단어-역문서 빈도 벡터화를 통한 한국 걸그룹의 음반 메타 정보 군집화

[1] 현준서, [2*]조재혁

Clustering Meta Information of K-Pop Girl Groups Using Term Frequency-inverse Document Frequency Vectorization

[1]JoonSeo Hyeon, [2*]JaeHyuk Cho

요 약

2020 년대 K-Pop 시장은 보이그룹보다 걸그룹이, 3 세대보다 4 세대가 전반에서 주목받았다. 해당 논문은 걸그룹의 세대가 바뀌기 시작했는지 알아보고자 가사 군집화에 대한 방법과 결과를 제시한다. 2013 년부터 2022 년까지 발표된 47 개 그룹의 1469 곡에 대한 메타정보를 수집하여 가사 정보와 가사 외 메타정보로 분류하여 각각 수치화했다. 가사 정보는 선행연구를 기반으로 단어-역문서 빈도 벡터화를 적용한 뒤 상위 벡터 값만 선정하는 전처리를 하였다. 가사 외 메타정보는 가사 정보만 사용했을 때의 편향성을 줄이고 더 좋은 군집화 결과를 보여주기 위해 One-Hot Encoding 으로 전처리하여 적용했다. 전처리된 데이터에 대한 군집화 성능은 Spherical K-Means 의 Silhouette Coefficient, Calinski-Harabasz Score 가 Hierarchical Clustering 에 비해 각각 129%, 45% 더 높았다. 본 연구는 한국 대중가요 발전사와 걸그룹 가사 분석 및 군집화 연구에 기여할 수 있을 것으로 기대된다.

Abstract

In the 2020s, the K-Pop market has been dominated by girl groups over boy groups and the fourth generation over the third generation. This paper presents methods and results on lyric clustering to investigate whether the generation of girl groups has started to change. We collected meta-information data for 1469 songs of 47 groups released from 2013 to 2022 and classified them into lyric information and non-lyric meta-information and quantified them respectively. The lyrics information was preprocessed by applying word-translation frequency vectorization based on previous studies and then selecting only the top vector values. Non-lyric meta-information was preprocessed and applied with One-Hot Encoding to reduce the bias of using only lyric information and show better clustering results. The clustering performance on the preprocessed data is 129%, 45% higher for Spherical K-Means' Silhouette Score and Calinski-Harabasz Score, respectively, compared to Hierarchical Clustering. This paper is expected to contribute to the study of Korean popular song development and girl group lyrics analysis and clustering.

Keywords: *TF-IDF, Spherical K-Means, Clustering Evaluation, Lyric Clustering, K-Pop girl group*

[1] *전북대학교 소프트웨어공학과 학사과정 (hjs40111@jbnu.ac.kr)*
[2*] *교신저자 전북대학교 소프트웨어공학과 교수 (chojh@jbnu.ac.kr)*

Received: Apr. 04, 2023, Revised: May. 11, 2023, Accepted: May. 16, 2023

I. 서론

‘아이돌 세대론’은 국내 아이돌의 역사를 설명할 때 아이돌 팬덤, 온라인 커뮤니티, 신문 기사 등에서 등장하는 단어이다. 이견이 있지만 통상적으로 7 년 단위로 구분하며 1990 년대는 1 세대, 2000 년대는 2 세대, 2010 년대는 3 세대로 정의한다[1]. 2020 년대 음원시장은 브레이브걸스의 미니 4 집 타이틀 곡인 ‘Rollin``이 4 년 만에 재등장하면서 코로나로 침체된 음원시장을 걸그룹 강세라는 판도로 바꿨다. 여기에 레드벨벳, BLACKPINK, Twice 등의 3 세대 걸그룹의 중흥과 소녀시대, KARA, EXID 등의 2 세대 걸그룹의 컴백으로 걸그룹이 음원차트를 꾸준하게 점유했다.

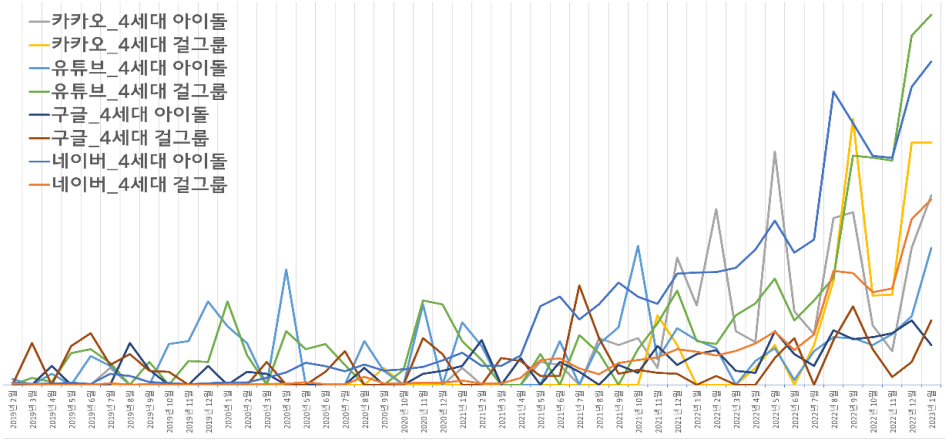

Figure 1. Transition of Searching Trend about ‘4th Generation Idol’ in Jan.2018. ~Jan.2023.
그림 1. 2018 년 1 월부터 2023 년 1 월까지의 ‘4 세대 아이돌’에 대한 검색 트렌드 추이

그 중에서도 가장 두각을 드러낸 것은 2020 년대에 데뷔한 STAYC, aespa, IVE, LE SSERAFIM, NewJeans 등의 신인들이었다. 이들은 ‘4 세대 아이돌’ 혹은 ‘4 세대 걸그룹’이라 불렸는데 2020 년 6 월 12 일, 아이돌 전문 평론 웹진인 ‘Idology’에서 “아이돌 세대론: 2020 아이돌팝 세대론”이라는 글에서 ‘4 세대 아이돌’이라는 단어를 처음 사용한 이후 ‘그림 1’에서처럼 주요 검색 포탈에서 언급되는 빈도가 기하급수적으로 증가했다[2]. 이는 대중들이 ‘4 세대 아이돌’이라고 불리는 신인들에 대한 관심이 시간이 지날수록 증가하고 있다는 것을 방증한다. 이러한 대중들의 기대심리는 4 대 대형 기획사인 YG, SM, JYP, HYBE 의 주가 상승에도 반영됐으며 2023 년 2 월 1 일 한국경제의 기사에 따르면 최근 3 개월간 주가 상승에 4 분기 실적보다 더 큰 영향을 미쳤다고 평가했다[3].

