시계열 모델을 활용한 위치 데이터의 시간적 패턴 분석

송하윤, 정준우, 이다솜
홍익대학교 컴퓨터공학과
hayoon@hongik.ac.kr, jwx999@naver.com, 000dasom@naver.com

Analysis on Temporal Pattern of Location Data with Time Series Model

Ha Yoon Song, Da Som Lee, Jun Woo Jung
Dept. of Computer Engineering, Hong-Ik University

요 약

시계열 분석은 이전 시점들의 데이터를 기반으로 미래 시점의 데이터를 예측하는 기술을 제공하며, SARIMA는 이러한 시계열 분석에서 활용되는 통계 모델의 일종이다. 본 연구는 직접 수집한 실시간 위치 데이터에 SARIMA를 적용하여 개인의 이동 패턴을 추출하고 이를 예측에 활용하는 전반적인 프로세스를 제작하였다. 첫째, DB에 업로드된 위치 데이터를 비지도 학습의 일종인 EM-clustering을 활용해 핵심 방문 장소들로부터의 거리에 따라 군집화했다. 둘째, 해당 장소에 입장하고 퇴장하는 시간 간격에 SARIMA를 적용해 주기성을 추출했다. 마지막으로, 이 주기성들을 군집의 중요도에 따라 순차적으로 분석하여 유의미한 예측 결과를 도출해냈다.

1. 서론

최근 모바일 기기의 기술적 발전과 보급률 증가로 인해 이를 활용한 위치 기반 응용 서비스의 사용이 증가하였다. 이에 따라 실시간으로 수집되는 위치 데이터의 양이 증가하였고, 이를 활용한 분석들이 국가 기관과 기업들에서 이루어지고 있다.

여러 데이터 분석 방법 중 본 논문에서 사용하고자 하는 방법은 시계열 분석이다. 시계열 분석이란 “예측될 변수의 과거의 자료에서 규칙적인 패턴을 탐색하고 미래에도 그러한 패턴이 반복된다는 가정하에 모델을 설정하여 미래 데이터를 예측하는 방법이다.” [1] 시계열 분석은 다양한 분야에 유용한 방법론을 제시하고 있는 만큼, 많은 주목을 받고 있는 분야이다.

시계열 분석은 인간의 이동 패턴에서 규칙적인 패턴을 발견하여 미래 위치에 대한 예측을 가능하게 한다. 이러한 인간의 이동 패턴 예측은 전염병 관리, 도시 계획, 교통량 예측뿐만 아니라 여러 모바일 애플리케이션에서 유용하게 사용된다.

본 연구에서는 직접 수집한 개인의 실시간 위치 데이터에 시계열 모델을 적용하여 주기성을 추출하고, 이를 바탕으로 개인의 미래 이동을 예측하는 일련의 프로세스를 제작하였다.

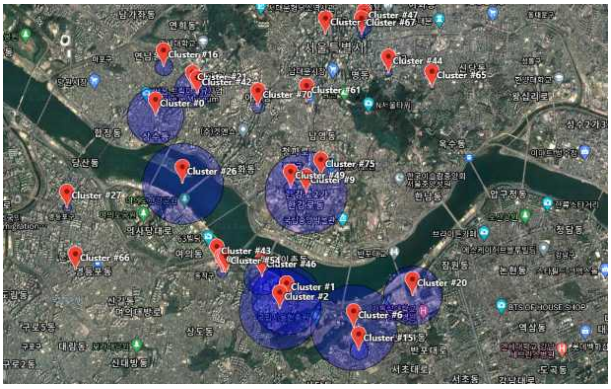
2. 자료 수집 및 전처리

(그림 1) 연구 프로세스

본 연구는 (그림 1)의 과정을 통해 진행하였다. 위치 데이터는 PEM 연구소에서 직접 Sports Tracker라는 스마트폰 애플리케이션을 이용하여 수집하였다. 수집된 데이터는 모두 PEM Data Base(DB)에 저장한다. DB에서 원하는 사용자의 데이터를 불러와 EM-clustering을 적용하면 사용자가 자주 방문하는 위치가 cluster로 도출된다. EM- clustering 결과를 활용하여 시계열 분석이 가능한 형태로 전처리한 뒤, 시계열 모델을 통해 데이터를 분석한다. 분석 결과는 다시 DB에 업로드 하여 다른 연구에 재사용될 수 있게 한다.

