Mobile Data Traffic 특성 분석을 위한 시계열 K-평균 군집화

이상연, 유현민, 김병석*, 유흥렬*, 홍인기
경희대학교, *(주)케이티

sangyeon@khu.ac.kr, yhm1620@khu.ac.kr, *b.kim@kt.com, *heungryeol.you@kt.com, ekhong@khu.ac.kr

Time-Series K-means Clustering for Mobile Data Traffic

Sang-Yeon Lee, Hyun-Min Yoo, Byungsuk Kim*, Heung-Ryeol You*, Een-Kee Hong
Kyunghee University, *KT

요 약

통신이 발전하면서 모바일 데이터 트래픽이 점점 증가하고 있으나, 이를 수용하기 위한 기지국 용량에는 한계가 있다. 따라서 한정된 자원을 바탕으로 모바일 데이터 트래픽을 수용하기 위해서는 모바일 트래픽의 특성을 분석하고, 이를 바탕으로 효율적인 자원할당이 필요하다. 따라서 모바일 트래픽의 특징을 살펴보기 위해 각 기지국들의 모바일 데이터 트래픽들을 군집화(Clustering)하여 같은 지역끼리 분류가 되는지 확인하였다. 클러스터링 결과는 강남역과 인천공항이 지역적으로 가장 잘 나뉘었고, 코엑스(쇼핑몰)와 광화문(사무실)이 같은 클러스터에 묶였다.

Ⅰ. 서 론

모바일 기술과 셀룰러 네트워크의 지속적인 발전에 따라 스마트폰, 태블릿, 웨어러블 기기와 같은 모바일 장비의 수와 휴대 전화 가입자의 수가 빠르게 늘어나고 있다. 모바일 장비 및 휴대 전화 가입자 수의 증가와 더불어 수많은 기기에서 생성된 데이터 또한 폭발적으로 증가하고 있다. 2022년 에릭슨(ericsson)의 조사의 따르면 2022년 3분기 월평균 트래픽 데이터양은 107.58EB(엑사바이트)로 지난 1분기 대비 7%의 상승폭이 있었고, 2년 전과 비교하면 약 2배로 트래픽이 증가했다. 2015년 1분기의 월평균 트래픽 데이터양은 3.58EB 이므로 7년 동안 모바일 트래픽이 30배나 오른 것이다.[1] 이처럼 폭발적인 트래픽 증가로 인해 모바일 네트워크의 정체는 빈번하게 일어나고 이를 수용하기 위한 시스템 용량과 효율적인 트래픽 운영 방식에 대한 필요성이 증가하고 있다. 본 논문은 모바일 트래픽의 특성을 확인하기 위해 k-means와 k-shape를 이용해 분류기준을 바꾸면서 어떤 경우가 가장 분류가 잘 되는지를 확인했다.[2][3][4]

Ⅱ-1. 데이터 설명

데이터 셋은 한국 이동통신 A사 기지국들의 모바일 데이터 트래픽을 사용하였다. 해당 데이터셋은 지역적으로 크게 강남역, 코엑스(쇼핑몰), 인천공항, 김포공항, 광화문(사무실) 5가지로 나뉜다. 데이터 수집 기간은 2022.09.17.~202210.16으로, 한 달간의 트래픽이 1시간 단위로 구성되어 있다. 강남역은 총 89개의 기지국이 있고, 쇼핑몰은 총 38개의 기지국, 인천 공항은 48개의 기지국, 김포공항은 9개의 기지국, 사무실은 8개의 기지국이 있다. 해당 데이터를 이용하여 라벨링 없이 클러스터링 했을 때, 어떻게 지역적으로 나뉘는 지 확인해볼 것이다.

Ⅱ-2. 데이터 분석 결과

데이터 분류는 클러스터의 개수(4, 5, 6, 8), 클러스터링 방법(K-means, K-shape), 시계열비교 (Euclidean, DTW), 시계열 길이(1주, 4주), 데이터 가공 여부(스케일, 원본)를 기준으로 하였다. 위의 기준을 통해 분석한 결과, 일주일 단위로 스케일된 데이터를 euclidean으로 K-means에서 k를 4 혹은 8로 클러스터링 했을 때, 지역적으로 가장 잘 분류되었다.