이에 반해 보이그룹은 걸그룹과 달리 세대교체가 더딘 모습을 보인다. 이것을 두고 보이그룹의 병역 의무 이행 문제, 멤버들의 잦은 사건사고, 방탄소년단의 압도적인 대중성에 따른 시장의 잠잠함 등으로 제시한 위험요소가 대중들이 걸그룹을 주목하는 이유라고 분석한 2022 년 9 월 6 일 Invest Chosun 의 기사가 있다[4]. 또한 보이그룹이 소비자 타겟을 팬덤으로 상정한 나머지 대중성을 확보하지 못해 걸그룹과 달리 세대 교체의 난항을 겪는다고 분석한 2022 년 5 월 31 일 한국일보의 기사도 있다[5]. 상기한 이유들로 인해 보이그룹에 대한 대중들의 현 주목도는 상대적으로 적기 때문에 세대교체가 더딘 것으로 보인다.

이러한 유동적인 K-Pop 의 시장의 실황을 분석하는 것은 쉽지 않다. 이를 마케팅적 관점에서 접근한 연구는[1] K-Pop 의 시장을 평가할 수 있지만 시장 환경을 분석하는 관점이기 때문에 실제로 아이돌의 컨셉이나 정체성의 변화를 알기 어렵다. 또한 대중음악 가사에 대한 감정분석 연구는[6][7] 대중가요 발전사에 도움이 되기 어렵다. 때문에 본 연구는 현 K-Pop 시장에서 주목받는 ‘4 세대 걸그룹’이 음반 메타정보의 군집화를 통해 3 세대와 구분되는지 알아보고자 한다. 군집화의 결과가 올바르다면 K-Pop 시장에서 걸그룹은 3 세대와 4 세대로 구분하여 설명할 수 있을 것이다. 이를 검증하기 위한 음반 메타정보의 수치화와 군집화 방법에 대해 3 장에서, 데이터 군집

화 결과의 분석과 2023 년 신곡에 대한 모델의 정확도를 4 장에서 다룬다.

II. 선행연구

2.1 Term Frequency – inverse Document Frequency

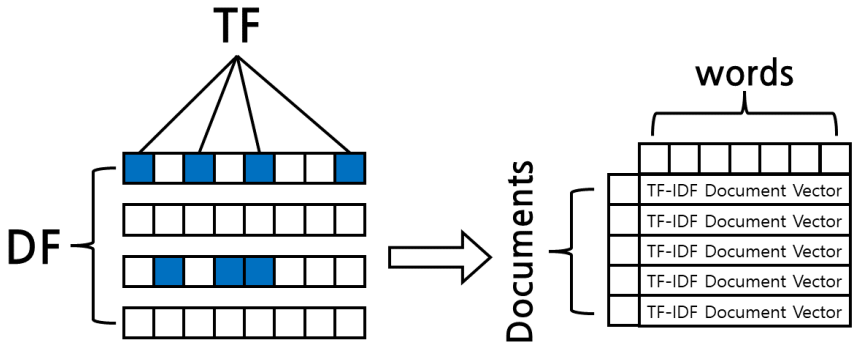

Figure 2. Schematized calculating TF-IDF
그림 2. TF-IDF 계산의 도식화

Terms Frequency-inverse Document Frequency(이하 TF-IDF)는 문서와 단어 간의 관계를 표현하는 방법 중 하나이며[8] 문서를 확실하게 표현하는 단어가 전체 문서에서는 낮은 빈도로 나타날 것이라는 가정을 단어 가중치에 반영한다[9]. 문서집합 $D = \{d_1, \dots, d_n\}$과 집합 D 의 각각의 모든 문서의 단어집합 $T = \{t_1, \dots, t_m\}$에 대하여 $tf(t,d)$는 문서 d 에 대해 단어 t 의 빈도를 의미한다.

[8]과 [9]는 각 $tf(t,d)$값을 전체 단어 빈도수로 나눈 상대적 빈도를 사용했고 [10]과 [11]에서는 절대적 빈도를 사용했다. $idf(t)$는 단어 t 가 나타나는 문서의 빈도인 $df(t)$의 역수를 의미한다.

$df(t)$가 0 인 경우가 있을 수 있어 $idf(t)$는 $\log\left(\frac{n}{df(t)+1}\right)$로 나타낸다. TF 만 사용하게 되면 중요하지 않은 단어가 단순히 빈도가 높다는 이유로 높은 값을 가진다는 문제점이 있으며[10] IDF 만 사용하게 되면 대부분의 단어들이 비슷한 값을 가진다는 문제점이 있다[8]. 때문에 TF 와 IDF 의 곱인 $tfidf(t,d)$로 표현한다. 이 값을 모든 문서와 단어에 대응하는 행렬로 나타내면 Document-Terms Matrix(이하 DTM)이 되며 모든 단어의 개수와 같은 차원수를 가지는 문서 벡터의 집합으로도 볼 수 있다. 이러한 전체적인 과정은 '그림 2'와 같이 도식화 할 수 있다[8][9][10][11].

그럼에도 TF-IDF 와 같은 단순 빈도수 기반의 방식은 문장의 단어 의미를 고려하지 못하고 새로운 단어를 해석하지 못해 효율성이 떨어진다. 또한 고차원 희소 행렬인 DTM 의 문제를 해결하는 것도 기술적 과제이다[11]. 선행연구에서는 이를 해결하기 위해 소설의 구조 별 텍스트의 TF-IDF 값을 구조 별 중요도에 따라 가중치를 달리하거나[9], 기사 제목의 TF-IDF 와 기사 내용의 TF-IDF 의 합으로 선정한 키워드의 벡터 값과 Document to Vector 방식으로 선정된 키워드의 유사도를 점별 상호 정보량으로 나타내어 보다 정확한 키워드를 추출하거나[12], 말뭉치 내에는 있지만 문서에는 존재하지 않는 단어의 예측을 위해 변형된 TF 와 변형된 IDF 의 곱으로 나타내거나[6], L2 정규화를 통해 Spherical K-Means 를 이용했다[10][11][13].

