도로망 정보를 반영한 DeepWalk 기반의 개인 이동데이터 엠베딩[1)]

전희준○, 신종화 ,송하윤
홍익대학교 컴퓨터공학과
yeonso16385@naver.com, troas96@naver.com, hayoon@hongik.ac.kr

DeepWalk Based Personal Geolocation Data Embedding by Reflecting Road Network

Heejun Jeon○, Jonghwa Shin, Hayoon Song
Dept. of Computer Engineering, Hongik University

요 약

도심지역에서 측정한 개인 이동 데이터(위도, 경도, 시간)는 대부분 실제 위치와 차이가 있고, 위치와 시간 외에 정보를 담지 않기 때문에 활용이 제한된다. 특히 이동 패턴 예측과 같은 정밀한 분석이 필요한 업무에서 해당 데이터의 한계가 명확하다. 이에 본 논문은 이동데이터가 도로간의 연결관계 정보를 포함하도록 변환한다. 구체적으로, 도로간의 연결관계를 그래프 형태인 도로 네트워크(Road Network)로 표현하고, 각 간선을 DeepWalk 모델을 기반으로 엠베딩 벡터로 변환한다. 맵 매칭(map matching)을 통해 이동 데이터를 적절한 간선으로 대응시키면 최종적으로 이동데이터는 벡터로 표현된다. skip gram 모델을 이용한 실험을 통해, 변환된 벡터가 도로 단위로 수행하는 업무에서 유의미한 결과를 도출할 수 있음을 보였다.

1. 서 론

어떤 대상의 이동데이터는 기본적으로 대상의 위치(위도, 경도)와 시간으로 구성되고, 속도, 가속도 등 대상의 움직임에 관한 정보를 추가적으로 포함한다. 하지만, 이동 패턴 생성과 같은 특정 업무는 대상의 움직임 뿐만 아니라 도로간의 연결관계, 방문한 장소의 유형 등 외재적인 요소에도 영향을 받기 때문에, 원시적인(raw) 이동데이터만으로는 수행에 한계가 있다. 또한, 측정한 이동데이터는 낮은 측정 빈도, 측정 에러 등의 문제로 실제 위치와 차이가 발생할 수 있고, 이는 이동데이터 활용을 더 어렵게 만든다.

따라서, 대부분의 연구는 업무에 따라 데이터를 적절히 변환한다. 예를들어, 이동 패턴 예측의 경우 다음과 같은 변환을 고려할 수있다. 먼저 전체 이동데이터에 대해 위치 클러스터링을 진행한다. 각 데이터를 자신이 속한 클러스터 번호로 변환하면 이동데이터는 클러스터 번호 시퀀스(sequence)가 된다. 이 시퀀스를 이용해, 클러스터간의 이동 패턴을 분석하고, 예측 업무를 수행할 수 있다.

이러한 변환 과정에서, 본 논문은 도로간의 연결정보를 포함시키려 한다. 대상이 이동했던 지역의 도로들을 모아 그래프를 생성하고, 이 그래프를 도로 네트워크(Road Network)라 한다. DeepWalk[1] 모델을 이용해 도로 네트워크의 각 간선(도로)을 벡터로 변환한다. 이 과정에서 도로간의 연결정보가 벡터에 포함된다. 맵 매칭(Map matching)을 이용해 각 이동 데이터를 적절한 간선에 대응시키면 최종적으로 이동 데이터는 벡터로 표현된다. skip gram[2] 모델을 이용한 실험을 통해, 엠베딩 벡터가 도로 단위로 수행하는 업무에서 유의미한 결과를 도출할 수 있음을 보였다.

