
2023 전기 졸업과제 착수 보고서



팀주제 : HMM을 이용한 Trajectory 예측 시스템

팀명 : 산순이와 정둘이

팀 번호 : 18

팀원 : 이성무, 하연지, 정재원

목차

1. 과제 배경 및 목표
2. 요구조건 분석 및 설계 문서
 - 2.1) 요구 조건 분석
 - 2.1.1) 데이터 전처리
 - 2.1.2) 시각화
 - 2.2) 개발 환경
 - 2.3) 전체 구상도
3. 진행 방안
 - 3.1) 데이터 전처리
 - 3.2) 모델 구성
 - 3.3) 시각화
4. 기대효과
5. 현실적 제약 사항
 - 5.1) 사용자의 개인적 특성 정보 반영의 어려움
 - 5.2) 관심지점 개수 선정의 까다로움
 - 5.3) 관심지점 선정 시 반경 설정의 까다로움
6. 개발 일정 및 역할 분담
 - 6.1) 개발 일정
 - 6.2) 역할 분담
7. 참고 문헌

HMM을 이용한 Trajectory 예측 시스템

팀명 : 산순이와 정둘이

팀 번호 : 18

팀원 : 이성무, 하연지, 정재원

1. 과제 배경 및 목표

2023년 5월 5일 세계보건기구(WHO)가 코로나 19 국제 공중보건 비상사태를 해제하고 엔데믹을 선언했다. 이로 인해 여행 수요가 늘어나 여행 산업이 활력을 띠 것으로 기대되고 있다(연합뉴스, 2023. 05. 15). 특히 맞춤형 여행에 대한 수요가 증가하고 있지만 여행 관련 정보의 과부하로 인하여 의사 결정에 어려움을 겪을 수 있다.

본 과제에서는 관심지점 선정을 위해 차량의 체류 빈도 수 데이터에 K-평균 군집화(K-Means Clustering)를 적용한다. 그 후 Haversine 공식을 통해 관심지점과의 거리를 계산하여 관심지점 방문 여부를 결정함으로써 차량의 이동 경로 데이터를 생성한다. HMM(Hidden Markov Model)을 활용하여 학습된 데이터를 기반으로 개인에게 최적화된 다음 지점을 추천하고, FMM(Fast Map Matching)으로 다음 지점으로 가는 경로의 최단 거리를 제안하는 기능을 연구할 것이다. 이는 여행자들에게 최상의 여행 경험을 제공하는 데 도움이 될 것이다.

이러한 문제점을 해결하기 위해 실제로 이동한 경로 데이터와 체류 지점을 파악, 정확한 정보를 확보한 후 이를 기반으로 소비자에게 경로(Trajectory)를 추천한다. 이를 통해 여행자들에게 더욱 정확하고 신뢰할 수 있는 맞춤형 여행 경로를 제공하고자 한다.

2. 요구조건 분석 및 설계 문서

2.1 요구 조건 분석

2.1.1 데이터 전처리

- 제주도 월별 렌터카 GPS 데이터의 이상치 및 결측치 처리
- 체류지점을 활용한 관심 지점 선정
- GPS 데이터에서 무의미한 이동 거리의 차량 정보 제외