본 연구에서는 익명으로 수집하여 LJS라 이름 붙인 개인의 위치 데이터를 이용하여 이동 패턴을 분

석하였다. LJS의 위치 데이터에 EM-clustering을 적용하게 되면, (그림 2)에서 보듯 오랜 시간 머물러 있던 장소가 클러스터로 선정된다. LJS의 경우 cluster #0에 해당하는 홍익대학교(이하 학교)와 cluster #1,#2에 해당하는 자택에 가장 오랜 시간 머물렀기 때문에 본 연구에서는 이 두 가지 장소를 기준으로 이동 패턴을 예측하였다.



(그림 2) LJS 위치 데이터의 클러스터링 결과

애플리케이션을 이용하여 수집한 위치 데이터는 시각, 위도, 경도로 구성되어 있다. 수집된 데이터의 경우 초 단위로 기록이 돼 있어서 시간적 패턴을 추출하기에는 데이터의 양이 너무 방대했고, 또 누락된 부분도 많았기에 추가적인 전처리 과정이 필수적이었다. 본 연구에서는 두 가지 방법으로 데이터를 전처리하여 예측에 사용하였다.

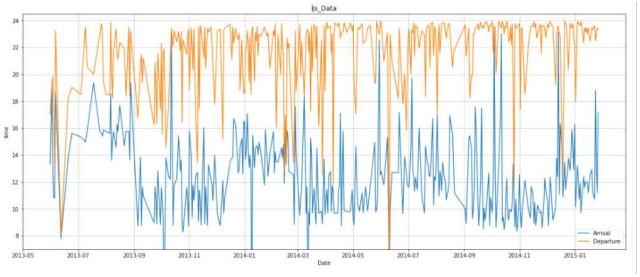
첫 번째 전처리 방법으로는, 해당 클러스터에 입장하고 퇴장하는 시각을 추출하였다. 그 결과로 (그림 3) 과 같은 결과를 얻을 수 있었다.

한편, (그림 3)의 데이터를 자세하게 살펴보면 입장 시간이 너무 늦거나, 퇴장 시간이 너무 빠르거나, 학교에 머무른 시간이 너무 짧은 데이터가 존재한다.

이러한 데이터는 시계열 모델과의 편차를 증가시키며 모델의 예측 정확도를 저하시킨다. 따라서 이러한 이상치를 제거하기 위해 머신러닝의 기법중 하나인 DBSCAN을 이용하였다.[2]

DBSCAN 적용 결과, 365일의 데이터 중 23일의 데이터를 제거할 수 있었다. (그림 3)과 (그림 4)를 비교해보면 이상치가 많이 사라진 것을 확인할 수 있다.

두 번째 전처리 방법으로는, 해당 클러스터에 방문하는 시간 간격을 추출하였다. 이 처리 방식에서는 이상치 제거를 적용하지 않았다. 첫 번째 전처리



(그림 3) LJS의 학교 방문 시각 데이터

(그림 4) 이상치를 제거한 데이터

방식은 이상치가 있는 것이 일일 방문 시각의 평균을 구하는 데에 방해가 됐었기에 이상치를 제거하는 것이 유의미했다. 하지만 두 번째 전처리 방식에서 이상치의 제거는 특정 방문 기록을 삭제하여 비정상적으로 시간 간격이 넓은 데이터를 만들어낸다. 따라서 이는 시계열 모델의 예측 정확도를 더욱 떨어뜨릴 뿐만 아니라 데이터의 의미를 훼손하는 것이 되므로 이상치를 제거하지 않고 진행했다.

