관광지 추천을 위한 클러스터링 최적화 군집수 결정

여해진 [1], 조인휘 [2]
[1] 한양대학교 컴퓨터 · 소프트웨어학과 석사과정
[2] 한양대학교 컴퓨터 · 소프트웨어학과 교수

hanhee1203@hanyang.ac.kr, iwjoe@hanyang.ac.kr

Clustering Optimization Cluster Count Determination for Tourist Destination Recommendation

Hae-Jin Yeo[1], In-Whee Joe[2]
[1]Dept. of Computer Science, Han-Yang University
[2]Dept. of Computer Science, Han-Yang University

요 약

factor 들이 많은 데이터의 군집화는 어려움을 요한다. K-means 클러스터링을 사용하여 군집화를 할 때, 각 데이터들이 가진 factor 의 개수가 상이한 경우 비슷한 성향을 가진 데이터임에도 불구하고 클러스터링이 적합하게 되지 않는 현상이 발생한다. 이러한 문제점을 해결하기 위해 최적의 군집화 개수를 결정하는 실루엣 기반 방법을 제안하고 제안기법의 성능을 평가한다.

1. 서론

여행에 대한 일정 생성 시 비슷한 성격을 띄는 관광지를 추천해주면 사용자는 더욱 만족스러운 관광을 경험할 수 있다. 위에 대한 방법으로 POI(Point Of Interest) 데이터 클러스터링을 통해 유사한 관광지들의 추천에 활용한다. POI 데이터는 클러스터링을 할 수 있는 factor 들을 가지고 있지 않은, 즉 정제되지 않은 데이터셋이다. 따라서 POI 에 factor 들을 설정해야 한다. 그러나 각 POI 들의 factor 를 직접 설정한다면 개인이 방문했던 곳이더라도 주관적인 의견이 담길 수 있다. 이에 근거하여 Naver 리뷰 태그를 웹 크롤링 방식을 사용해 POI 들의 특성을 알 수 있는 factor 들을 추가한다.

최종적으로 POI 들의 factor 들을 산출한 뒤 클러스터링을 하기 위해 K-means 클러스터링 방식을 사용했다. K-means 클러스터링 방식은 K 개의 클러스터로 묶는 알고리즘이며 초기 K 값을 5 로 설정하고 클러스터링을 실행한 결과, POI 가 가진 factor 가 상이할 때 클러스터링이 잘 되지 않는 현상을 발견했다. 따라서 최적의 K 값을 찾고, 적절한 클러스터링에 대한 분할 방법론에 대해 결정한다.

2. 관련 연구

K-means 클러스터링에서 군집 개수를 정하기 위해 보편적으로 사용하는 2 가지 방식이 있다. Elbow method[1]는 클러스터 간 거리의 합을 나타내는 파라미터 값이 급격히 감소하는 구간을 K 값으로 설정한 뒤 클러스터링 하는 방식이다. Silhouette coefficient method[2]는 각각의 데이터가 할당된 클러스터링 내 데이터들과 얼마나 근접하게 군집화 되어있는지, 다른 클러스터의 데이터들과는 어느 정도의 격차가 있는지에 대한 정보를 수치로 나타내는 방식이다. 이 방식은 K 값을 Silhouette coeifficient 를 이용하여 1 과 가까우면 클러스터링이 잘 되었다 판단한다. 그림 1 과 2 는 각각 Elbow method 와 Silhouette coefficient

```
k :  5  / inertia :  1811.3230217657251
k :  6  / inertia :  1811.3230217657251
k :  7  / inertia :  1811.3230217657251
k :  8  / inertia :  1811.3230217657251
k :  9  / inertia :  1811.3230217657251
k :  10  / inertia :  1811.3230217657251
k :  11  / inertia :  1811.3230217657251
k :  12  / inertia :  1811.3230217657251
k :  13  / inertia :  1811.3230217657251
k :  14  / inertia :  1811.3230217657251
k :  15  / inertia :  1811.3230217657251
```