본 연구에서는 두 가지 방식으로 TF-IDF 에 대한 문제점을 개선했다. 첫 번째는 모든 가사에 대해서 각 가사별로 상위 10 개 TF-IDF 값을 갖는 단어들만 종합하여 이에 대한 TF-IDF DTM 으로 나타내었다. 두 번째는 가사 외 음반 정보를 특성으로 선정하여 각각의 가중치를 달리하였다.

[7]과 [9]에서는 분석 대상 정보를 소설 내용이나 가사에만 국한시키지 않았다. 머리말, 맺음말, 대화문과 비대화문으로 구조를 나눠 소설의 주제를 예측하거나 노래의 구조, 음정, 빠르기, 감정 등의 정보에 각각 다른 가중치를 주어 노래를 추천하는 시스템 모두 좋은 성능을 보여줬다. 마찬

가지로 본 연구도 가사 외 음반 메타정보에 각각 다른 가중치를 주었으며 3.3 과 4.1 에서 설명한다. 선정한 특성에는 앨범 발매 연도, 앨범 종류, 트랙 번호, 소속사, 작사가, 작곡가, 편곡가이며 해당 논문에서는 가사데이터는 TF-IDF 로, 그 외에는 특성 혹은 메타데이터로 정의했다.

2.2 Spherical K-Means Clustering

Spherical K-Means 군집화는 K-Means 군집화와 달리 데이터의 거리를 계산할 때 Cosine Similarity 를 이용하는 알고리즘이다. Cosine Similarity 는 두 벡터의 내적을 각 벡터의 크기의 곱으로 나눈 값이다.

K-Means 가 데이터 간 거리를 계산할 때 Euclidean Distance 를 쓰는 것과 달리 Spherical K-Means 는 Cosine Similarity 를 이용한다. 마찬가지로 K-Means 가 Centroid 를 수정할 때 전체 벡터의 합을 행의 개수로 나누는 것과 달리 Spherical K-Means 는 L2 Normalization 이 적용된 벡터의 합을 행의 개수로 나눈다. Spherical K-Means 는 고차원 희소 행렬에 대해 Euclidean K-Means 보다 더 좋은 성능을 보인다. 대다수 문서의 단어 수는 전체 문서의 단어 수에 비하면 5~1% 혹은 그 미만의 비율로 나타나기 때문에 DTM 은 고차원 희소 행렬로 나타난다. 따라서 DTM 을 군집화할 때는 Spherical K-Means 가 좋은 성능을 보이며 이는 선행연구를 통해 검증되었다[10][11][14][15].

2.3 Metrics of Unsupervised Clustering

2.3.1 Silhouette Coefficient

Silhouette Coefficient 는 밀도가 높고 분명하게 나눠지는 군집을 찾는데 유용하다. Silhouette Coefficient 를 계산하기 위해서는 군집과 데이터 지점 간의 근접도를 모두 나타낸 집합이 필요하다. 각 지점 i 에 대한 Silhouette Score s(i)의 공식은 아래와 같다.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

a(i)는 지점 i 가 속한 군집 내 다른 모든 데이터 지점과의 평균적인 거리이다. b(i)는 지점 i 가 속한 군집을 제외한 모든 군집 C 에 대해 계산한 d(i, C)에 대한 최대값인데 d(i, C)는 군집 C 가 가진 모든 데이터 지점에 대한 지점 i 와의 평균적인 거리이다. 즉, d(i, C)에서 d 의 값은 지점 i 가 속한 군집과 가장 먼 군집과의 거리로 계산될 것이다. 공식에 따라서 Silhouette Score 는 아래와 같은 세 가지 경우로 생각해볼 수 있기 때문에 -1 과 1 사이의 값을 가진다[16].

$$s(i) = \begin{cases} 1 - \dfrac{b(i)}{a(i)} & if\ a(i) > b(i) \\ 0 & if\ a(i) = b(i) \\ \dfrac{a(i)}{b(i)} - 1 & if\ a(i) < b(i) \end{cases}$$

Silhouette Coefficient 는 모든 Silhouette Score 의 평균값이다. 군집의 Silhouette 너비의 평균인 SilhouetteW 도 제시되었으나 크기가 큰 군집 몇 개와 크기가 작은 군집 여러 개로 나눠지는 식으로 군집의 크기가 불균등하다면 Silhouette Coefficient 보다 현저히 좋지 않은 성능을 보여줬다[17]. 본 연구는 Silhouette Score 와 Silhouette Coefficient 만으로 평가했다.

Silhouette Coefficient 에 대한 판단기준은 참조한 논문에서 언급되지 않았으나 삽입된 그림에서 적절한 군집 수를 가지고 Silhouette Coefficient 값이 높을수록 데이터의 Silhouette Score 값이 고르고 군집 당 데이터 수 또한 고른 것으로 나타난다.

2.3.2 Calinski-Harabasz Score

Calinski-Harabasz Score(이하 CH)는 군집 내 거리의 합(이하 WCSS)에 대한 군집 간 거리의 합(이하 BCSS)의 비율로 나타낸 지표이다. CH 값이 증가하는 것은 WCSS 가 감소하는 것이기 때문에 군집의 밀도가 높아지는 것을 의미한다. 특히 K-Means, 계층적 군집화(특히 Ward) 알고리즘에

정확한 결과를 나타내며 CH 가 다른 지표들에 비해 평균적으로 더 좋은 성능을 보인다고 했다[18]. 각 군집의 WCSS 를 군집 내 데이터의 수로 나눈 총합을 나타낸 BallHall 도 제시되었으나 SilhouetteW 가 나쁜 성능을 보여주는 조건과 같은 조건에서는 CH 보다 나쁜 성능을 보여줬다[17].

III. 연구방법

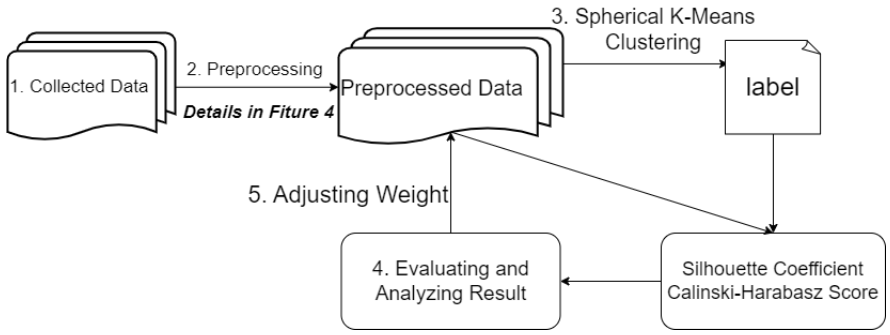

Figure 3. Process of Methodology
그림 3. 연구 과정

'그림 3'은 전체적인 연구과정을 나타낸 것으로써 데이터 수집은 바로 아래 문단에서, 전처리는 '그림 4'에서, 모델 선정과 결과 분석, 가중치 조정은 4 장에서 설명한다.