3. ARIMA 모델과 SARIMA 모델

시계열 분석의 대표적인 통계적 예측 모델로 ARIMA와 SARIMA가 있다. 이는 일반적인 선형 회귀 모델에 쓰이는 데이터와 시계열 데이터의 구별되는 특징 때문에 별도로 연구되어온 모델이다. 시계열은 t시점과 t-h시점의 측정값이 연관되어 있는 경우가 많다. 이런 경우, 일반적인 회귀 연산을 적용하기 까다로워지므로 차분을 진행해야한다. 차분이란 t시점의 데이터에서 t-h시점의 데이터를 뺀 수열을 시계열 분석의 대상으로 치환하는 과정을 의미한다.[3]

차분을 통해 시계열이 정상성을 확보하게 되면 비로소 자기회귀(AR)와 이동평균(MA) 식을 선형 결합한 ARIMA 모델을 구성할 수 있다.

자기회귀(AR) 모형은 현재 관측자료(y_t)가 과거 자료($y_{t-1}, y_{t-2}, \cdots$)들과 비관측영향(α_t)의 선형결합으로 표시되는 모형이다.

$$AR(p): y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots \phi_p y_{t-p} + \alpha_t \quad (1)$$

이동평균(MA) 모형은 현재 관측자료(y_t)가 현재와 과거의 비관측영향(α_t)의 선형결합으로 설명되는 모형이며 이때 비관측영향(α_t)는 IID를 따른다.

$$MA(q): y_t = \alpha_t - \theta\alpha_{t-1} - \cdots \theta_q\alpha_{t-q} \quad (2)$$

AR 모형과 MA 모형은 시차변수를 선형 결합하는 유사한 형태를 띈다. ARIMA 모형은 차분을 적용한 데이터를 대상으로 AR 모형과 MA 모형을 선형 결합한 모델이다. 편의를 위해 $\phi_p(B), \theta_q(B), \theta_0$를 사용했으며 뜻은 아래와 같다.

$$ARIMA(p,d,q): \phi_p(B)(1-B)^d y_t = \theta_0 + \theta_q(B)\alpha_t \quad (3)$$
$$\phi_p(B) = (1 - \phi_1 B - \cdots \phi_p B^p)$$
$$\theta_q(B) = (1 - \theta_1 B - \cdots \theta_p B^q)$$
θ_0는 deterministic trend이다.

ARIMA는 훌륭한 예측 능력을 보여준다. 하지만 본 연구에서 사용한 2년간의 위치 데이터가 가지는 월별 또는 분기별 패턴을 잡아내기에는 부족하다고 판단했다. 따라서 이러한 데이터를 더 잘 적합시키는 모델을 적용하기 위해 ARIMA에 계절성을 추가한 SARIMA 모델을 활용했다.

$$SARIMA(p,d,q)(P,D,Q): \quad (4)$$
$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)y_t = \theta_q(B)\Theta_Q(B^s)\alpha_t$$

SARIMA(p,d,q)(P,D,Q,s) 식의 s는 주기성을 판별하기 위한 단계의 수를 가정하는 것으로, 분기별(s=4), 주별(s=7), 월별(s=12) 주기를 가지는 것으로 각각 가정하여 모델을 생성한다. Φ_P, Θ_Q는 계절성을 포함한 AR, MA 연산자이다.[4]

4. 서로 다른 분석법에 따른 예측 결과

ARIMA(p,d,q) 모델이 정확도가 높으면서 동시에 과적합을 회피하기 위해서는 Box-Jenkins의 graphical method를 적용하여 AIC가 낮으면서 동시에 p,d,q의 차수가 낮은 모델을 선출해야 한다. 본 연구에서는 SARIMA(p,d,q)(P,D,Q,s) 모델의 p,d,q에 대해서도 이와 유사한 방식으로 파라미터를 설정했으며, P,D,Q는 s에 따라 AIC가 가장 낮은 모델을 선정하였다. 이때 AIC 추정식은 다음과 같다.[4]

