Asia-pacific Journal of Convergent Research Interchange
Vol.10, No.2, February 29 (2024), pp.25-34
http://dx.doi.org/10.47116/apjcri.2024.02.03

Analysis and Discussion of Public Sports Data based on Clustering Model

클러스터링 모델 기반의 체육 분야 공공 데이터 분석 및 고찰

Seo-Bin Hong[1], Bong-Hyun Kim[2]

홍서빈[1], 김봉현[2]

[1] Master Course, Department of Information and Communication Engineering, Seowon University, Republic of Korea, tjqls691@naver.com
[2] Professor, Department of Computer Engineering, Seowon University, Republic of Korea, bhkim@seowon.ac.kr

Corresponding author: Bong-Hyun Kim

Abstract: In a situation where effective analysis and understanding are emerging as important tasks due to the rapidly increasing amount and diversity of data in modern society, this paper identifies problems with data quality and consistency through clustering public data and seeks ways to improve them. The data used was general in physical education provided by data.go.kr. Nouns were extracted from text data and clustering was performed using TF-IDF vectorization. Performance was evaluated by comparing K-means, DBSCAN, and GMM algorithms and keyword extraction methods, and problems with data consistency and quality were analyzed. As a result of the study, it was confirmed that stopword processing and choice of keyword extraction method had a significant impact on clustering results. Additionally, data length, format, and keyword quality affect clustering performance. It was concluded that data imbalance, lack of consistency, and lack of standards can affect clustering results and that standardized guidelines and research are needed to solve these problems. We identify the diversity of data through clustering, suggest ways to improve data collection and analysis strategies through this, and emphasize the importance of improving data quality and active use of clustering techniques for the effective use of public data.

Keywords: Public Data, K-means, DBSCAN, GMM, Komoran

요약: 본 논문은 현대 사회에서 빠르게 증가하는 데이터의 양과 다양성으로 인해 효과적인 분석과 이해가 중요한 과제로 부각되고 있는 상황에서, 공공데이터의 클러스터링을 통해 데이터 품질과 일관성에 대한 문제점을 도출하고 개선 방안을 모색하는 것을 목표로 하였다. 데이터는 data.go.kr에서 제공되는 체육 분야의 공공데이터를 사용하며 텍스트 데이터에서 명사를 추출하고 TF-IDF 벡터화를 사용하여 클러스터링을 수행하였다. K-means, DBSCAN, GMM 알고리즘과 키워드 추출 방법을 비교하여 성능을 평가하고, 데이터의 일관성과 품질에 대한 문제점을 분석하였다. 연구 결과, 불용어 처리와 키워드 추출 방법의 선택이 클러스터링 결과에 큰 영향을 미침을 확인하였다. 또한 데이터의 길이, 형식, 키워드 품질 등이

Received: October 28, 2023; 1st Review Result: November 30, 2023; 2nd Review Result: January 04, 2024
Accepted: January 25, 2024

클러스터링의 성능에 영향을 미치는 것으로 나타났다. 데이터의 불균형과 일관성 부재, 표준 부재 등이 클러스터링 결과에 영향을 미칠 수 있으며, 향후에는 이러한 문제를 해결하기 위한 표준화된 가이드라인과 연구가 필요하다는 결론을 도출하였다. 클러스터링을 통해 데이터의 다양성을 파악하고, 이를 통해 데이터 수집과 분석 전략을 개선하는 방안을 제시하며, 공공데이터의 효과적인 활용을 위해 데이터의 품질 향상과 클러스터링 기법의 적극적 활용이 중요하다는 점을 강조한다.

핵심어: 공공데이터, K-means, DBSCAN, GMM, Komoran

1. 서론

현대 사회에서는 빠르게 증가하는 데이터의 양과 다양성으로 인해 효과적인 분석과 이해가 더욱 핵심적인 과제로 떠오르고 있다. 이러한 데이터 중에서도 공공데이터는 다양한 분야에서의 활용이 기대되며 사회 발전에 미치는 영향이 상당히 크다. 공공데이터는 정부 및 공공기관에서 수집된 정보를 포함하고 있어, 이를 분석하고 활용함으로써 정책 수립, 사회 문제 해결, 혁신적인 기술 개발 등 다양한 영역에서 긍정적인 결과를 낼 수 있다. 이에 대한 효과적인 분석과 이해가 필수적이다. 특히, 공공데이터는 다양한 분야에서 활용되며 사회 발전에 기여할 수 있는 중요한 자원이다. 그러나 이러한 공공데이터를 효과적으로 활용하기 위해서는 데이터의 일관성과 품질에 대한 이해가 필요하다. 과거의 연구에서는 공공데이터의 다양한 측면에서의 활용에 대한 논의가 진행되었다. 그러나 데이터의 일관성과 품질에 중점을 둔 연구는 제한적이었다[1]. 본 연구는 이전 연구들에서 미처 다루지 못한 데이터 품질의 측면에 주목하여, 공공데이터의 클러스터링을 통해 데이터 간의 패턴을 도출하고 문제점을 도출하고자 한다.