데이터 구축을 위해 선정된 47 그룹은 아래와 같은 기준으로 선정했다.

1. 2013 년 이후에 데뷔
2. 2013 년에서 2022 년까지 6 개의 음악방송(SBS 인기가요, SBS THE SHOW, MBC 음악중심, MBC SHOW Champion, KBS 뮤직뱅크, Mnet M countdown)에서 1 위 후보로 선정된 경우가 한 번이라도 있는 경우
3. 혼성 그룹, 발라드가 주 장르인 그룹 미포함
4. 유닛 그룹 포함

해당 그룹이 발표한 전체 앨범에서 가사가 존재하는 1487 개 수록곡의 그룹 데뷔연도, 그룹 소속사, 앨범 발매연도, 앨범타입, 수록곡 번호, 타이틀곡 여부, 작곡가, 작곡가, 편곡가, 가사를 종합하여 데이터를 구축했다. 이중 2023 년 1 월에 발매한 18 곡은 4.3 장에서 신곡에 대한 모델의 검증용 데이터로 사용했다. 구축된 데이터를 전처리하고 Spherical K-Means 로 군집화를 진행해 Silhouette coefficient, CH 로 평가했으며 최종적인 군집의 개수, 메타데이터, 가사데이터의 가중치를 각각 2, 0.1, 0,9 로 설정했다.

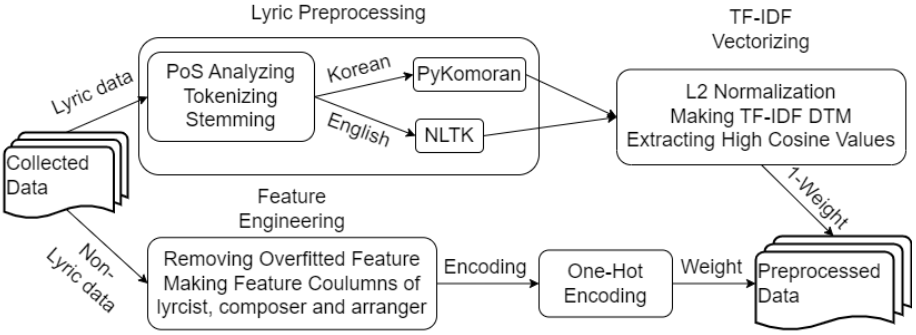

Figure 4. Process of Data Preprocessing
그림 4. 데이터 전처리 과정

‘그림 4’는 수집된 데이터를 수치화된 데이터로 변환하는 과정을 나타낸 것이다. 가사 데이터는 각 언어별로 품사분석, 어간추출, 토큰화를 거쳐 TF-IDF 벡터화를 통해 TF-IDF DTM 을 생성했고 메타데이터는 One-Hot Encoding 을 적용한 행렬로 변환했다. 전처리된 두 데이터의 가중치를 다르게 적용한 뒤 결합해 수치화된 데이터를 준비했다.

한국어 전처리에는 PyKomoran 을 이용했다. 먼저 품사 분석을 통해 실질형태소에 해당하는 14 개 품사에 해당하지 않는 단어들은 전부 불용어로 처리했다. 이 과정에서 사용자가 정의한 단어나 어절 단위의 품사분석 사전이 사용됐다. 동시에 품사 분석 결과가 외국어인 경우 영어로 취급했으며 ‘Déjà vu’, ‘Me gustas tu’ 등의 외국어도 모두 영어로 일괄 취급했다. 품사 분석 후에는 용언의 활용을 무시하기 위해 어간 추출까지 진행했다.

영어 전처리에는 NLTK 를 사용했다. NLTK 에서 비교적 잘 인식하지 못하는 은어나 줄임말 등의 단어를 표준어로 바꿔 준 후 한국어 전처리와 마찬가지로 품사 분석과 어간 추출을 진행했으며 형용사, 명사, 부사, 동사에 해당하지 않는 모든 단어를 불용어로 처리했다.

Table 1. Words of 10 Highest Cosine Value
표 1. 코사인 값이 높은 상위 10 개의 단어

word	TF-IDF Value	Improved TF-IDF Value	word	TF-IDF Value	Improved TF-IDF Value
Matter	0.500689	0.505920	가까이	0.167506	0.169256
Want	0.189460	0.191439	Spotlight	0.162145	0.163839
마음	0.177964	0.179824	좋아하다	0.155839	0.157467
Special	0.172516	0.174318	피하다	0.148214	0.149762
못	0.168009	0.169764	떨리다	0.144146	0.145652

전처리된 가사는 공백을 기준으로 토큰화했다. 이 데이터를 가지고 TF-IDF 를 계산한 후 L2 Normalization 이 적용하여 (1487, 9426)의 TF-IDF DTM 을 생성했고 다시 검증용 데이터 18 곡을 제외시켜 (1469, 9426) DTM 으로 변환했다. 2.1 장과 2.2 장에서 설명한 바와 같이 고차원 희소 행렬이라는 문제점을 해결하기 위해 본 연구는 각 문서의 높은 TF-IDF 값을 가지는 단어를 10 개씩 추출하여 TF-IDF DTM 을 재생성하였으며 문서 벡터를 4548 차원으로 축소시켰다. ‘표 1’은 프로미스나인의 미니 4 집 ‘Midnight Guest’의 타이틀 곡인 ‘DM’을 예로 든 것으로 TF-IDF 값이 낮은 단어들을 제거함으로써 높은 값을 가지는 단어들이 해당 곡을 더욱 대표할 수 있게 했다.