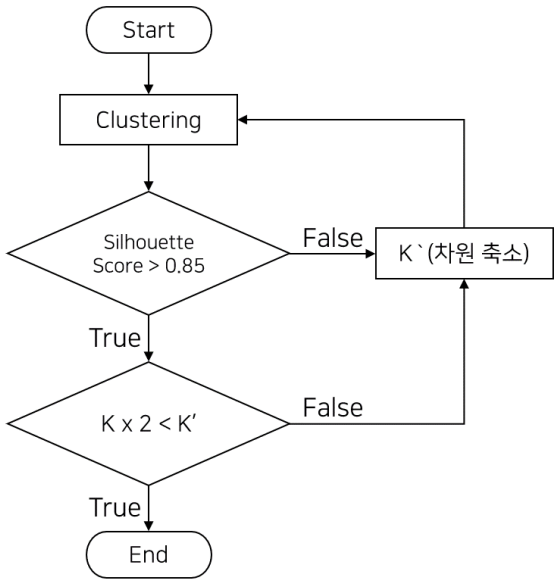
(그림 1) K-means Elbow method

```
k :  5  / score :  0.25954277833441314
k :  6  / score :  0.25954277833441314
k :  7  / score :  0.25954277833441314
k :  8  / score :  0.25954277833441314
k :  9  / score :  0.25954277833441314
k :  10  / score :  0.25954277833441314
k :  11  / score :  0.25954277833441314
k :  12  / score :  0.25954277833441314
k :  13  / score :  0.25954277833441314
k :  14  / score :  0.25954277833441314
k :  15  / score :  0.25954277833441314
best n :  5  | best score :  0.25954277833441314
```

(그림 2) K-means Silhouette coefficient method

method 를 통해 나온 값이다. 그림에서와 같이 클러스터링 개수에 따른 결과값들이 동일하다는 것은 적절한 K 값이 도출이 되지 않았다는 것을 의미한다.

3. 제안 방법



(그림 3) 제안 방식의 Flow chart

K-means 클러스터링의 K 값을 얻는 직관적인 방법으로는 K 값을 직접 조절하며 클러스터링이 잘 되었는지 확인할 수 있지만 위의 방법들에 비해 비효율적인 방법이다. 따라서 많은 factor 들 중 상이한 factor 를 가졌을 때 적절한 군집 개수를 정하는 것은 까다롭다. 이에 착안하여 새로운 군집 개수를 정하는 방법에 대해 제안한다.

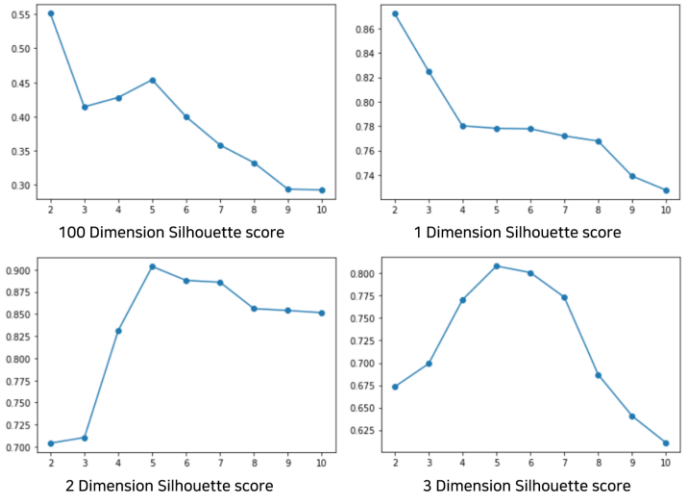
그림 3 은 제안 방식의 흐름도를 나타낸다. K-means 를 통해 임의의 K 값을 설정한 뒤 클러스터링을 시행한다. 초기 설정된 100 차원에서 Silhouette score 를 평가하게 되는데, Silhouette score 는 차원에 따라 적절한 K 값을 평가해주는 지표이다. Silhouette score 가 전체 그래프에서 0.85 라는 값을 넘기지 않았을 때엔 최적의 K 값이 없다고 판단하여 차원의 축소가 일어나게 되는 것이다. 초기 K 값은 100 차원에서의 최대 Silhouette score 를 나타낸 K 값으로 설정한 뒤, 차원의 축소를 통해 초기 K 값의 2 배인 값과 Silhouette score 가 0.85 이상인 최적의 K 값을 찾는 것이다.

K x 2 < K'는 군집 개수 K 의 증가가 가능하다면 서론에 대한 결과로 관광지의 추천이 더욱 세분화되어 사용자의 취향에 알맞은 관광지를 제안하는 것이 이루어질 것이라 사료된다.

결과론적으로 군집화에 적합한 최적의 K 값과 최적의 차원을 도출해낼 수 있는 것이다.

