이동 사용자의 다음 장소 예측을 위한 맵리듀스 기반의 분산 데이터 마이닝

김종환, 이석준, 김인철
경기대학교 컴퓨터과학과
e-mail:{click7254, dltjrwns4127, kic}@kgu.ac.kr

A MapReduce-Based Distributed Data Mining Approach to Next Place Prediction for Mobile Users

Jong-Hwan Kim, Seok-Jun Lee, In-Cheol Kim
Dept of Computer Science, Kyonggi University

요 약

본 논문에서는 휴대용 기기 사용자들의 이동 궤적을 기록한 대용량의 GPS 위치 데이터 집합으로부터 각 사용자의 이동 패턴 모델을 학습해내고, 이 모델을 적용하여 각 사용자의 다음 방문 장소를 효율적으로 예측할 수 있는 맵리듀스 기반의 분산 데이터 마이닝 시스템을 소개한다. 본 시스템은 크게 사용자별 이동 패턴 모델을 학습하는 후단부와 실시간으로 다음 방문 장소를 예측하는 전단부로 구성된다. 이 중에서 후단부는 주요 장소 추출, 이동 궤적 변환, 이동 패턴 모델 학습 등 총 3개의 맵리듀스 작업 모듈들로 구성된다. 이에 반해, 본 시스템의 전단부는 이동 경로 후보군 생성, 다음 장소 예측 등 총 2개의 맵리듀스 작업 모듈들로 구성된다. 그리고 본 시스템을 구성하는 각각의 작어마다 분산 처리를 극대화할 수 있도록 맵과 리듀스 함수를 설계하였다. 끝으로, 대용량의 GeoLife 벤치마크 데이터 집합을 이용하여 본 논문에서 소개한 시스템의 예측 성능을 분석하기 위한 실험을 수행하였고, 이를 통해 본 시스템의 높은 성능을 확인할 수 있었다.

1. 서론

최근 들어 스마트폰과 같은 휴대용 기기(mobile device)의 보급률이 높아지고 모바일 센서 기술이 발전함에 따라, 대용량의 GPS 위치 데이터를 이용한 위치 기반 서비스(location-based service, LBS)에 관한 관심이 증가하고 있다. 예를 들면, 사용자의 다음 방문 장소를 미리 예측하여 주변 관광지나 맛집 등을 추천해주거나, 또는 운전자의 다음 이동 장소를 예측하여 사고 지역이나 정체 구간 등과 같은 실시간 교통정보를 미리 제공함으로써 운전자에게 도움을 줄 수도 있다. 이동 중인 휴대용 기기 사용자에게 이와 같은 유용한 위치 기반 서비스를 제공하기 위해서는, 사용자 개개인의 과거 이동 궤적(trajectory)을 기록한 대용량의 GPS 위치 데이터 집합으로부터 각 사용자의 이동 패턴 모델을 학습해내고, 이를 이용하여 각 사용자의 다음 방문 장소를 실시간으로 예측할 수 있어야 한다.

본 논문에서는 휴대용 기기 사용자들의 이동 궤적을 기록한 대용량의 GPS 위치 데이터 집합을 분석하여 각 사용자의 다음 방문 장소를 효율적으로 예측할 수 있는 맵리듀스 기반의 분산 데이터 마이닝 시스템을 소개한다. 본 시스템은 크게 사용자별 이동 패턴 모델을 학습하는 후단부(back-end)와 실시간으로 다음 방문 장소를 예측하는 전단부(front-end)로 구성된다. 이 중에서 후단부는 3개의 맵리듀스 작업 모듈들로 구성된다.

첫 번째 작업은 사용자들의 이동 궤적을 기록한 대용량의 GPS 위치 데이터 집합에 K-평균 군집화(k-means clustering)를 적용함으로써, 각 사용자들이 방문한 적이 있는 주요 장소(point of interest, POI)들을 추출하는 기능을 수행한다. 두 번째 작업에서는 GPS 위치 데이터들의 시퀀스 형태로 표현되어 있는 각 사용자들의 이동 궤적을 주요 방문 장소들의 시퀀스 형태로 변환한다. 세 번째 작업에서는 주요 방문 장소들의 시퀀스 형태로 변환된 각 사용자별 훈련 데이터 집합을 이용하여, 그 사용자의 이동 패턴을 표현할 수 있는 은닉 마코프 모델(hidden markov model, HMM)을 학습한다. 한편, 본 시스템의 전단부는 2개의 맵리듀스 작업 모듈들로 구성된다. 첫 번째 작업에서는 이동 중인 사용자의 현재까지 이동 궤적에 다음에 방문 가능한 주요 장소들을 결합하여 가능한 모든 이동 경로 후보군을 생성한다.