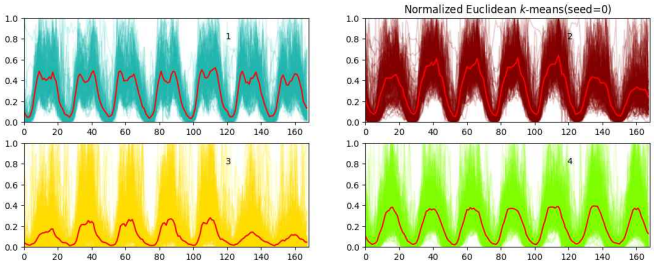

그림 1 k-means euclidean에서 k=4일 때의 각 클러스터의 평균

	강남역[%]	쇼핑몰[%]	인천공항[%]	김포공항[%]	사무실[%]
cluster1	1.12	0.66	51.56	25.00	0.00
cluster2	46.07	1.97	1.56	0.00	0.00
cluster3	21.35	56.58	37.50	38.89	90.63
cluster4	31.46	40.79	9.38	36.11	9.38

표 1 k-means euclidean k=4에서의 지역대비 각 클러스터의 포함률

그림1과 표1은 일주일 단위로 스케일된 데이터를 거리를 euclidean으로 계산하고, k=4로 하였을 때의 K-mean로 클러스터링한 결과이다. 1번 클러스터의 경우 전체 인천공항의 51%를 포함하고 있고, 다른 지역의 비율이 낮으므로 해당 클러스터는 인천공항의 지역적 특성이 있다고 판단된다. 2번 클러스터의 경우 전체 강남역의 41%를 포함하고 있고, 다른 지역의 비율이 낮으므로 해당 클러스터는 강남역의 지역적 특성을 가진다. 3번 클러스터의 경우 전체 쇼핑몰의 46%, 전체 사무실의 90%를 차지하고 있다. 하지만, 1번, 2번 클러스터와 달리 다른 지역의 트래픽도 포함하고 있으므로 해당 클러스터가 확실하게 쇼핑몰, 사무실의 특성이 있다고 하기는 힘들다.

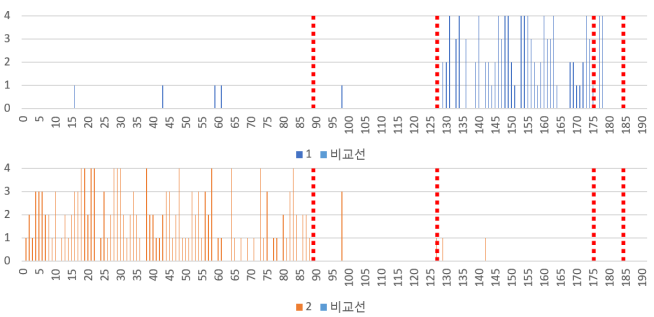

그림 2 k-means eucdliean k=4에서 1,2번 클러스터

그림2는 점선을 기준으로 왼쪽부터 강남역, 코엑스(쇼핑몰), 인천공항,

김포공항, 광화문(사무실)을 의미한다. 앞서 설명한 것처럼 1번, 2번 클러스터는 특정 지역의 기지국들이 모여 있는 것을 알 수 있다.

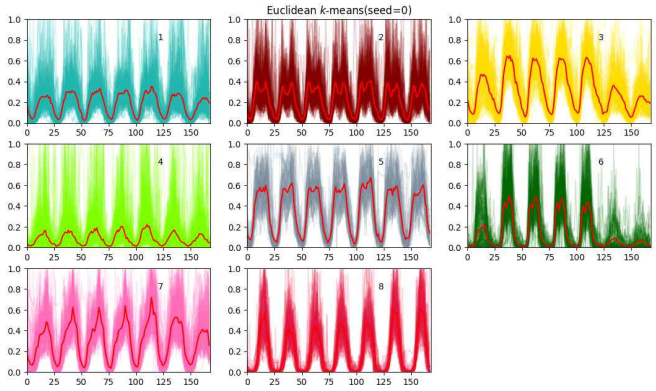

그림 3 k-means euclidean에서 k=8일 때의 각 클러스터의 평균

	강남역[%]	쇼핑몰[%]	인천공항[%]	김포공항[%]	사무실[%]
cluster1	28.65	4.61	1.04	0.00	0.00
cluster2	1.40	0.66	46.88	8.33	6.25
cluster3	24.44	0.00	3.65	0.00	0.00
cluster4	12.36	35.53	28.13	22.22	34.38
cluster5	1.12	1.97	16.15	16.67	0.00
cluster6	4.78	25.00	2.60	16.67	59.38
cluster7	25.84	0.66	1.56	0.00	0.00
cluster8	1.40	31.58	0.00	36.11	0.00