Table 2. Example Applied One-Hot Encoding to Fromis_9’s 5 Songs
표 2. 프로미스나인 5 곡에 대해 One-Hot Encoding 방식을 적용한 예시

Song Name	Debut_ 2018	Release _2021	Type_M	Track_1	Agency_HYBE	서지음	이우민 "collapsedone"	Justin Reinstein
Feel Good	1	0	1	1	1	1	1	1
WE GO	1	1	0	0	1	0	1	0
Talk & Talk	1	1	0	1	1	0	0	0
DM	1	0	1	0	1	0	1	1
Stay This Way	1	0	1	0	1	1	1	1

메타데이터는 모델의 과적합을 일으키는 특성인 그룹 데뷔연도와 타이틀곡 여부를 제외하고 One-Hot Encoding 방식으로 전처리했다. 앨범 타입, 트랙 번호, 소속사, 발매연도는 category-encoder 라이브러리의 One-Hot Encoder 를 이용했다. 그러나 작곡가, 작사가, 편곡가는 리스트 자료형으로 되어있어 바로 적용할 수 없었기 때문에 곡의 참여자 전체를 열로 하고 각 곡을 행으로 하는 이진행렬을 생성했다. 상기한 방식을 프로미스나인의 5 곡에 대해서 예시로 들면 ‘표 2’와 같이 나타나며 이 행렬에 TF-IDF DTM 을 결합하여 수치화된 데이터를 준비했다.

IV. 연구결과

4.1. 정량적 지표 평가

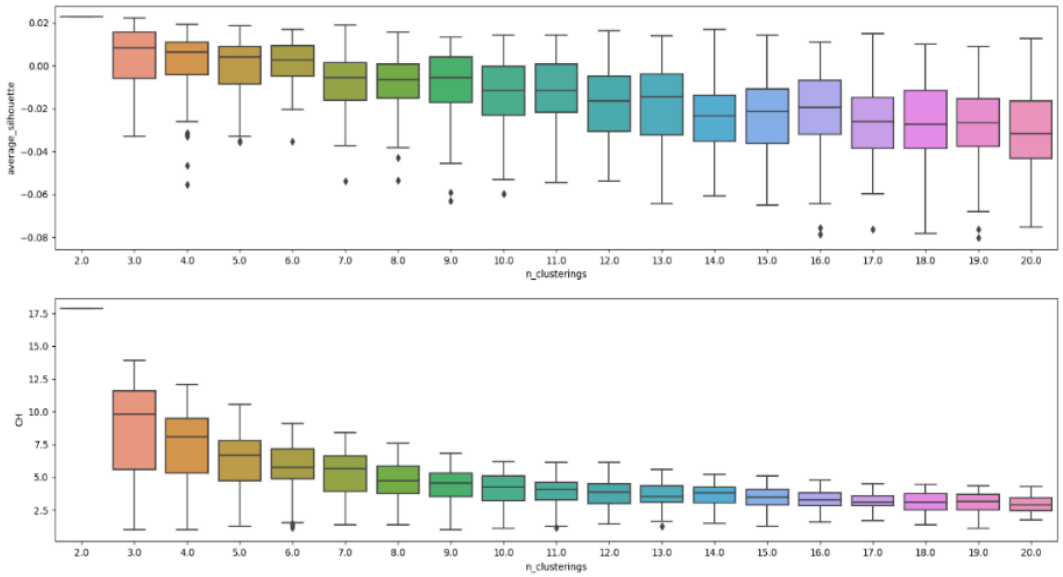

Figure 5. Box Plot for Silhouette Coefficient and CH along Number of Clustering
그림 5. 군집의 개수에 따른 정량적 지표 값의 그래프

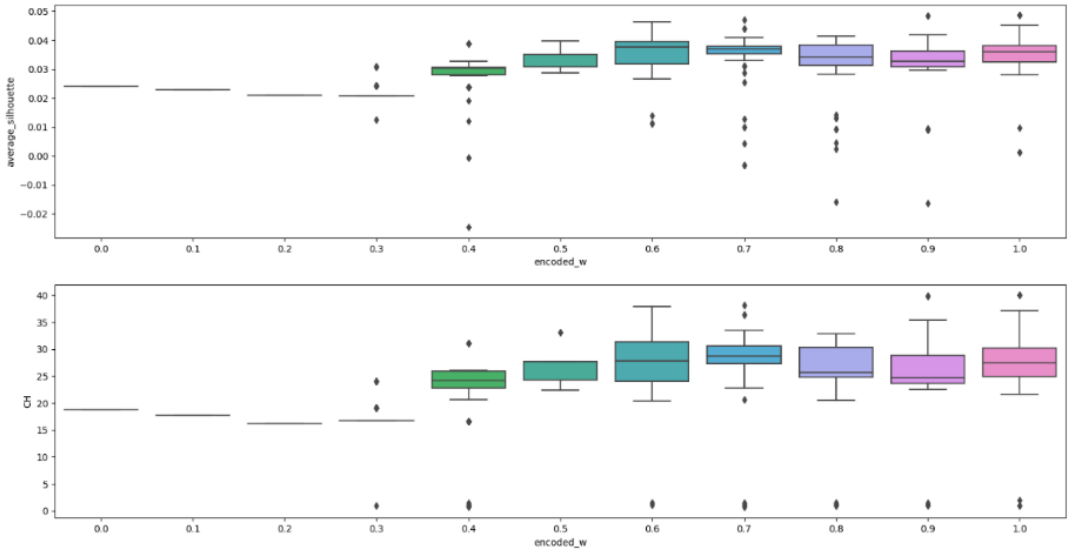

Figure 6. Box Plot for Silhouette Coefficient and CH along Weight of Feature Data
그림 6. 메타데이터의 가중치에 따른 정량적 지표 값의 그래프

‘그림 5’ 군집의 개수에 따른 Silhouette Coefficient, CH 값의 상자수염 그래프를, ‘그림 6’는 가사 외 메타데이터의 가중치에 따른 Silhouette Coefficient, CH 값의 상자수염 그래프를 그린 것이다. ‘그림 5’을 통해 군집의 수가 2 개인 것이 가장 적절하다고 판단했다. ‘그림 6’에서 메타데이터의 비율이 늘어날수록 군집화 결과의 이상치나 극단치를 자주 보여주지만 중앙 값을 기준으로

0.6 일 때 좋은 성능을 보여줬다. 이와 별개로 본 연구는 4.2 와 4.3 의 결과로 해석하기 위해서는 메타데이터의 가중치가 0.1, 가사데이터의 가중치가 0.3 일 때가 가장 적합하다고 판단했다.