본 논문에서는 공공데이터의 클러스터링을 통해 데이터 품질과 일관성에 대한 문제점을 제시하고, 이를 통해 데이터 수집과 분석 전략을 개선하는 방안을 모색한다. 특히, 서로 다른 키워드 추출 방법과 클러스터링 알고리즘을 비교하여 성능을 평가하고, 공공데이터의 활용을 높이기 위한 실질적인 지침을 도출한다. 이를 통해 현대 사회에서 급변하는 데이터 환경에 능동적으로 대응하고, 공공데이터의 가치를 최대한 끌어올리는 데 기여하고자 한다.

2. 관련이론

2.1 클러스터링 알고리즘

2.1.1 K-means

데이터를 K개의 군집으로 나누기 위한, 거리 기반 클러스터링 알고리즘이다. 같은 집단 내 데이터들은 비슷한 특징을 가지고 있고, 다른 집단의 데이터와는 데이터적으로 상반된 특징을 가지고 있다는 것을 가정한다. 즉, 동일 집단의 군집화를 고려하는 것 뿐만 아니라, 타집단과의 관계도 고려한다. K-means는 간단하면서도 효과적인 클러스터링 알고리즘으로 널리 사용되며, 데이터를 구조화하고 유사한 패턴을 찾는 데에 유용하게 활용된다[2][3].

2.1.2 DBSCAN(밀도 기반 클러스터링)

클러스터링 알고리즘으로 Multi Dimension의 데이터를 밀도 기반으로 서로 가까운 데이터 포인트를 함께 그룹화하는 알고리즘이다. 밀도가 다양하거나 모양이 불규칙한 클러스터가 있는 데이터와 같이 모양이 잘 정의되지 않은 데이터를 처리할 때 유용하게 사용 가능하다. DBSCAN은 이상치에 민감하지 않고 다양한 형상의 클러스터를 감지하는 데에 유용하며, 데이터 내 구조를 파악하는 데에 활용된다[4].

2.1.3 GMM(Gaussian Mixture Model, 가우시안 혼합 모델)

GMM은 데이터를 여러 개의 가우시안 분포를 가진 혼합 요소로 모델링하는 확률적 생성 모델이다. 각 가우시안 분포는 데이터 내의 서로 다른 클러스터를 대표하며, 데이터 포인트는 이러한 분포들 중 하나에 속할 확률을 가진다. GMM은 EM(Expectation-Maximization) 알고리즘을 사용하여 모수 추정을 수행하며, 데이터의 분포를 정확하게 모델링할 수 있다. 이 모델은 유연하면서도 강력하여 데이터 내의 숨겨진 구조를 파악하고, 복잡한 데이터를 효과적으로 클러스터링할 수 있다[5][6].

본 논문에서는 어떤 알고리즘이 데이터에 적합한지 알 수 없어 다양한 알고리즘을 활용해 보았다. 우선 특정 유형을 명확하게 분류하고자 k-means를 선택하였으며, 특정 단어들의 밀집된 정도를 분석하기 위해 DBSCAN를 활용하였다. 더불어, 다양한 특성을 가지는 군집을 찾기 위해 GMM을 사용하여 체육분야의 공공데이터를 분석하였다.

2.2 평가지표

2.2.1 Elbow Method

최적의 클러스터 개수를 찾기 위해 클러스터 개수 $\boldsymbol{k}$를 변화시키면서 $\boldsymbol{J(k)}$를 계산하고 Elbow Method를 적용한다. 각 데이터 포인트 $\boldsymbol{x_i}$와 해당 클러스터 중심 $\boldsymbol{y_j}$간의 거리의 제곱을 모두 합한 후, 모든 데이터 포인트와 클러스터에 대해 이 거리의 제곱을 더하여 총 제곱 거리의 합을 얻는다. 이 거리의 제곱은 유클리드 거리의 제곱을 의미하며, 클러스터링에서는 각 데이터 포인트가 속한 클러스터 중심까지의 거리를 측정하는데 사용된다[7][8].

$$d(x,y)^2 = \sum_{j=1}^{m}\left(x_j - y_j\right)^2 \tag{1}$$

$$J(k) = \sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_i - y_j\right)^2 \tag{2}$$

2.2.2 Sihouette Score

각 데이터 포인트의 클러스터 내 응집성과 클러스터 간 분리도를 고려하여 클러스터링의 품질을 평가하는 지표이다. 각 데이터 포인트는 자신이 속한 클러스터와 다른 클러스터와의 거리를 계산하여 실루엣 값으로 표현된다. 이 지표는 -1부터 1까지의 값을 가지며, 1에 가까울수록 클러스터가 잘 형성되었음을 나타낸다[9].