2. 관련 기술

(1) DeepWalk : 어떤 그래프를 $G=(V,E)$라 표기하자. 이때 V는 노드의 집합이고, E는 간선의 집합이다. DeepWalk는 그래프의 노드를 벡터로 엠베딩하는 기술로 노드간의 이웃관계를 최대한 보존하는 것을 목표로 한다. 이때 노드 $u \in V$ 의 이웃이란 단순히 직접 연결된 노드가 아니라, u를 중심으로 이웃 셈플링 전략 S를 가지고 선택한 노드들의 집합이고, 이를 $N_S(u) \in V$ 라 표기한다. 각 노드를 $|V|$차원의 원 핫 인코딩(one hot encoding)으로 표현한 후, 해당 인코딩을 d차원 벡터로

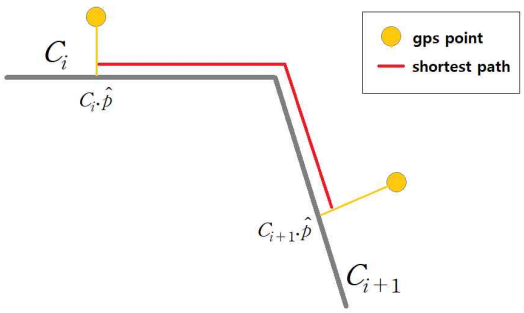
1) This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2019R1F1A1056123)

변환하는 엠베딩 행렬을 f라 하자. f의 크기는 $|V| \times d$ 이다. 노드 u의 엠베딩 벡 터 $f(u)$가 주어졌을때 u의 이웃노드 $N_s(u)$ 를 출력할 확률이 높도록 f가 학습되고, 다음 (식1)과 같은 f를 학습하는 것이 목적이다.

$$\max_{f} \sum_{u \in V} \log \Pr(N_s(u) | f(u)) \qquad \text{(식1)}$$

DeepWalk는 이웃 셈플링 전략에서 Random Walk 방식을 이용한다. 노드 시퀀스를 만드는 방식으로, 먼저 한 소스 노드에서 시작해 현재 방문한 노드와 연결된 노드들 중 확률적으로 하나를 선택해 다음 순서로 방문한다. 이 과정을 L번 반복하면 길이가 L인 노드시퀀스가 완성된다. 해당 시퀀스에서 한 노드 u를 선택하고, 그 노드의 앞 뒤로 길이 C안에 있는 노드끼리 묶어 context라 표현한다. context 안에 있는 노드들끼리 서로 이웃관계로 정의한다.

(2) 맵 매칭(Map Matching): 도로의 연결관계를 그래프 형태로 표현한 것을 도로 네트워크라 한다. 도로 네트워크에서 노드는 교착점 또는 엔드포인트를 의미하고, 간선은 도로를 의미한다. 맵 매칭은 이동 데이터의 각 포인트를 도로 네트워크의 간선에 매칭하는 것이다. 여러 알고리즘이 제안되었지만, 주로 은닉 마코프 모델(hidden markov model)을 이용한 알고리즘이 사용된다. 본 논문에선 구체적으로 [3]에서 제시한 알고리즘을 이용했다. 은닉 마코브 모델 측정된 위치 포인트를 관측값으로, 실제 대상이 위치하고 있는 간선을 은닉상태(hidden state)로 모델링하고 다음과 같이 진행한다.(그림1)



(그림1) 맵 매칭 알고리즘

먼저 한 이동데이터 포인트를 중심으로 반경이 r인 원안에 있는 도로를 해당 포인트의 후보로 채택한다. C_i는 i 시점 포인트의 후보중 하나를 의미한다. 포인트와 후보 사이 최단거리에 위치한 점($C_i.\hat{p}$, $C_{i+1}.\hat{p}$) 을 이용해 방출확률과 전이확률을 계산한다. 방출확률은 실제 위치한 도로가 주어졌을때 관측값이 측정될 확률이다. 식은 보통 정규 분포를 따른다. 후보 Ci에 대한 방출확률은 (식2)와 같다.

$$C_i.ep = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(C_i.dist)^2/2\sigma} \qquad \text{(식2)}$$