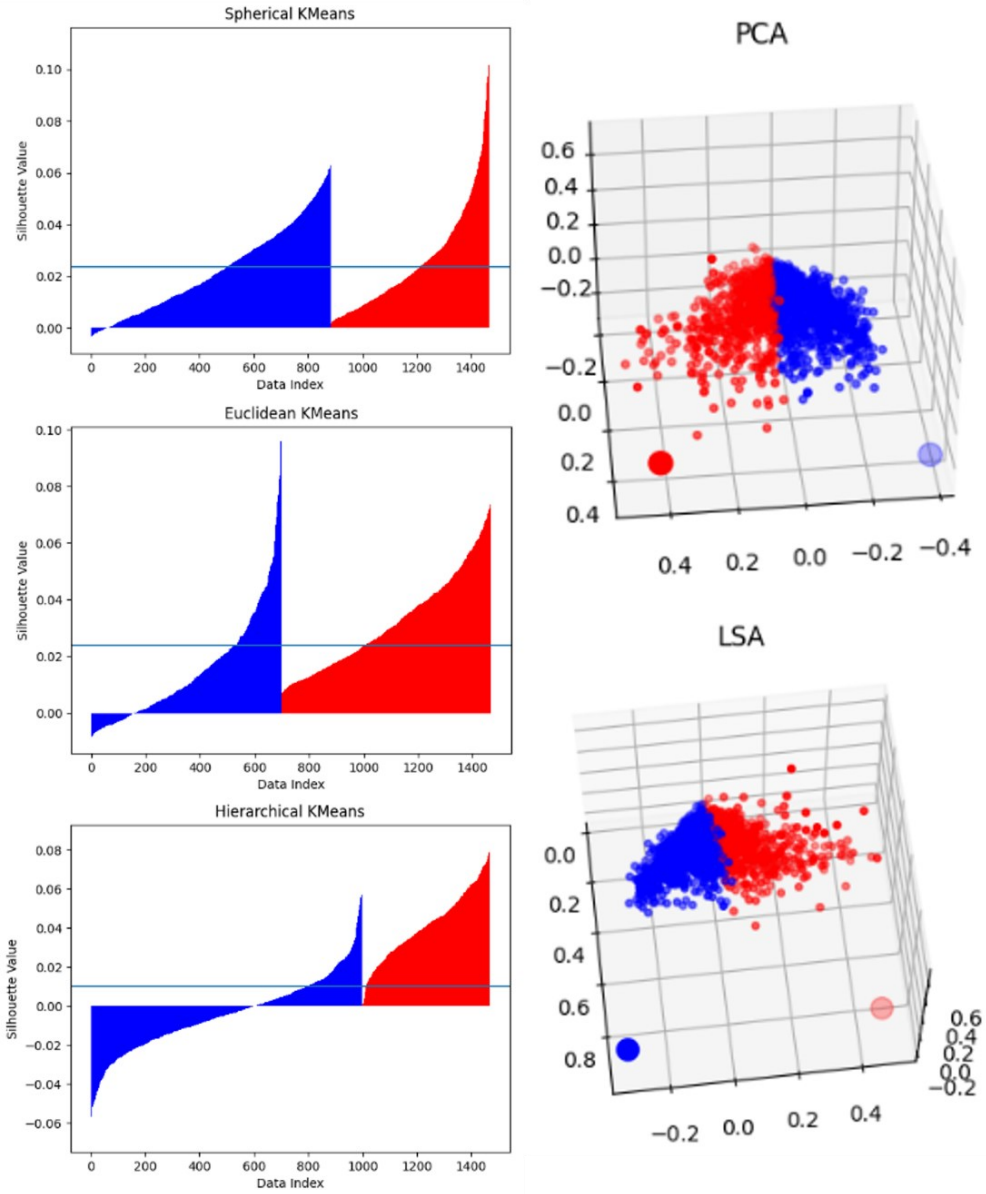

Figure 7. Silhouette Score Graph of Clustering Algorithms and Visualization through Dimension Reduction
그림 7. 군집화 알고리즘별 Silhouette Score 그래프와 차원축소를 통한 시각화

‘그림 7’는 3 장에서 언급한 매개변수를 각각 다른 군집화 알고리즘인 Euclidean K-Means, Spherical K-Means, Hierarchical Clustering 에 적용한 결과이다. Euclidean K-Means 와 Spherical K-Means 의 Silhouette Coefficient 는 비슷하지만 Silhouette Score 가 음수인 데이터가 Spherical K-Means 에서 더 적기 때문에 Spherical K-Means 가 더 적합한 방식이라고 볼 수 있다. 한편 Spherical K-Means, Euclidean K-Means, Hierarchical Clustering 의 CH 값은 각각 18.455, 18.484, 12.852 으로 Spherical K-Means 와 Euclidean K-Means 의 점수 차이는 거의 없으나 Hierarchical Clustering 와의 차이는 큰 것으로 나타났다.