2.2.3 Davies-Bouldin Index

각 클러스터의 분리도와 응집도를 고려하여 클러스터의 품질을 측정하는 지표이다. 각 클러스터 $\boldsymbol{i}$와 $\boldsymbol{j}$간의 상대적 거리를 계산한다. 상대적 거리는 클러스터 내의 데이터

포인트와 해당 클러스터 중심 간의 평균 거리와 클러스터 간의 중심 거리를 고려한다. 클러스터$\boldsymbol{i}$내의 모든 다른 클러스터$\boldsymbol{j}$에 대해 상대적 거리를 비교하여 최대값을 찾는다. 이 최대 상대 거리 $\boldsymbol{D_i}$는 클러스터$\boldsymbol{i}$의 분리도를 나타낸다. 각 클러스터에 대한 최대 상대 거리 $\boldsymbol{D_i}$를 모두 더한 후, 클러스터 개수로 나누어 평균을 계산한다. 따라서 Davies-Bouldin Index는 클러스터 간의 거리가 클러스터 내의 거리보다 작을수록 지수 값이 낮아지며, 값이 작을수록 더 좋은 클러스터링을 나타낸다[10][11].

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \tag{4}$$

$$D_i = \max_{j \neq i} R_{i,j} \tag{5}$$

$$DB = \frac{1}{N}\sum_{i=1}^{N} D_i \tag{6}$$

2.2.4 Calinski-Harabasz Index

클러스터 간의 평균 중심 $\bar{\boldsymbol{X}}_{\boldsymbol{i}}$과 전체 데이터의 평균 중심 $\bar{\boldsymbol{X}}$ 간의 거리를 고려하여 클러스터 간 분산과 각 클러스터 내의 데이터 포인트의 평균 중심$\bar{\boldsymbol{X}}_{\boldsymbol{ij}}$과 해당 클러스터의 평균 중심$\bar{\boldsymbol{X}}_{\boldsymbol{i}}$ 간의 거리를 모두 합하여 클러스터 내 분산을 계산한다. 클러스터 간 분산을 클러스터 내 분산으로 나눈 비율을 구한다. Calinski-Harabasz Index 값은 클러스터 간의 거리가 멀고, 클러스터 내의 분산이 작다는 것을 나타내며 최적의 클러스터링 결과는 값이 최대가 되는 경우로 해석한다[12].

$$\text{Between Cluster Variance} = \sum_{i=1}^{k} n_i \, ||\bar{X}_i - \bar{X}||^2 \tag{7}$$

$$\text{Within Cluster Variance} = \sum_{i=1}^{k}\sum_{j=1}^{k} ||\bar{X}_{ij} - \bar{X}_i||^2 \tag{8}$$

$$\text{Calinski} - \text{Harabasz Index} = \frac{\text{Between Cluster Variance}/(\text{k}-1)}{\text{Within Cluster Variance}/(\text{n}-\text{k})} \tag{9}$$

2.2.5 AIC(Akaike Information Criterion)

AIC는 통계 모델의 적합도를 평가하는 지표 중 하나이다. 모델의 복잡도와 적합도 사이의 균형을 나타내며, 모델의 예측 성능을 측정하는 데 사용된다. 값이 작을수록 모델의 적합도가 높은 것을 나타낸다[13].

2.2.6 BIC(Bayesian Information Criterion)

BIC는 AIC와 비슷한 지표로, 모델의 복잡도를 고려하여 적합도를 평가합니다. AIC와 마찬가지로 값이 작을수록 모델의 적합도가 높다고 판단된다. BIC는 AIC와 달리 더 많은 데이터에 대한 페널티를 부여하여 모델의 과적합을 방지한다[13].

3. 연구 방법 및 결과

본 연구의 핵심 목표는 data.go.kr에서 제공하는 공공데이터를 클러스터링하여 데이터 간의 패턴을 도출하고 유의미한 정보를 추출하다. 이를 통해 공공데이터의 특성과 유형을 파악하고 문제점을 도출하는 것이 주된 목표이다. 또한, 공공데이터를 클러스터링하여 서로 유사한 유형의 데이터들 활용하므로써 공공데이터의 활용 가능성을

높이고자 한다.

3.1 데이터 수집

본 연구에서는 공공데이터 포털인 data.go.kr에서 제공되는 공공데이터 정보를 활용하였다. 데이터 수집은 XML 형식의 데이터를 크롤링하여 진행되었으며, 수집된 데이터는 https://www.data.go.kr/data/[dataId]/fileData.do 에서 제공되었다. 그 중 분류가 '체육'인 것을 데이터셋으로 사용하였으며, 데이터의 속성은 '제목', '기관', '테마', '데이터형식', '발행일', '수정일', '갱신주기', '키워드', '설명', '요청수(다운로드수)', '파일명', '분류', '동적/정적' 이다. 14개의 열로 이루어진 963개의 데이터를 사용하였다.

3.2 데이터 전처리

Komoran을 사용하여 '설명'에서 명사(NNG: 보통명사, NNP: 고유명사)만 추출 하였다. 설명이 너무 짧거나 Nan인 값을 제외하고 학습시켰다. 텍스트 데이터로 단어 간의 유사성을 계산하고 클러스터링을 수행하기 위해 TF-IDF 벡터화를 했다. TfidfVectorizer를 사용하였고 단어의 n-gram 범위와 최소 문서 빈도 등을 설정하고 추출한 키워드 TF-IDF 벡터로 변환하였다[14].