전이확률은 대상이 실제로 이전 시점(t-1)에서 특정 도로에 위치했을때, 다음 시점(t)에서 특정 도로로 이동할 확률이다. (그림1)과 같이 연속된 두 측정 포인트의 거리를 edist, 각 포인트가 후보에 대응되는 포인트($C_i.\hat{p}$, $C_{i+1}.\hat{p}$) 사이 도로를 따라 이동한 최단거리를 spdist라 하면, 전이확률 tp는 다음 (식3)과 같이 계산된다

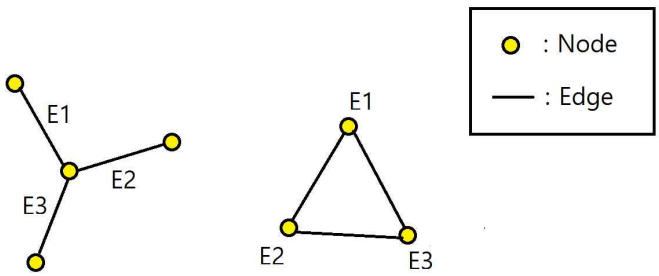
$$tp(C_i, C_{i+1}) = \frac{\min(edist, spdist)}{\max(edist, spdist)} \qquad \text{(식3)}$$

마지막으로 각 시점에서 나올 수 있는 모든 후보지에 대해 다음과같은 점수(식(4))를 최대로 만드는 후보 조합을 계산한다. 계산된 조합이 이동데이터에 추정 경로이다.

$$score = \prod_{i=1}^{n-1} C_{i+1}.ep \;\cdot\; tp(C_i, C_{i+1}) \qquad \text{(식4)}$$

3. 변환 알고리즘

본 논문은 맵 매칭과 DeepWalk을 이용해 이동 데이터를 d차원의 벡터로 변환한다. 맵 매칭은 간선을 기준으로 수행되기 때문에, 아래 그림과 같이 도로 네트워크의 간선을 새롭게 노드로 모델링 한다.



(그림2) (좌) 그래프 (우) 왼쪽의 그래프에서 간선을 노드로 모델링해 만든 그래프

DeepWalk를 이용해 각 간선마다 특정 벡터로 변환한 후, 맵 매칭을 이용해 이동 데이터를 적절한 간선에 매칭 하면 최종적으로 이동 데이터는 d차원 벡터로 변환된다.

4. 실험

실험의 구체적인 구현 사항은 다음과 같다. 먼저 DeepWalk에서 노드 시퀀스 길이 $L = 20$, context 길이 $C_D = 10$으로 사용했다. 그 다음 맵 매칭에서 후보지를 결정할때 사용하는 반지름 $r = 0.0005$, 방출확률에서 표준편차로 $\sigma = 0.0005$를 사용했다. 실험에서 사용할 데이터는 2015년 3월 1일 부터 2015년 5월 31일(약 92일) 동안 측정한 이동 데이터 중 홍익대학교 근처로 측정된 약

20000개의 데이터를 이용했다. 해당 데이터 중, 연속된 두 데이터 사이 시간 간격이 10분 이내로 측정된 좌표들을 모아 하나의 Trajectory(Tr)를 구성했고, 총 493개의 Tr을 만들어 맵 매칭을 진행했다. 맵 매칭 결과 Tr은 간선 번호 시퀀스가 되고, 각 번호를 엠베딩 벡터에 대응시키면, 최종적으로 Tr은 엠베딩 벡터 시퀀스가 된다. 실험에서 사용할 도로 네트워크는 Open Street Map [4] 에서 제공하는 홍익대학교 근처 1619개의 간선을 가진 무방향 그래프를 이용했다. 간선의 수가 적으므로 엠베딩 벡터의 차원은 $d = 32$로 고정했다.