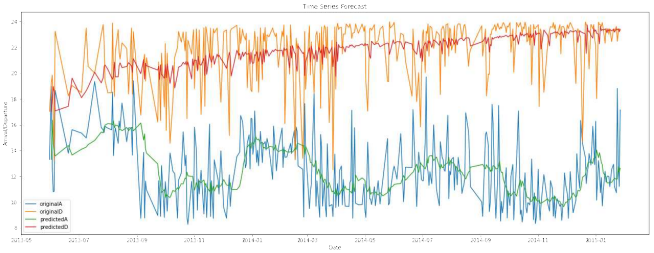
$$AIC = -2 \times \ln(likelihood) + 2 \times k \quad (5)$$

4-1. 일일 방문 시각

첫 번째 전처리 방식, 즉 일일 방문 시각을 기준으로 ARIMA와 SARIMA 모델을 적합해보았다. (그림 5)과 (그림 6)은 입장과 퇴장 시간을 시계열 모델에 적합시킨 그림이다. 일일 '퇴장' 시각을 기준으로 하였을 때 각 모델의 RMSE, MAPE, AIC 값은 다음과 같다. 이때 p-value는 0.000001로, 본 실험이 유의미함을 나타낸다.

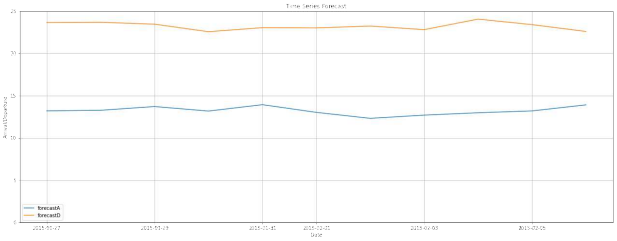
모델	RMSE	MAPE	AIC
ARIMA(1,1,3)	2.158	8.127	1561.861
SARIMA(2,1,3)(2,1,3,4)	2.014	6.904	1537.257
SARIMA(1,1,3)(2,1,3,7)	2.094	7.071	1537.387
SARIMA(2,1,3)(2,1,3,12)	2.067	7.138	1524.292

위 표에 근거하여 우리는 AIC가 유의미하게 작은 SARIMA(2,1,3)(2,1,3,12) 모델을 바탕으로 개인의 이동 패턴을 예측하였다.

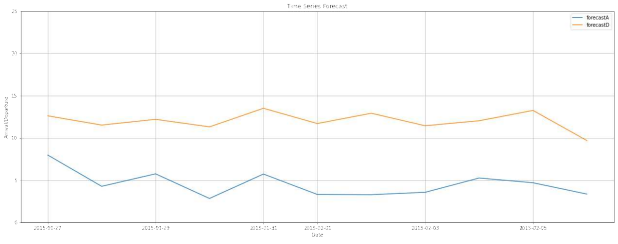


(그림 5) 학교 입/퇴장 시각 ARIMA(1,1,3)

(그림 6) 학교 입/퇴장 시각 SARIMA(2,1,3)(2,1,3,12)