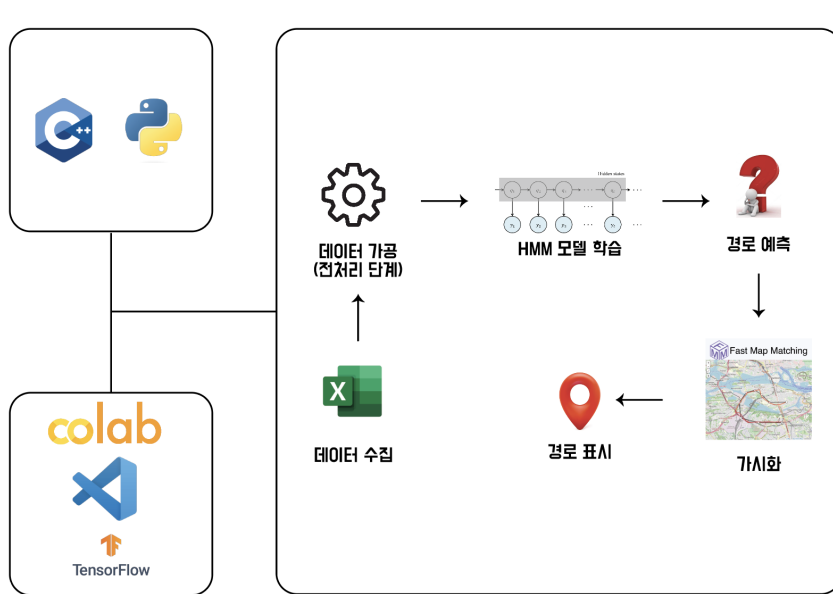
2.1.2 시각화

- 완성된 모델 시각화

2.2 개발 환경

- 개발 언어
 - Python
 - C++
- 개발 도구 및 실행 환경
 - Visual Studio Code, Iterm2 (FMM 활용)
 - TensorFlow, Colab

2.3 전체 구상도



3. 진행 방안

3.1) 데이터 전처리

- 이상치 및 결측치 처리

제주도 데이터 허브에서 “[교통, 안전] 월별 렌터카 체류 거점 위치 정보”에서 수집한 GPS 데이터의 이상치 및 결측치를 처리한다.

- 관심지점 선정

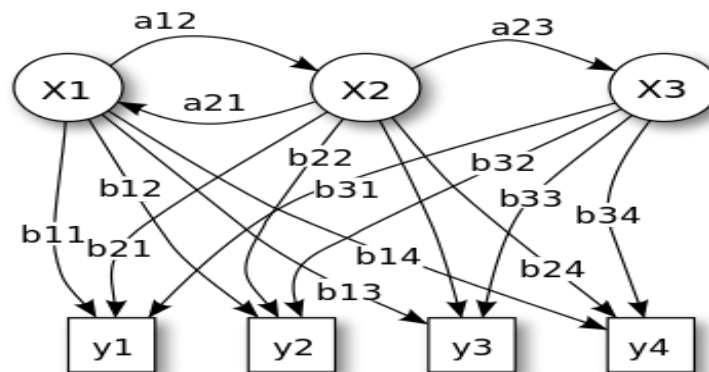
여행경로를 추천하기 위해 관심지점(POI)를 선정해야한다. 이를 위해 제주도 데이터 허브에서 “[교통, 안전] 월별 렌터카 체류 빈도”에서 수집한 데이터를 K-평균 군집화(K-Means Clustering)를 통해 관심지점의 수를 10개, 30개, 50개 등 여러 상황을 지정하여 모델에 적용하고 성능을 평가한다.

- GPS 데이터에서 경로 데이터 생성

이상치 및 결측치 처리와 관심 지점 선정이 끝나면 GPS 데이터를 관심지점의 경로로 나타낸다. 위도, 경도를 통해 거리를 계산할 수 있는 Haversine 공식을 통해 한 사용자의 위치가 관심지점과의 거리가 1km 이내일 때, 그 관심지점을 방문한 것으로 처리한다.

3.2) 모델 구성(HMM)

HMM(Hidden Markov Model)은 은 관측치 뒤에 은닉되어 있는 상태(Hidden State)를 추정하는 모델이며 시간의 흐름에 따라 변화하는 시스템의 패턴을 인식하는 작업에 유용하다고 알려져있다. 우리는 관심지점을 Hidden State(아래 그림에서 x_i)로 정의하고 시간에 따른 GPS 데이터(아래 그림에서 y_j)를 HMM에 적용한다. 또한, 은닉 상태(Hidden State) 간의 전이확률(Transition Probability, 아래 그림에서 a_{ij})은 GPS 데이터의 통계를 내 측정하고, 특정 은닉 상태(Hidden State)에서 관측치가 발생할 방출확률(Emission Probability, 아래 그림에서 b_{ij})은 거리에 따라 추정할 예정이다.