실험에서 해결하고자 하는 업무는 이동 데이터가 어떤 도로 A에 위치했을때, 과거 혹은 미래에 위치할 가능성이 높은 다른 도로를 찾는것이다. 본 논문은 skip gram 모델을 변형해 업무를 해결하려했다. 먼저 493개의 Tr중 데이터의 약 70%에 해당하는 350개를 훈련데이터로, 나머지 30%인 143개를 테스트 데이터로 구성했다. 모델의 학습은 DeepWalk와 유사하다. Tr에서 중심 도로를 선택한 후 앞 뒤로 길이 $C_S = 20$ 이내에 있는 다른 주변도로를 묶어 context를 구성한다. 학습과정은 중심도로를 모델에 입력했을때, 주변도로가 출력될 확률을 높게 만들도록 진행된다. 따라서 실제로는 Tr 데이터에서 (중심도로, 주변도로) 쌍을 뽑아내고 이를 학습 데이터로 이용한다. 실험에선 약 80000개 정도의 쌍을 만들어 학습을 진행했다. 구체적인 모델 구조는 다음과 같다. 은닉 층은 크기가 $d \times d$행렬이고 2개의 층으로 구성된다. 마지막 출력층의 크기는 $d \times |E|$($|E|$는 도로의 개수)이고 활성화 함수 softmax를 거치면 각 도로당 확률값이 출력된다.

5.결과

학습을 마친 후, 다음과 같은 성능 평가를 진행했다. 테스트 데이터로 분류한 Tr에서 훈련 데이터와 마찬가지로 중심 도로와 길이가 C_T인 context를 구성했다. 중심도로 A를 모델에 입력했을때, 출력 확률값이 큰 N개의 도로들의 집합을 $TOP(A,N)$이라 하고, A의 context를 $A.context$라 하자. 성능평가는 가능한 모든 중심 도로 A에 대해, $TOP(A,N)$중 어떤 한 원소도 $A.context$에 포함되지 않을 경우의 수 q, 하나 이상의 원소를 포함할 경우의수 p를 기준으로 Odds를 계산한다. 구체적인 수식은 다음(식5)와 같다.

$$Odds = \frac{p}{q} \qquad \text{(식5)}$$

(식5) 값이 클수록, 모델이 훈련데이터에서 의미있는 이동 패턴을 학습해 이를 새로운 데이터에 적용할 수 있음을 의미한다. $N = 10$, $C_T = 20$ 으로 실험을 진행했고, 실험 결과 q는 114, p는 1470으로 Odds값은 약 12.8947이 나왔다.

또한, 348번 도로에대해, $TOP(348,30)$과 비슷한 모양을 가진 Tr 2개를 관찰했다. 이는 (그림4)와 같다.

(그림4)

(좌) $TOP(348,30)$: 주변도로(빨간색 선), 348번 중심도로(파란색 선)

(우) 유사한 Tr: GPS 좌표(점), 매칭된 도로(선)

왼쪽에 위치한 그림은 모두 $TOP(348,30)$를 의미한다. 왼쪽 아래에 위치한 그림에서 파란색 선이 중심도로인 348번 도로이다. 모델이 이동 패턴을 학습하므로 $TOP(348,30)$은 오른쪽 Tr들과 상당히 유사하다. 주목할 만한 점은 훈련 데이터중 348번 도로를 포함한 Tr은 오른쪽 하단(회색)의 Tr이 유일하다는 점이다. 단순히 이동 패턴만 고려한다면 회색 Tr만 학습하지만, 지리적 특성을 고려해 회색과 가까운 오른쪽 상단(초록색) Tr도 학습했다. 이러한 점에서 DeepWalk기반 엠베딩 방식은 지리적 유사성을 학습하는데 도움이 된다고 볼 수 있다. 도로망 정보를 벡터 자체에 내포해서 모델이 자연스럽게 지리적 특성을 학습할 수 있다.

참고문헌

[1] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 701-710. ACM, 2014.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In ICLR Workshop Papers, 2013a.

[3] Can Yang and Gyozo Gidofalvi, Fast map matching, an algorithm integrating hidden Markov model with precomputation. IJGIS , Vol. 32, 3 (2018), 547--570, 2018

[4] https://www.openstreetmap.org/#map=7/35.948/127.736