	설명	키워드추출
4	성남시내 학교체육시설에 대한 자료이며, 학교명, 시설종류, 학교위치, 설치년도, 비...	성남시, 학교, 체육, 시설, 학교, 시설, 종류, 학교, 위치, 설치, 비고, 항...
5	성남시의 하천인 탄천에 있는 시설물 현황에 대한 데이터로 관리주체, 시설명, 위치,...	성남시, 하천, 탄천, 시설물, 관리, 주체, 위치, 시설, 면적, 단위
6	수원도시공사에서 운영중인 장안구민회관, 종합운동장의 강좌 및 스포츠 프로그램 안내 ...	원도, 시공사, 운영, 중인, 장안구, 민, 회관, 종합운동장, 강좌, 스포츠, 프...
8	서대문문화체육회관 주요시설 현황 및 프로그램 현황 제공	서대문, 문화, 체육, 회관, 시설, 프로그램
10	남동구도시관리공단에서 운영하는 남동국민체육센터 및 남동수영장의 체육 강좌 정보 제공	남동구, 도시, 관리, 공단, 운영, 남동, 국민, 체육, 센터, 남동, 수영장, ...

[그림 1] 키워드 추출 예시

[Fig. 1] Keyword Extraction Example

```
(1, 132)      0.2729834677133435
(1, 1363)     0.1802074280491075
(1, 1076)     0.09480034238675493
(1, 1047)     0.10259548979624174
(1, 274)      0.3504430498413725
(1, 405)      0.19122080612788622
:     :
(960, 1047)   0.2967734451595529
(960, 1253)   0.19072495091068262
(960, 887)    0.2517098832799956
(960, 691)    0.1554698674651855
(960, 1238)   0.14362122180656733
(961, 519)    0.40389267159114245
(961, 515)    0.2942365069277685
(961, 514)    0.2877607054382758
```

[그림 2] 텍스트 데이터를 변환한 TF-IDF 피처 매트릭스 예시

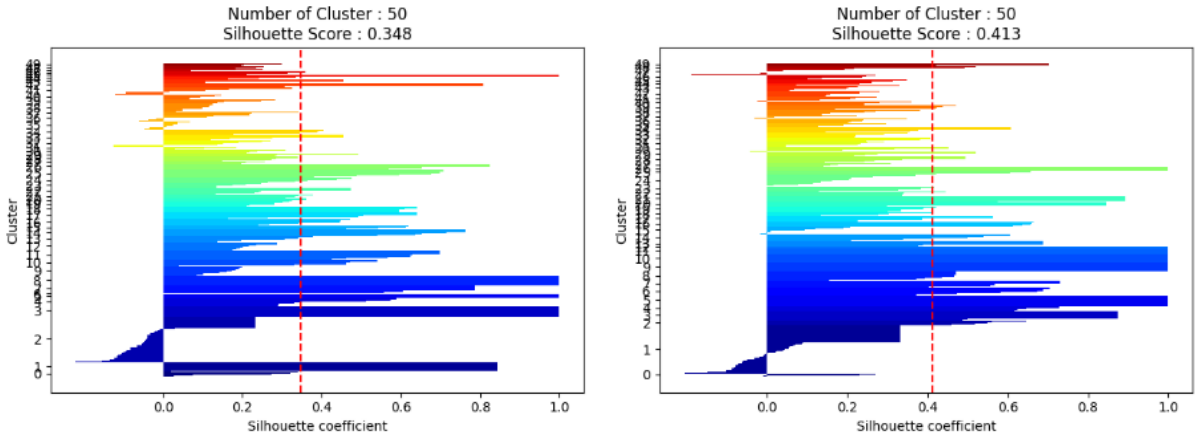
[Fig. 2] Example of TF-IDF Feature Matrix Converted from Text Data

3.3 모델 비교 및 결과

3.3.1 불용어 처리 후 클러스터링 성능 변화

데이터에서 클러스터링에 방해가 되는 '정보', '내용', '관련', '데이터', '제목', '설명', '자료',

'제공'등의 단어들을 불용어 처리한 후 클러스터의 수를 50개로 설정했을 때, 실루엣 점수가 0.348에서 0.413으로 향상되었다. 불용어 처리가 클러스터의 의미를 더욱 명확하게 드러나게 하고, 각 클러스터의 내부 응집력을 강화하여 클러스터링의 성능이 향상된 것을 확인했다.