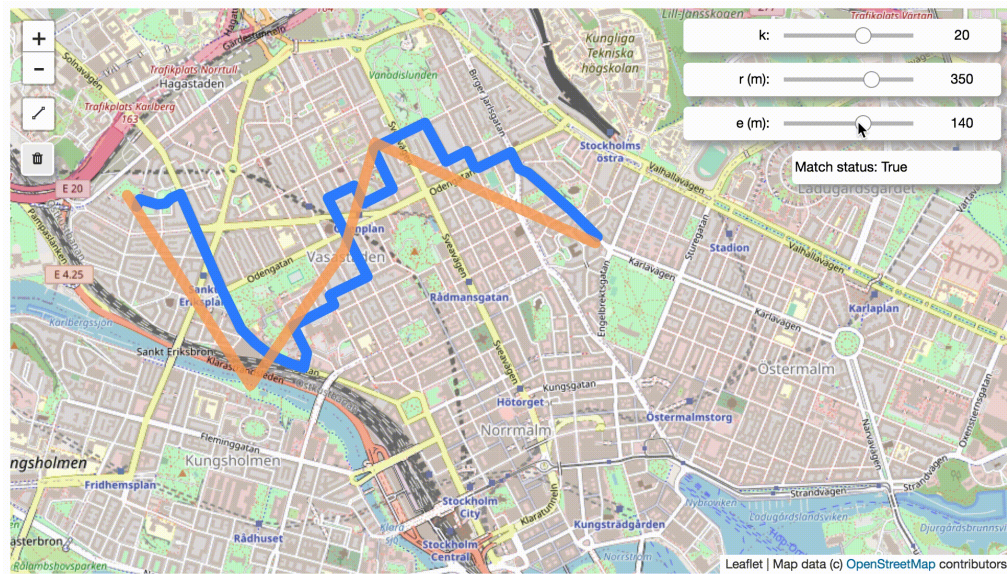


3.3) 시각화(FMM)

FMM(Fast Map Matching)은 GPS 데이터를 도로네트워크와 일치시켜 시각적으로 나타내주는 C++, Python의 오픈 소스 맵매칭 프레임워크이다.

FMM은 도로 네트워크와 관측된 GPS 데이터에 대한 사전계산을 수행하여 매칭 과정에서 빠른 접근과 검색이 가능하다. 또한 사전계산된 정보와 위치 데이터의 특성을 고려하여 가장 정확한 지점을 매칭 시켜준다. 그리하여 우리는 FMM을 통해 사용자들이 이용할 수 있는 경로를 추천해주는 시스템을 구축하려한다.

아래의 그림이 GPS 데이터를 도로 네트워크에 매칭시키는 예시를 보여준다.



4. 기대효과

사용자들의 이동 경로 데이터를 기반으로 HMM을 구축하여 다른 사용자의 이동 경로 데이터가 주어졌을 때, 해당 사용자의 다음 이동 지점과 경로를 예측하는 알고리즘으로 기존과 달리 실제 방문 데이터를 활용한 다음 경로 예측이 가능하다. 이는 여행 계획 수립을 어려워하는 사용자들에게 유용할 것으로 기대된다.

5. 현실적 제약 사항

본 과제와 데이터와 사용 알고리즘의 특성에 따라 다음과 같은 몇 가지 현실적 제약 사항이 발생한다.