두 번째 작업에서는 이동 경로 후보군에 해당 사용자의 학습된 이동 패턴 모델을 적용함으로써, 사용자의 다음 방문 장소를 예측한다. 본 시스템을 구성하는 각각의 작업마다 분산 처리를 극대화할 수 있도록 맵(map)과 리듀스(reduce) 함수를 설계한다. 또한 본 논문에서는 대용량의 GeoLife 벤치마크 데이터 집합을 이용하여 본 논문에서 소개하는 분산 데이터 마이닝 시스템의 예측 성능을 분석하기 위한 실험을 수행하고, 그 결과를 소개한다.

※ 본 연구는 미래창조과학부 및 한국산업기술평가관리원의 SW컴퓨팅산업원천기술개발사업(SW)의 일환으로 수행하였음. [10044494, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발]

2. 관련 연구

<표 1> 기존 연구 비교

	W. Mathew et al. (2012)[1]	G. Sébastien et al. (2012)[2]	M. Mikołaj. (2007)[3]
주요 장소 추출	고정 사이즈의 삼각형 구역	구역 반경, 머문 시간, 방문 횟수를 고려	고정 사이즈의 직사각형 구역
사용자 궤적 변환	각각 구역에 해당하는 값들의 시퀀스로 표현	각각 장소에 해당하는 값들의 시퀀스로 표현 동일한 장소가 반복해서 나타나면 병합	각각 구역의 모서리에 해당하는 2차원 벡터들의 시퀀스로 표현
다음 장소 예측	확률 그래프 모델 (Hidden Markov Model)	확률 그래프 모델 (Markov Chains Model)	연관 규칙 분석

<표 1>은 이동 사용자의 다음 방문 장소를 예측하는 대표적인 기존 연구들을 간략히 비교하였다. 이 표에서 보듯

이 Mathew의 연구[1], Sébastien의 연구[2], Mikolaj의 연구[3] 등은 주요 장소 추출 방법, 사용자 궤적 변환 방법, 다음 장소 예측 방법 등 크게 3 가지 관점에서 서로 차이점과 공통점을 나타내고 있다.

3. 사용자 이동 패턴 모델을 이용한 다음 장소 예측

이 절에서는 사용자의 이동 궤적을 나타내는 대용량의 GPS 위치 데이터 집합을 분석하여, 사용자의 다음 방문 장소를 예측하는 방법을 설명한다.

3.1 주요 장소 추출

이 절에서는 휴대용 단말기 사용자들의 과거 이동 궤적들이 (식 1)의 p_i 와 같은 GPS 위치 데이터들의 시퀀스 형태로 저장되어 있다고 가정하고, 이 데이터 집합으로부터 각 사용자들이 방문한 적이 있는 주요 장소(point of interest, POI)들을 추출하는 방법을 설명한다.

$$p_i = (Latitude, Longitude, Date, Time),\ p_i \in P \qquad \text{(식 1)}$$

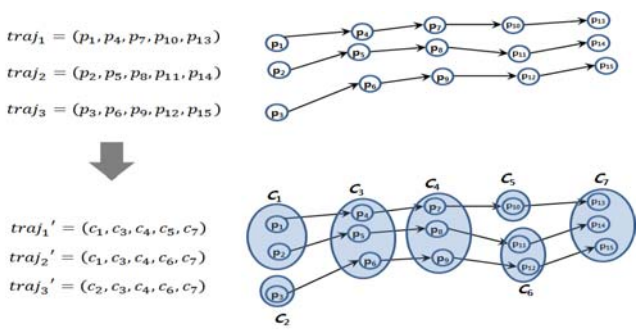
주요 장소 추출을 위해 본 논문에서는 사용자들의 이동 궤적 데이터 집합에 포함된 GPS 위치 데이터들에 K-평균 군집화(K-means clustering)를 적용하여 K개의 주요 장소를 구한다. 주요 장소 추출에 사용되는 K-평균 군집화는 구조가 간단하여 직관적으로 이해하기 쉽고, 굉장히 효율적이라는 장점이 있다. 하지만 K-평균 군집화에서는 군집의 개수를 미리 정해주어야 하고, 군집들의 초기 중심값을 어떻게 정해주느냐에 따라 군집화의 최종결과에 영향을 많이 받는 단점도 존재한다. 이 문제를 극복하기 위해서 본 연구에서는 군집의 개수, 군집 대상 데이터의 차원 및 크기에 제한이 없는 캐노피 군집화(canopy clustering)를 먼저 수행하여 K-평균 군집화의 군집 개수와 초기 중심 값을 결정하는 방법을 제안한다.