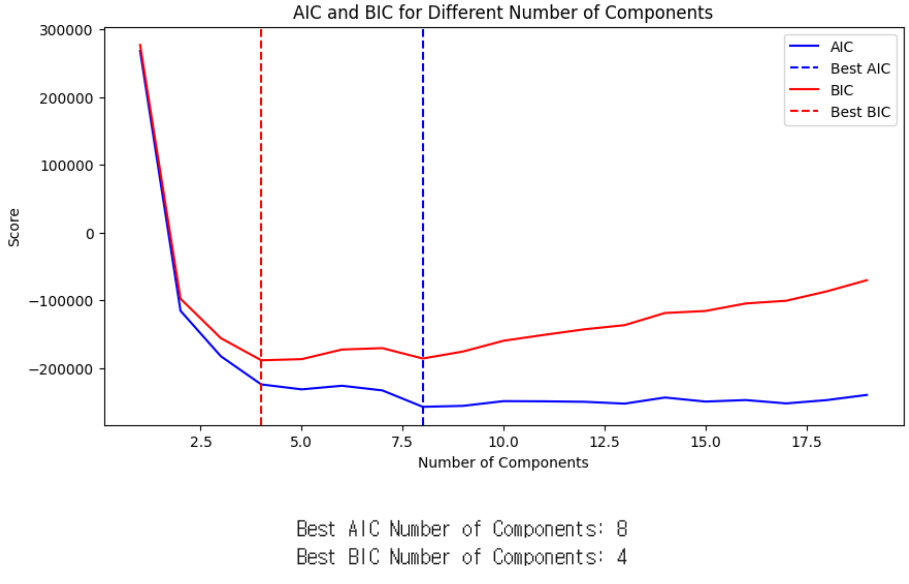


[그림 3] 불용어 처리 후 클러스터링 성능 비교

[Fig. 3] Clustering Performance Comparision after Stopword Processing

3.3.2 GMM(AIC, BIC)비교

GMM 모델로 수행한 클러스터링 결과, AIC 값이 최소일 때는 Components가 8이고, BIC 값이 최소일 때는 4임을 확인했다. 이는 각각 정보 기준을 최소화하는 데 가장 적합한 클러스터의 개수를 나타낸다. 또한 Components가 10 미만일 때는 클러스터링이 잘 이루어졌으나, 그 이상의 값에서는 불규칙한 클러스터링 결과가 나타났다.



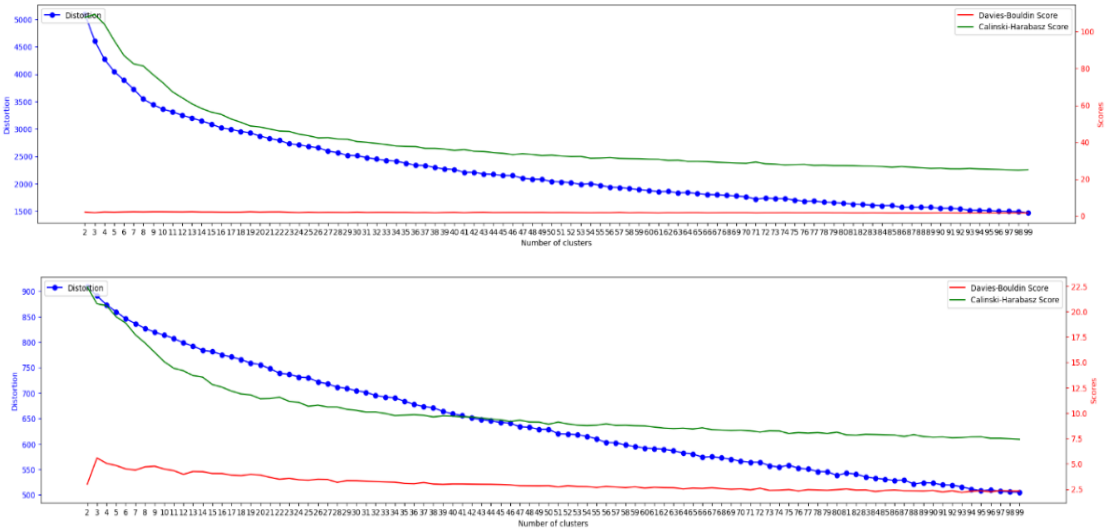
[그림 4] Components 에 따른 AIC and BIC 비교

[Fig. 4] AIC and BIC Comparison according to Components

3.3.3 k-means(konlpy vs ChatGPT keword 추출)

descripsion에서 추출한 키워드의 평균 수는 konlpy가 15.67이고, ChatGPT는 4.19인 것을

확인하였다. Elbow Method, Davies-Bouldin Index, Calinski-Harabasz Index를 사용하여 두 방법을 비교한 결과, konlpy로 추출한 키워드를 활용한 클러스터링이 ChatGPT보다 더 우수한 결과를 보였다. 이를 통해 키워드 추출 개수 및 추출 방법이 클러스터링의 품질에 영향을 미침을 확인 할 수 있었다.