5.1) 사용자의 개인적 특성 정보 반영의 어려움

- 본 알고리즘 개발 시 이용하는 데이터는 제주 데이터 허브에서 제공하는 월별 렌터카 위치정보로, 해당 데이터는 렌터카 아이디, 위도, 경도 및 해당 위치 포인트가 기록된 시간이다. 이와 같은 정보는 사용자의 이동 경로만 제공할 뿐 그 외 다른 정보를 포함하고 있지 않기 때문에 개인적 특성을 반영하지 못해 사용자의 이동 경로에만 의존하여 다음 경로를 예측할 뿐 개인화된 특성에 맞추어 경로를 예측하지 못함.
 - 월별 렌터카 위치정보 데이터에서 렌터카 사용자의 해당 위치 포인트가 기록된 시간을 오전, 오후, 저녁으로 나누어 HMM 구축에 반영해볼 수 있다.
 - 제주 데이터 허브의 다른 자료들을 열람하여 사용할 수 있는 추가 정보가 있는지 살펴본다.

5.2) 관심지점 개수 선정의 까다로움

- 김중환 *et al.*, 2014 에 따르면, 관심지점 개수가 많을 수록 주어진 데이터의 장소 집합을 잘 표현할 수 있지만 계산 시간이 증가하고, 관심지점 개수가 적을 수록 계산 시간은 감소하나 예측 정확도는 떨어질 수 있다.
 - 관심지점 개수를 달리하여 성능을 평가하여 보고, 그 결과 가장 좋은 성능을 보이는 개수를 선정한다. 이때, 관심지점 기준 반경 설정과 조합 시의 성능을 고려한다.

5.3) 관심지점 선정 시 반경 설정의 까다로움

- 관심지점을 설정한 후 매칭을 위하여 관심지점으로 설정한 위치 포인트 기준 몇 m 내에 사용자의 위치 정보가 속할 때 해당 관심지점을 방문한 것으로 볼 것인지에 따라 예측 성능의 차이가 발생할 것이다. 김종환 *et al.*, 2014 에 따르면, 반경이 작을 수록 불필요하게 세분화된 후보 장소가 생성되어 예측 성능이 떨어질 수 있다.
 - 반경의 크기를 달리하여 관심지점 개수와 조합하였을 때 성능을 평가하여 보고, 그 결과 가장 좋은 성능을 보이는 관심지점 개수를 선정한다.

6. 추진 체계 및 일정

6.1) 개발 일정

	5월			6월					7월					8월					9월			
주	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4
	데이터 전처리																					
				모델 구성																		
									모델 최적화													
									시각화													
											중간 보고서 작성											
														데이터 보충 가능 여부 확인								
														오류 확인 및 최종 테스트								
																			최종 보고서 작성 및 발표준비			

6.2) 역할 분담

이름	역할
이성무	1.데이터 전처리 2. Hidden Markov Model 구성 3. 학습 및 검증
정재원	1.데이터 전처리 2. 클러스터링을 활용해 관심 지점 군집 선정 3. 데이터 시각화
하연지	1. 데이터 수집 및 전처리 2. 학습 데이터 구축 3. 결과 분석 및 모델 최적화
공통 역할	1. 필요한 지식 학습 2. 보고서 작성

7. 참고 문헌

- 1) "코로나 터널 끝났다"...주요 여행사, 1분기 일제히 흑자전환, 2023. 05. 15, 연합뉴스, <https://www.yna.co.kr/view/AKR20230515129400003?input=1195m>
- 2) <https://www.vw-lab.com/104>
- 3) <https://github.com/cyang-kth/fmm>
- 4) 「Trajectory learning using posterior hidden Markov model state distribution」, Asmaa A.E. Osma 외 4명, Department of Information Technology, Faculty of Computers and Information, Cairo University, Egypt, 2017
- 5) 「이동 사용자의 다음 장소 예측을 위한 맵리듀스 기반의 분산 데이터 마이닝」, 김종환 외 2명, 한국정보처리학회 2014년도 춘계학술발표대회, 2014
- 6) 「Fast map matching, an algorithm integrating hidden Markov model with precomputation」, Can Yang, Gyözö Gid'ofalvi, INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2018