3.2 사용자 이동 궤적 변환

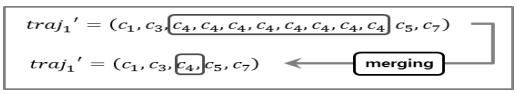
이 절에서는 GPS 위치 데이터들의 시퀀스 형태로 표현된 각 사용자의 이동 궤적을 주요 장소들의 시퀀스 형태로 변환하는 방법을 설명한다. 사용자의 이동 궤적을 구성하는 각 GPS 위치 데이터를 만나면, 이것과 주요 장소를 나타내는 K개의 군집 중심과의 거리를 각각 계산하여, 가장 가까운 주요 장소 식별자로 해당 GPS 위치 데이터를 변환해준다. (그림 1)은 GPS 위치 데이터들의 군집화를 통해 추출된 주요 장소들을 이용하여, 각 사용자의 이동 궤적들을 변환하는 방법을 예시하고 있다. 변환된 사용자 이동 궤적은 (식 2)와 같이 시간 순서에 따라 가변 길이를 가지는 주요 장소들의 시퀀스로 표현된다.

$$traj_i' = (c_1, c_2, \ldots, c_{t-1}, c_t),\ traj_i' \in Trajectory \qquad \text{(식 2)}$$

이 시퀀스에서 동일한 장소를 나타내는 식별자가 연속해서 나타나는 경우에는 그 사용자가 한 장소에 계속 머무르는 것을 의미하므로, 이들을 (그림 2)처럼 하나의 식별자로 병합하여 시퀀스 데이터를 간략화 한다.



(그림 1) 사용자 이동 궤적 변환

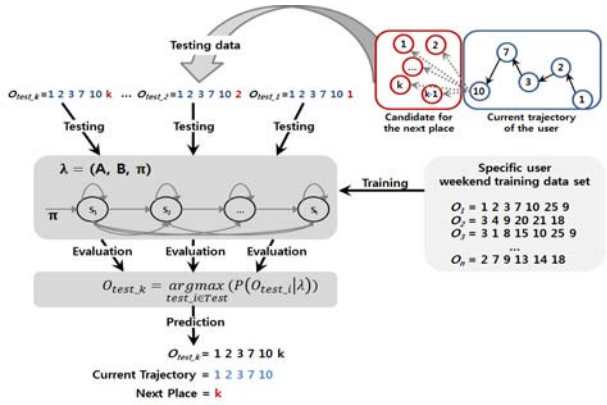


(그림 2) 주요 장소 병합

3.3 이동 패턴 모델 학습

이 절에서는 주요 장소들의 시퀀스 형태로 표현된 각 사용자의 궤적 데이터들을 이용하여 사용자별 이동 패턴(mobility pattern)을 표현할 수 있는 은닉 마코프 모델(HMM)을 학습하는 방법을 설명한다. 은닉 마코프 모델은 관측 변수와 은닉 상태 변수간의 상호 의존성을 잘 표현할 수 있는 확률 그래프 모델의 하나로, 시계열 데이터의 패턴을 표현하는데 매우 우수하다고 알려져 있다. 본 연구에서는 동일 사용자라도 일반적으로 주중과 주말에 보여주는 이동 패턴이 매우 다르다는 가정에 따라, 논문에서는 각 사용자의 이동 궤적 데이터 집합을 주중과 주말로 다시 나눈 다음, 각각의 훈련 데이터 집합(training dataset)에 대해 별도의 은닉 마코프 모델을 학습한다.