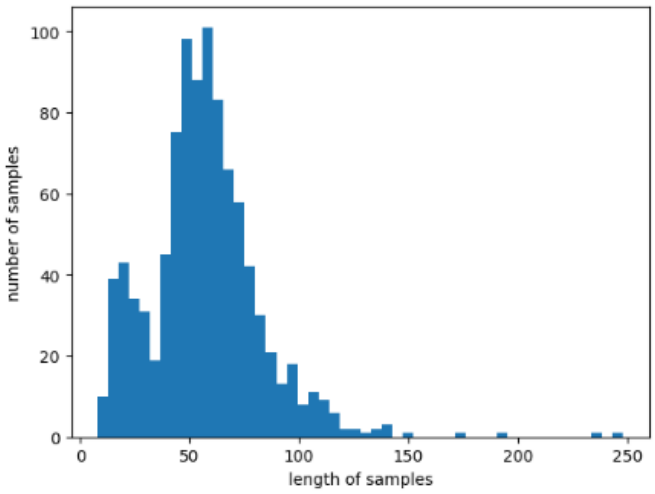


[그림 5] konlpy와 ChatGPT 키워드 추출 k-평균 성능 비교

[Fig. 5] konlpy vs ChatGPT Keyword Extraction k-means Performance Comparision

3.3.4 공공데이터 문제점 분석

descripsion의 최소 길이가 8글자로 짧고, 최대 길이가 248글자로 큰 차이가 있어 클러스터링 과정에서 왜곡을 일으킬 수 있다. 또한 평균 길이가 56.9로, 최대 길이와의 차이가 크므로 데이터 간의 거리나 유사성을 측정하는 데에도 영향을 미친다. 최소 길이가 8글자로 비교적 짧은 편이기 때문에 문장의 문맥이 충분히 반영되지 않을 수 있다. 이는 키워드 추출 및 클러스터링 결과에 일부 중요한 정보가 누락될 수 있다는 것을 알 수 있다.



[그림 6] 설명 길이 분포

[Fig. 6] Description Length Distribution

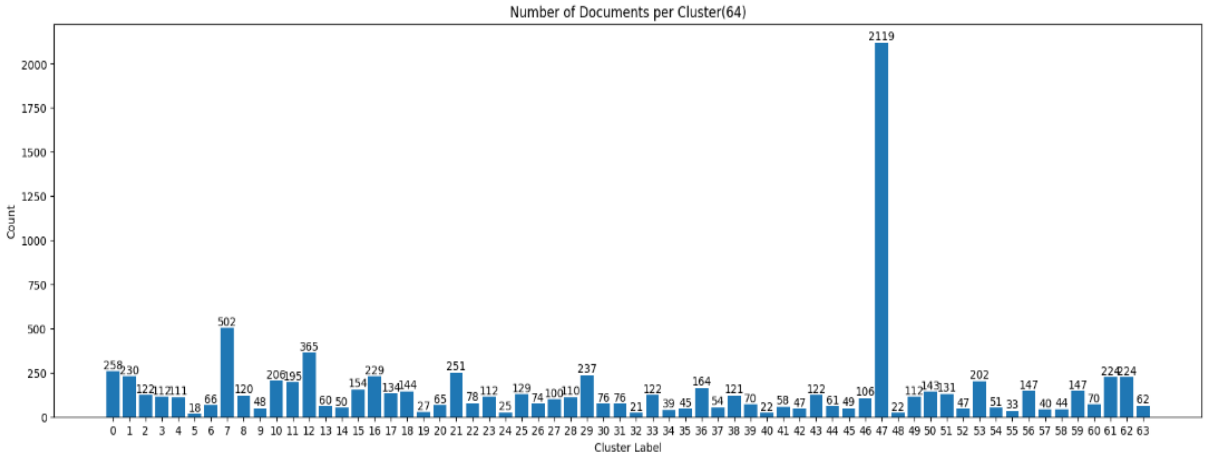
[표 1] 설명 최대, 최소 및 평균 값

[Table 1] Descripsion Max, Min and Avg Value

	Description Length
Max	248
Min	8
Avg	56.9

또한 k-means, DBSCAN, GMM으로 클러스터링 한 결과 모두 무분별한 cluster가 있었다. 각각의 cluster의 상위 10개의 feature를 확인한 결과 '정보', '내용', '관련', '데이터', '제목', '설명', '자료', '제공', '목록정보', '상세정보', '목록', '조회' 등과 같은 키워드로 구성 된 것을 확인 했다. 각 데이터들이 일관적으로 구체적이고 유의미한 키워드들이 빠져있다면, 각 클러스터가 실제로 나타내는 내용이나 특징을 파악하는데 어려움을 겪을 수 있다[10].

클러스터 간의 특징이 유의미하게 차이나지 않는다면, 클러스터링 알고리즘이 데이터의 다양성을 충분이 파악하지 못했을 가능성이 있다. 따라서 의미 없는 단어나 일반적인 단어를 제거하여 각 클러스터의 유일성을 강조할 필요가 있다. 또한 공공데이터를 등록할 때 descripsion의 최소 길이, 최대 길이의 제약을 주어 데이터의 불균형을 줄이는 것이 중요하다. 또한 의미 없는 단어나 일반적인 단어를 제거하여 평가지표이 성능이 향상된 것을 알 수 있었다.