4.2. 메타데이터의 특성과 군집 간 상관관계 분석

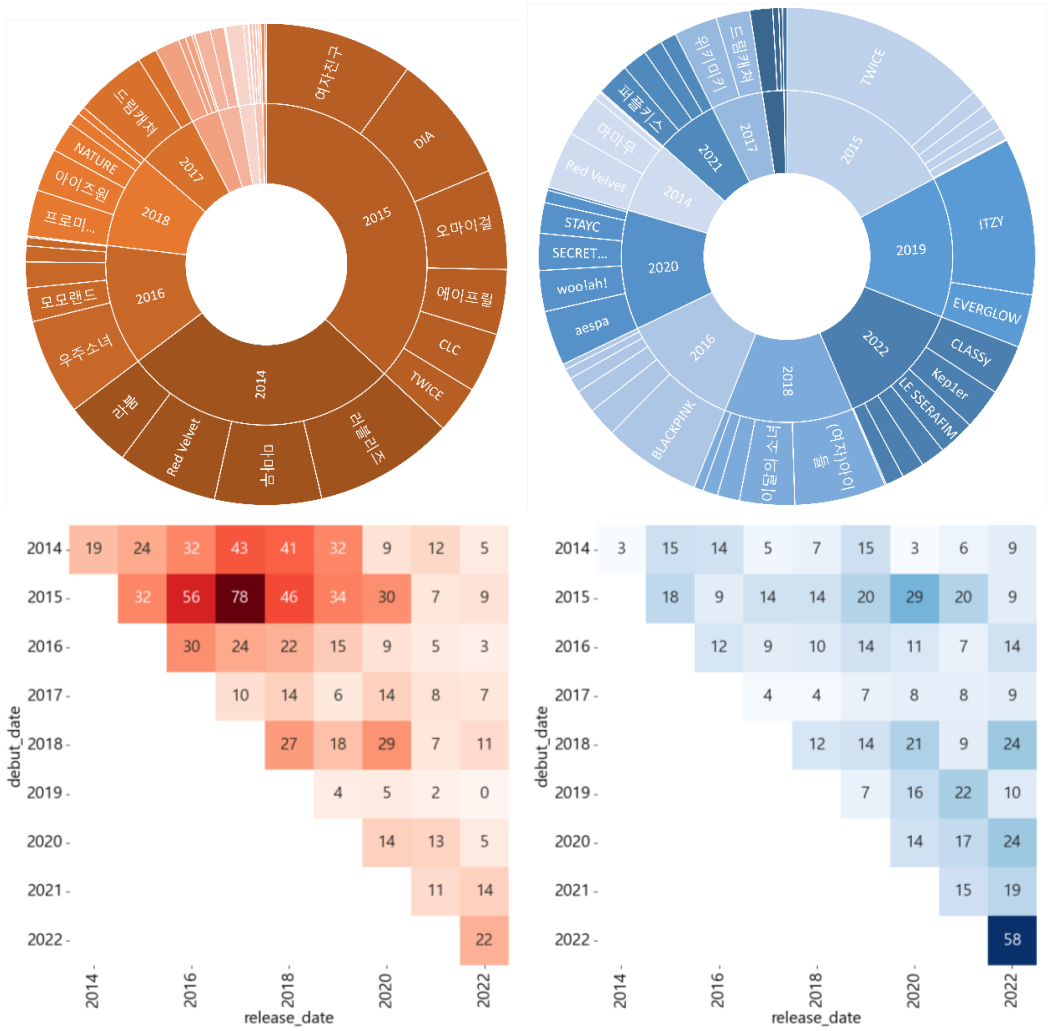

Figure 8. Visualizing Distribution of 3rd, 4th Generation along group debut year and album releasing year
그림 8. 그룹의 데뷔연도와 노래 발매연도에 따른 3 세대, 4 세대 분포 그래프

'그림 8'의 좌측 그래프는 3 세대를, 우측 그래프는 4 세대를 나타내며 Sunburst Chart 는 그룹이 발표한 곡 중 해당 세대에 속한 곡의 비율을, Heatmap Chart 는 곡의 발매연도와 그 그룹의 데뷔연도가 일치하는 곡의 수를 나타낸 것이다. 3 세대는 2015, 2014, 2016 년에 데뷔한 그룹의 곡들이 전체의 약 75%를 차지하며, 4 세대는 2019, 2022, 2018, 2020 년에 데뷔한 그룹의 곡들이 전체의 약 50%를 차지한다. 4 세대 Sunburst Chart 에서 주목할만한 점은 TWICE 와 BLACKPINK 가 높은 비율을 보이는 것이다. TWICE 는 전체 그룹 중 가장 많은 곡을 발표했기 때문에 3 세대와 4 세대 구분하지 않고 군집에 속한 곡의 절대적인 수가 커서 그런 것으로 해석된다. 한편, BLACKPINK 의 경우 곡의 수가 압도적으로 많은 것도 아니고 2016 년에 데뷔한 다른 그룹에 비해 유독 비율이 크기 때문에 활동 시기는 3 세대로 분류되어도 곡의 특성은 4 세대로 분류되는 것으로 해석할 수 있다. 반대로 활동시기는 4 세대로 분류되어도 곡의 특성이 3 세대로 분류되는 것으로 해석되는 그룹에는 '첫사랑'이 있었다.

이외의 메타데이터와 군집 간의 관계에 대해서는 뚜렷한 연관성을 찾을 수 없었다.

4.3. 2023 년 신곡 검증

상기된 그룹 중 2023 년 1 월에 신곡을 발매한 그룹은 다섯 그룹이며 총 18 곡이다. 해당 곡들은 모델이 '표 3'과 같이 분류하였다.

Table 3. New Songs Released Jan. 2023. and Clustered Label of Them
표 3. 2023 년 신곡과 군집 라벨

debut_year	artist	album	song	album_type	track	agency	label
2022	GOT the beat	Stamp On It	Stamp On It	M 1	1	SM	4th
			Goddess Level		2		4th
			Alter Ego		3		4th
			가시		4		4th
			Outlaw		5		4th
			MALA		6		4th
2022	ILY:1	A Dream of ILY:1	별꽃동화	M 1	1	FC	3rd
			Secret Recipe		2		3rd
			Tasty		3		3rd
			Thanks to…		4		3rd
2022	NewJeans	OMG	OMG	S 1	1	HYBE	4th
2022	VIVIZ	VarioUS	PULL UP	M 3	1	BPM	3rd
			Blue Clue		2		4th
			Love or Die		3		4th
			Vanilla Sugar Killer		4		3rd
			Overdrive		5		3rd
			So Special		6		3rd
2017	DREAMCATCHER	REASON	REASON	DS	1	DREAMCATCHER	3rd

'DREAMCATCHER', 'GOT the beat', 'NewJeans'의 신곡은 속한 세대에 일치하게 분류됐다. 'VIVIZ'는 3 세대와 4 세대가 비슷한 비율로 나타났는데 이는 2015 년도에 데뷔한 '여자친구'가 해체되고 일부 멤버가 재결성한 그룹이기 때문에 3 세대의 특성도 가지는 것으로 보인다. 한편 'ILY:1'의 신곡은 활동시기가 4 세대에 해당함에도 불구하고 3 세대로 분류됐는데 이는 그룹의 컨셉인 청순함이 3 세대에 활동하던 그룹들과 비슷했기 때문인 것으로 보인다. 결론적으로 해당 모델은 그룹이 속한 세대와 곡의 세대가 전반적으로 일치하는 모습을 보여줬다.

V. 결론

본 연구에서는 메타데이터와 노래의 특성에 따라 가장 적합한 군집화 방법을 찾고 이를 통해 한국의 현대 대중음악을 3 세대와 4 세대로 구분하였다. 일부 곡이 그룹의 세대에 속하지 않았지만 전반적으로 3 세대와 4 세대로 나뉘는 결과를 보여줬다. 그러므로 본 연구는 아이돌 세대론과 걸그룹 가사 군집화를 연구하기 위한 선구적 연구 방법론으로 활용될 수 있다.