표 2 k-means euclidean k=8에서의 지역대비 각 클러스터의 포함률

표2, 그림3과 그림4는 일주일 간격으로 스케일된 데이터를 거리를 euclidean으로 계산하고, k=8로 하였을 때의 K-means로 클러스터링한 결과이다. 1번, 3번, 7번 클러스터는 다른 지역에 비해 강남역의 비율이 높다. 해당 클러스터들을 합산하면 전체 강남역의 78.93%를 차지한다. 이를 통해 해당 클러스터들은 강남역의 지역적 특성을 가지고 있음을 알 수 있다. 2번, 5번 클러스터는 다른 지역에 비해 인천공항의 비율이 높은 것을 알 수 있다. 해당 클러스터는 전체 인천공항의 63%가 포함되어 있고, 그 다음으로 김포공항의 비율이 높다. 즉, 해당 클러스터는 인천공항의 특성을 지닌다. 4번, 6번, 8번 클러스터는 전체 쇼핑몰의 92.11%, 전체 사무실의 93.76%를 차지한다. 해당 클러스터는 쇼핑몰이나 사무실이 묶여있는 것을 보아 인도어의 특성을 띄는 클러스터라고 할 수 있다. 클러스터의 갯수를 4개, 8개일 경우를 비교하여 봤을 때, 클러스터를 1개씩 보면 4개일 경우가 지역대비 클러스터 포함률이 높지만, 특정 지역의 포함률이 높은 클러스터들을 합쳐서 본다면 클러스터가 8개일 경우가 특정 지역이 모이는 포함률이 높은 것을 알 수 있다.

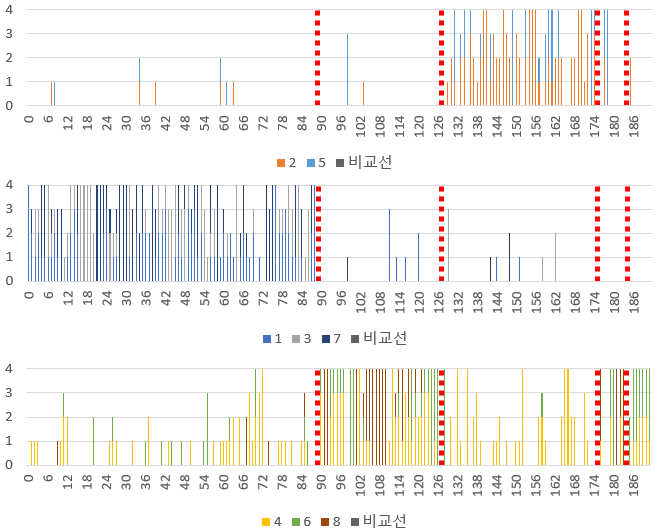

그림 4 k-means euclidean k=8에서 클러스터

Ⅲ. 결론

본 논문에서는 여러가지 분류기준 중에서 지역적인 특성이 가장 잘 드러난 2가지 클러스터링 결과를 보여주었다. 해당 결과를 통해 강남역, 코엑스(쇼핑몰), 인천공항, 김포공항, 광화문(사무실) 중에서 강남역과 인천공항은 확실히 분리되었고, 쇼핑몰과 사무실은 같은 클러스터에 묶이는 것을 확인하였다. 쇼핑몰과 사무실은 모두 인도어 트래픽이므로, 해당 클러스터를 인도어의 특성이 있는 클러스터라 생각한다면, 해당 클러스터에 있는 강남역의 기지국 역시 인도어의 특성을 띄는 기지국일 것이다. 향후 연구에는 해당 기지국들을 조사하여 인도어의 특성을 띄는 위치인지 확인할 것이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2021-0-02046)

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획 평가원의 지원을 받아 수행된 연구임(No. 2018-0-01659, 5G NR 기반 지능형 오픈 스몰셀 기술 개발)

참 고 문 헌

[1] Ericsson, “Ericsson Mobility Report,” 2022.[Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/november-2022

[2] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the royal statistical society. series c (applied statistics) 28.1 (1979): 100-108.

[3] Keogh, Eamonn, and Chotirat Ann Ratanamahatana. "Exact indexing of dynamic time warping." Knowledge and information systems 7.3 (2005): 358-386

[4] Hamerly, Greg, and Charles Elkan. "Learning the k in k-means." Advances in neural information processing systems 16 (2003).