[그림 7] 각 클러스터 당 개수 비교

[Fig. 7] Comparision of Numbers per Cluster

```
#####Cluster Num: 47

상위 10개 feature단어들:
 ['위치', '문화', '이미지', '대회', '공원', '홈페이지', '사진', '기관', '경상남도', '등록']

성남시, 하천, 탄천, 시설물, 관리, 주체, 위치, 시설, 면적, 단위
동대문구, 공동, 제조, 사업장, 창업, 지원, 센터, 입주, 기관
중랑구, 민, 체육, 센터, 중랑, 문화, 체육관, 면목, 체육관, 프로그램
용인시, 골프장
전국, 국립공원, 자연, 경관, 동식물, 모습, 촬영, 여, 이미지, 파일, 구성, 평균, 해상도, 픽셀, 웹 디자인, 활용, A4, 크기, 인쇄물, 제작
국립공원, 탐방, 자연, 관찰, 탐방, 안내소, 야영장, 공중화장실, 안내, 표지판, 시설물, 통계
국립공원, 공단, 촬영, 동영상, 경관, 생태, 탐방, 민간, 사업자, 국립공원, 주제, 홍보물, 제작, 때, 사용
마포구, 상암동, 소재, 마포, 주민, 편익, 시설, 일반, 체육, 프로그램
염, 리, 생활, 체육관, 일반, 프로그램
```

[그림 8] Cluster 47 상위 10개 feature단어 및 데이터 예시

[Fig. 8] Cluster 47 Top 10 Feature Words and Data Examples

다양한 출처에서 온 데이터는 다양한 형식과 길이를 가질 수 있다. 일관된 가이드라인을 적용하여 데이터의 일관성을 유지하면 데이터의 길이, 형식, 그리고 특정 키워드들의 표준화를 확보할 수 있다. 이를 통해 관련된 여러 공공데이터를 효율적으로 클러스터링하고, 관련 정보를 보다 쉽게 추출하여 활용할 수 있다.

3.4 향후 개선 방향 및 클러스터링 활용 방안 제시

공공데이터는 다양한 기관에서 제공되며, 이로 인해 데이터 품질과 일관성이 다르게 나타날 수 있다. 이는 클러스터링 결과에 부정적인 영향을 미칠 수 있다. 데이터의 품질을 검토하고, 표준화된 가이드라인을 통해 일관성을 확보해야한다. 특정 분야에 한정된 데이터나 부족한 양의 데이터로 인해 클러스터링 결과의 다양성과 신뢰성이 저하될 수 있다.

클러스터링 결과 무분별한 클러스터가 존재한다면 해당 클러스터에 속한 데이터들은 데이터들과의 유사성이 상대적으로 낮다고 판단 할 수 있으며 클러스터링을 통해 부족한 데이터를 파악할 수 있고, 이를 보완하거나 추가적인 조사가 필요한 부분으로 도출할 수 있다. 더 나아가 클러스터링을 적극 활용하면 데이터 수집 및 분석 전략을 개선하는 데에 도움을 줄 수 있다. 예를 들어 클러스터링을 통해 동일한 클러스터에 속하는 데이터들은 유사한 특성을 가지므로, 이들을 함께 활용함으로써 자원을 효과적으로 공유할 수 있으며, 클러스터링 결과에서 이상치를 식별하고, 이를 처리하여 데이터의 일관성을 높일 수 있다. 또한 키워드 추출 방법에 따라 클러스터링 성능이 상이하게 나타나는 것은 키워드의 양과 품질이 클러스터링에 중요한 역할을 하는 것이며 데이터 등록 시 키워드에 대한 강조나 표준화된 방법을 도입함으로써 클러스터링 뿐만 아니라 검색 및 접근성 향상, 데이터 이해와 분류, 데이터의 활용성 등 다양한 역할을 할 수 있을 것이다.

4. 결론

공공데이터의 클러스터링을 통해 데이터 품질과 일관성에 대한 문제점을 도출하고, 클러스터링이 데이터 수집과 분석 전략의 개선에 어떤 역할을 할 수 있는지를 탐색하였다. 공공데이터는 설명 길이, 형식, 일관성 부족, 표준 부재, 그리고 데이터 통합의 어려움 등의 문제에 대한 개선이 필요하다. konlpy로 추출한 키워드를 활용한 클러스터링이 ChatGPT보다 더 우수한 결과를 나타냈다. 키워드의 양과 품질이 중요한 역할을 한다는 것이며 데이터 등록 시 키워드에 대한 강조나 표준화된 방법의 도입이 필요하다. GMM 모델을 적용한 클러스터링에서는 Components가 10 미만일 때는 클러스터링이 잘 이루어졌으나, 그 이상의 값에서는 불규칙한 클러스터링 결과가 나타났다. 이는 해당 클러스터에 속한 데이터들은 상대적으로 유사성이 낮다고 판단할 수 있다. 이를 통해 부족한 데이터를 파악하고, 보완하거나 추가적인 조사가 필요한 부분으로 도출할 수 있다.