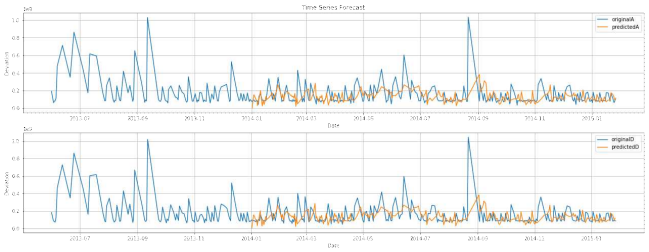
(그림 7) 학교 입/퇴장 시각을 10일 뒤까지 예측



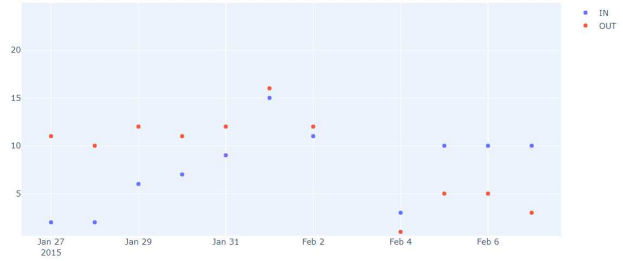
(그림 8) 집 입/퇴장 시각을 10일 뒤까지 예측

4-2. 방문 시간 간격

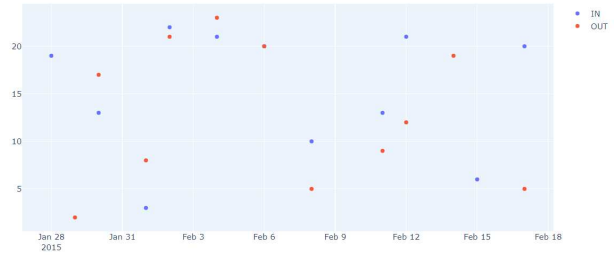
두 번째 전처리 방식, 즉 방문 시간 간격 데이터에도 4-1 방식과 유사하게 SARIMA 모델을 적용하였다. 방문 시간 간격에도 유의미한 주기성이 존재하는 것을 확인할 수 있었다.



(그림 9) 학교 입/퇴장 시각 SARIMA(3,1,3)(0,1,3,12)



(그림 10) 학교 입/퇴장 시각을 10일 뒤까지 예측



(그림 11) 집 입/퇴장 시각을 10일 뒤까지 예측

5. 연구 결과의 의미

4-1의 경우, 해당 클러스터에 방문하지 않은 날들에도 이전까지의 방문 시각 평균값이 적용되어 비교적 안정적인 예측결과를 얻을 수 있었다. 또한 집 퇴장 시각과 학교 입장 시각, 학교 퇴장 시각과 집 입장 시각이 유기적으로 연결된 것도 확인할 수 있었다. 하지만 이 방식으로는 클러스터에 방문하지 않는 날과 같은 비관측 오차를 포착하기 어렵다.

이러한 단점을 4-2의 방법으로 극복할 수 있다. 4-2의 경우, 클러스터에 입/퇴장하는 시각이 일정하지는 않지만 클러스터에 머무는 시간이 비교적 일정하고, 가까운 미래에 한해서는 적절한 정도의 비관측 오차를 포착한다는 점을 알 수 있다.

본 연구에서는 데이터를 수집하여 EM-clustering과 SARIMA 모델을 적용하고, 이를 바탕으로 분석한 개인의 이동 패턴으로 미래 이동을 예측하였다. 이 일련의 예측 프로세스는 DB에 실시간 위치 데이터가 업데이트될 때마다 자동적으로 이동 패턴을 갱신한다. 자동화된 위치 분석 시스템은 데이터 수집과 분석의 속도를 향상시킬 것이다. 본 연구에서는 학교와 집을 기준으로 사용자의 위치를 예측하였지만, 추가적인 조정이 있다면 다양한 장소와 개인에 관한 예측이 가능해질 것이다. 또한 군집간의 이동 패턴을 나타내는 Markov 모델과의 융합이 적절하게 이루어지면 이동 패턴 분석을 더 심화시킬 수 있을 것으로 예상한다.

사사

이 연구는 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행되었다. (NRF-2019R1F1A1056123)

참고문헌

[1] K. Zhao, S. Tarkoma, S. Liu and H. Vo, “Urban human mobility data mining: An overview,” 2016 IEEE International Conference on Big Data (Big Data), 2016, pp. 1911-1920, doi: 10.1109/BigData.2016.7840811.

[2] Tran Manh Thang. “ANOMALY DETECTION USING DBSCAN CLUSTERING WITH MULTIPLE PARAMETERS.” 국내석사학위논문 동국대학교, 2011. 대한민국

[3] Nielsen, Aileen. “Practical Time Series Analysis: Prediction with Statistics and Machine Learning”. Oreilly & Associates Inc. 2019

[4] 홍 정 열 · 한 은 룡 · 최 창 호 · 이 민 서 · 박 동 주. “SARIMAX 모형을 이용한 공공자전거 수요추정과 평가: 서울시의 COVID-19 영향을 중심으로.” 한국ITS학회논문지. Vol.20 No.1(2021). pp.10~21. February, 2021