연구 결과, 메타데이터의 가중치가 0.1 일 때 좋은 결과를 보이며, Spherical K-Means 알고리즘이 Euclidean K-Means 알고리즘보다 더 적합하다는 것을 확인할 수 있었다. 군집의 수를 정할 때에는 Silhouette Coefficient 와 CH 모두 좋은 지표로 사용됐다. 다만 선행연구를 참고하여 가사 외의 메타데이터의 가중치를 조절하는데 Silhouette Coefficient 와 CH 의 값이 아닌 3 세대와 4 세대의 분류 가능성에 초점을 두고 가중치를 선정했기 때문에 정량적 지표와 데이터의 가중치 간의 연관관계를 설명할 수 있는 후속연구가 필요해 보인다. 또한 본 연구를 통하여 4 세대가 시작되었음을 알 수 있었지만 신인 걸그룹이 발표한 곡의 수가 적기 때문에 차후 데이터가 더 늘어난다면

같은 방법으로 세대론을 검증하는 후속 연구도 필요하다고 여겨진다. 더하여 걸그룹만 연구 대상으로 삼았기 때문에 최근 한국 대중가요 전반의 흐름을 파악하기 위해선 장르를 불문하고 많은 가수들의 곡에 대한 가사 데이터와 메타 데이터의 수집과 분석이 필요하다.

VI. 감사의 글

본 연구는 정보통신기획평가원과 과학기술정보통신부의 지원을 받아 수행하였습니다. (열린 혁신 플랫폼 디지털 오픈랩, 과제번호 2021-0-00546)

VII. 참고문헌

[1] S.W. Choi, S. J. Limb (2019, Dec). The Third-Generation K-Pop Idols' Strategies: Focused on 'EXO', 'TWICE' and 'BTS'. Journal of Industrial Innovation. 35(4), pp.57~93. [Online]. Available: https://doi.org/ 10.22793/indinn.2019.35.4.003

[2] Dis1co (2020, Jun). Theory of Idol Generation: Theory of 2020 Idol Pop Generation. 『Idology』. [Online]. Available: https://idology.kr/13070

[3] S. M. Sim (2023, Jan). "New Girl Groups Debut One after Another in This Year"… Entertainment Companies Stocks are Bull Market by Growing Anticipation. 『HanKyung Korea Market』. [Online]. Available: https://www.hankyung.com/finance/article/2023020167281

[4] S. E. Lee (2022, Sep). 'Top Pick' of Entertainment Companies' Stocks is also Girl Groups …HYBE is just 'shaking' with BTS. 『Invest Chosun』. [Online]. Available: http://www.investchosun.com/site/data/html_dir/2022/09/05/2022090580662.html

[5] H. M. Hong (2022. May). [HI★Focus] 4th Generation Boy Groups, Review for Popularity. 『Hankook Ilbo』. [Online]. Available: https://hankookilbo.com/News/Read/A2022052610230002851

[6] Kornkanya Siriket, Vera Sa-ing, Subhron Khonthapagdee (2021, Mar). Mood classification from Song Lyric using Machine Learning. [Online]. Available: http://doi.org/10.1109/iEECON51072.2021.9440333

[7] J. H. Lee (2016, Oct). Popular Music Similarity Evaluation using Emotion and Structure Analysis on Lyrics. KIISE Transactions on Computing Practices. [Online]. Available: http://doi.org/10.5626/KTCP.2016.22.10.479

[8] Sabbah, Thabit, Selamat, Ali, Selamat, Md Hafiz, Al-Anzi, Fawaz S., Viedma, Enrique Herrera, Krejcar, Ondrej, Fujita, Hamido (2017, Apr). Modified frequency-based term weighting schemes for text classification. [Online]. Available: https://doi.org/10.1016/j.asoc.2017.04.069

[9] E.S. You, G. H. Choi, S. H. Kim (2016, Feb.). Study on Extraction of Keywords Using TF-IDF and Text Structure of Novels. [Online] Available: https://doi.org/10.9708/jksci.2015.20.2.121

[10] Anna Huang. "Similarity measures for text document clustering," In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)., Christchurch, New Zealand, 2008. pp. 9-56.

[11] Inderjit S Dhillon, Dharmendra S Modha (2000, Oct). Concept decompositions for large sparse text data using clustering. [Online]. Available: https://doi.org/10.1023/A:1007612920971

[12] M.J. Lim, J.H. Kim, J.H. Shin (2019, Nov). Method of Related Document Recommendation with Similarity and Weight of Keyword. [Online]. Available: https://doi.org/0.9717/kmms.2019.22.11.1313

[13] Mohammad Alodadi, Vandana P. Janeja (2015, Oct.). Similarity in Patient Support Forums. [Online]. Available: https://doi.org/10.1109/ICHI.2015.99

[14] S.Y. Bang, M. Y. Lee (2021, Mar.). A Study on Fashion Attribute Analysis Using Spherical K-means Clustering. [Online]. Available: https://doi.org/ 10.15843/kpapr.35.1.2021.3.137

[15] Kurt Hornik, Ingo Feinerer, Martin Kober, Christian Buchta (2012, Sep). Spherical k-Means Clustering. [Online]. Available: https://doi.org/10.18637/jss.v050.i10

[16] Peter J. Rousseeuw (1986, Nov.). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. [Online]. Available: https://doi.org/10.1016/0377-0427(87)90125-7
[17] Gagolewski, Marek, Bartoszuk, Maciej, Cena, Anna (2021, Oct). Are cluster validity measures (in) valid?. [Online]. Available: https:// 10.1016/j.ins.2021.10.004
[18] Y.S.Shim, J.W.Chung, I.C.Choi (2006, Mar.). “A Performance Comparison of Cluster Validity Indices based on K-means Algorithm.”, Asia Pacific Journal of Information Systems, Vol. 16, No. 1, pp. 127-144, Mar. 2006.

저자소개

현준서(*Joonseo Hyeon*)

2019 년 3 월 전북대학교 소프트웨어공학과 학사과정

관심분야 : 자연어처리(NLP), 데이터 분석

조재혁(*JaeHyuk Cho*)

2011 년 2 월 중앙대학교 모바일 및 임베디드 SW 박사
2022 년 3 월~현재 전북대학교 소프트웨어공학과 교수

관심분야 : Bigdata, IoT, SW Platform System, applied AI