본 논문에서는 공공데이터의 효과적인 활용을 위해 데이터의 품질 향상과 클러스터링 기법의 적극적 활용이 중요하다는 점을 강조하였다. 향후에는 표준화된 데이터 등록 가이드라인의 도입 및 추가적인 연구를 통해 이러한 문제에 대한 더 나은 해결책이 제시될 것으로 기대한다.

References

[1] Bo-ok Lee, A Study on Legal Improvement Measures to utilize the re-use of public sector information - Focusing on legal improvement measures through the enactment of the National Data Framework Act, SungKyunKwan Law Review, (2021), Vol.33, No.2, pp.623-688.
DOI: 10.17008/SKKLR.2021.33.2.019

[2] Tae-il Lee, Kwan-hyun Kim, Ji-hyun Lee, Su-chul Lee, An Efficient BotNet Detection Scheme Exploiting Word2Vec and Accelerated Hierarchical Density-based Clustering, Journal of Internet Computing and Services, (2019), Vol.20, No.6, p.11-20.
DOI: 10.7472/JKSII.2019.20.6.11

[3] Mi-young Min, Gun-Up Jang, Eui-Kyeong Hong, Korean Document Clustering with Frequency of words in K-means and DBSCAN, Korean Institute of Information Scientists and Engineers, (2009), Vol.36, No.2, pp.12-13

[4] Gyeong-Yong Heo, Seong-Hoon Kim, A New Clustering Method for Minimum Classification Error, Journal of the Korea Society of Computer and Information, (2014), Vol.19, No.7, p.1-8.
DOI: 10.9708/JKSCI.2014.19.7.001

[5] Jung-woo Lee, Sung-Jin Kim, Min-Kwan Kim, Jae-Young Yoo, Hyuk Hahn, Hun Park, Chang-Hee Han, Classification of Performance Types for Knowledge Intensive Service Supporting SMEs Using Clustering Techniques: Focused on the Case of K Research Institute, The Journal of Society for e-Business Studies, (2017), Vol.22, No.3, p.87-103.
DOI: 10.7838/JSEBS.2017.22.3.087

[6] Min-seo Jang, Su-jin Oh, Ung-Mo Kim, Article Analytic and Summarizing Algorithm by facilitating TF-IDF based on k-means, Proceedings of the Korea Information Processing Society, (2018), Vol.25, No.1, p.271-274.
DOI: 10.3745/PKIPS.Y2018M05A.271

[7] Yun-Hui Woo, Hyon-Hee Kim, Reconstruction of Categories on the National Petition Site Using K-Means clustering and Topic Modeling, Korea Information Processing Society, (2019), Vol.26, No.1, p.302-305.
DOI: 10.3745/PKIPS.Y2019M05A.302

[8] Shin-won Lee, Comparison of Initial Seeds Methods for K-Means Clustering, Journal of Internet Computing and Services, (2012), Vol.13, No.6, p.1-8.
DOI: 10.7472/JKSII.2012.13.6.1

[9] Seon-Ja Lim, Sung-Dae Youn, Data classification using K-means clustering, Proceedings of the Korea Information Processing Society, (2020), Vol.27, No.2, p.1087-1088.
DOI: 10.3745/PKIPS.Y2020M11A.1087

[10] Young-Soo Kim, Byoung-Yup Lee, Multi-class Support Vector Machines Model Based Clustering for Hierarchical Document Categorization in Big Data Environment, The Journal of the Korea Contents Association, (2017), Vol.17, No.11, pp.600-608.
DOI: 10.5392/JKCA.2017.17.11.600

[11] Jong-Sub Lee, Vocabulary Recognition Performance Improvement using k-means Algorithm for GMM Support, Journal of Digital Convergence, (2015), Vol.13, No.2, pp.135-140.
DOI: 10.14400/JDC.2015.13.2.135

[12] Eun-Suk Choi, Jeong-Hun Kim, Aziz Nasridinov, Sang-Hyun Lee, Jeong-Tae Kang, Kwan-Hee Yoo, Detection of the Defected Regions in Manufacturing Process Data using DBSCAN, The Journal of the Korea Contents Association, (2017), Vol.17, No.7, p.182-192.
DOI: 10.5392/JKCA.2017.17.07.182

[13] Byeong-Sik Kim, Ju-hyun Shin, A Method for Measuring Similarity Measure of Thesaurus Transformation Documents using DBSCAN, Journal of Korea Multimedia Society, (2018), Vol.21, No.9, p.1035-1043.
DOI: 10.9717/KMMS.2018.21.9.1035

[14] Yun-A Hur, Dong-Yub Lee, Kue-Kyeng Kim, Won-Hee Yu, Heui-Seok Lim, A System for Automatic Classification of Traditional Culture Texts, Journal of the Korea Convergence Society, (2017), Vol.8, No.12, pp.39-47.
DOI: 10.15207/JKCS.2017.8.12.039