4. 실험 결과

제안 방법에 기초하여 초기 데이터가 가진 factor 는 100 여개로 군집화를 하기에 어려울 것이라 판단하여 PCA(Principal Componenet Analysis)를 이용해 데이터를 차원 축소시켰다.



(그림 4) 차원 축소에 따른 Silhouette score

그림 4 는 Silhouette score 를 이용하여 클러스터링의 군집 개수를 정하는 그래프이며, 각각 초기 차원인 100 차원과 1, 2, 3 차원의 결과를 시각화한 그래프이다. 그림에서와 같이 100 차원과 1 차원의 군집 개수는 2 개로 설정이 되어있는 반면 2, 3 차원으로 축소한 데이터의 결과로는 적절한 군집 개수가 5 개로 설정이 되었다.

2, 3 차원의 군집수는 흐름도에서의 K x 2 < K'을 만족한다. 그러나 3 차원의 경우 표 1 에서와 같이 Best silhouette score 가 0.85 를 넘기지 못하는 수준에 그쳐 최적의 차원으로 적합하지 않다. 표 1 에서 볼 수 있듯이, 초기 차원의 Best silhouette score 는 가장 높은 값임에도 불구하고 0.5 에 달하는 Score 가 도출이 되었다.

Dimensions	K (best)	Silhouette Score (best)
100	2	0.5523
1	2	0.8721
2	**5**	**0.8725**
3	5	0.8081

(표 1) 차원 축소에 따른 Best silhouette score

이는 군집 개수 K 개가 선정이 되었다 하더라도 정확도가 떨어지는 즉, 적절히 군집화가 되지 않았다는 뜻을 의미한다. 1 차원과 2 차원의 Best silhouette score 는 유사하게 0.87 이라는 값을 나타내었지만, 1 차원의 Best K 값은 2 로 흐름도의 K x 2 < K'를 충족시키지 못하였다.

Dimensions	K (worst)	Silhouette Score (worst)
100	10	0.2768
1	**10**	**0.7275**
2	2	0.7036
3	10	0.6107

(표 2) 차원 축소에 따른 Worst silhouette score

표 2 는 표 1 과 상반되는 Worst silhouette score 를 나타낸다. 초기 차원은 이전 Best score 와 같이 가장 낮은 Silhouette score 값이 도출되었으며, 1, 2 차원에서는 Worst silhouette score 또한 0.7 이상의 높은 값이 반환되었다. 흐름도의 Silhouette score 의 Threshold 값을 감소시킨다면 1 차원의 K 값 또한 최적의 군집화 개수라 여겨질 수 있다.

5. 결론

POI 추천에 대한 다양성은 조금 더 섬세한 추천이 가능하며, 사용자에게 선택지가 주어져 원하는 곳을 손쉽게 선택하여 방문할 수 있다는 것을 뜻한다. 이에 따라 POI 들의 특성에 따른 클러스터링을 K-means 클러스터링 방법을 통해 진행하였다. 이때 최적의 군집화 개수 K 에 대한 방법으로 Silhouette score 과 차원의 축소를 통해 나타냈다. 축소된 차원에 따라 변하는 Silhouette score 를 확인하였으며, Worst silhouette score 또한 기준치에 가까운 값을 도출한 차원들도 존재했다. 따라서, 다양한 관광지 추천의 목적으로 구현하고자 했던 특성상 군집화의 개수가 많으며 정확도가 높은 차원을 선택하였다. 추후 구현의 목적이 상이할 경우에도 높은 Worst silhouette score 의 값들을 통해 Threshold 를 조절해가며 사용할 수 있다는 점을 시사한다.

본 논문은 2023 년도 한국콘텐츠진흥원의 지원을 받아 수행된 연구임 [R2022020116, AI 기반 관광객 상황인식 및 관광정보 큐레이션을 통한 맞춤형 관광 여정(Itinerary) 추천 플랫폼 기술개발]

참고문헌

[1] Syakur, M. A., et al. “Integration k-means clustering method and elbow method for identification of the best customer profile cluster”. In:IOP conference series: materials science and engineering. IOP Publishing, 2018. p. 012017.

[2] Shahapure, Ketan Rajshekhar; NICHOLAS, Charles. “Cluster quality analysis using silhouette score”. In:2020 IEEE 7th international conference on data science and advanced analytics (DSAA). IEEE, 2020. p. 747-748.