Journal of The Korea Academic Society Of Tourism And Leisure*, 27(7), 277-298.
- [9] S. K. Choi, J. G. Hoan & M. J. Im. (2017). A Study on Impacts Perception of the Tourism Festival on Urban Regeneration. *Korean Journal of Korean Tourism Industry Research Association*, 42(1), 125-145.
- [10] J. K. Kim & J. G. Hoan. (2015). Effects of Festival Service Quality on Visitor's Satisfaction and Festival Performance. *Korean Journal of Korea Society of Culture Industry*, 15(3), 93-101.

이 정 원(Jeongwon Lee)

[정회원]



- 2006년 2월 : 한양사이버대학교 컴퓨터공학과(공학사)
- 2009년 2월 : 공주대학교 교육대학원 컴퓨터교육과(교육학석사)
- 2014년 8월 : 공주대학교 대학원 컴퓨터교육과(박사수료)

- 2018년 3월 ~ 현재 : 한밭대학교 정보통신전문대학원 정보통신공학과(박사과정)
- 관심분야 : 빅데이터분석, 인공지능, 데이터베이스, 응용소프트웨어, 경영정보
- E-Mail : mentor1023@daum.net

이 중 호(Choong Ho Lee)

[정회원]



- 1985년 2월 : 연세대학교 전자공학과(공학사)
- 1987년 2월 : 연세대학교 대학원 전자공학과(공학석사)
- 1998년 3월 : 도호쿠대학 대학원 정보과학연구과(공학박사)

- 1987년 2월 ~ 2000년 2월 : KT 멀티미디어연구소 전임연구원
- 2000년 2월 ~ 현재 : 한밭대학교 정보통신공학과 교수
- 관심분야 : 디지털신호처리, 영상처리, 패턴인식, 인공지능, 응용소프트웨어
- E-Mail : chlee@hanbat.ac.kr