은닉 마코프 모델의 학습을 위해 결정해주어야 하는 요소로는 모델의 구조(structure), 상태(state) 수, 그리고 관측(observation) 수 등이 있다. 본 연구에서 은닉 마코프 모델의 구조는 주요 장소들의 반복으로 이루어진 이동 궤적을 표현하기에 적합한 어고딕(ergodic) 모델을 사용한다. 상태 수는 사용자들이 방문할 수 있는 실제 주요 장소들의 수를 암시하며, 이것은 적용 영역의 특성과 훈련 데이터 집합을 고려하여 적절히 선택되어야 한다. 반면에, 관측 수는 모두 군집의 수인 K로 고정된다. 은닉 마코프 모델의 학습은 Baum-Welch 학습 알고리즘을 이용한다.



(그림 3) 은닉 마코프 모델을 이용한 다음 장소 예측

3.4 이동 경로 후보군 생성

현재 이동 중인 한 사용자의 다음 방문 장소를 예측하기 위해서는 그 사용자의 현재까지 이동 궤적에다 다음 방문 가능한 주요 장소들을 붙여 가능한 모든 이동 경로 후보군을 생성한다. (그림 3)의 예처럼, 한 사용자가 1->2->3->7의 이동 궤적을 거쳐 현재 10의 장소에 도달하였다면, 이 궤적의 끝에 다음 방문 가능한 장소들의 식별자인 1~K의 붙여, 1->2->3->7->10->1 과 같은 총 K개의 이동 경로 후보군을 생성한다.

3.5 다음 방문 장소 예측

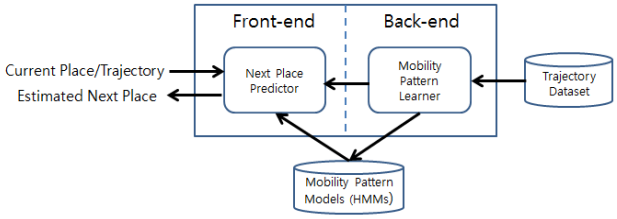
앞서 생성한 K 개의 이동 경로 후보군 각각에 해당 사용자의 이동 패턴을 나타내는 은닉 마코프 모델을 적용하여 가장 큰 우도 확률(likelihood probability)을 갖는 이동 경로를 찾아낸다. 그리고 이 경로의 맨 끝에 붙여진 장소가 바로 이 사용자 다음에 방문할 가능성이 가장 큰 장소로 예측한다.

$$O_{test\_k} = \underset{test\_i \in Test}{argmax}\left(P(O_{test\_i}|\lambda)\right) \qquad \text{(식 3)}$$

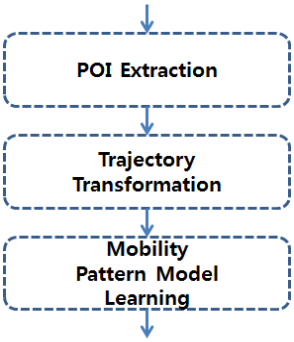
(식 3)은 학습된 은닉 마코프 모델(λ)을 기초로, 가장 큰 우도 확률을 갖는 이동 경로를 찾는 식을 나타낸다.

4. 분산 데이터 마이닝 시스템

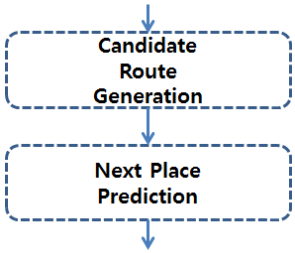
이 절에서는 이동 사용자의 다음 방문 장소를 효율적으로 예측하기 위한 맵리듀스 기반의 분산 데이터 마이닝 시스템 설계를 소개한다. 본 시스템은 (그림 4)와 같이 크게 사용자별 이동 패턴 모델을 학습하는 후단부(back-end)와 실시간으로 다음 방문 장소를 예측하는 전단부(front-end)로 구성된다. 후단부는 다시 (그림 5)와 같이 주요 장소 추출, 이동 궤적 변환, 이동 패턴 모델 학습 등의 작업들로 구성되고, 전단부는 (그림 6)과 같이 이동 경로 후보군 생성, 다음 장소 예측 등의 작업들로 구성된다.



(그림 4) 전체 시스템 구성



(그림 5) 후단부 작업 구성



(그림 6) 전단부 작업 구성

4.1 주요 장소 추출 작업

이 절에서는 사용자 이동 궤적을 기록한 대용량의 GPS 위치 데이터 집합으로부터 캐노피 군집화와 K-평균 군집화를 적용하여 주요 장소를 추출(POI Extraction)하는 작업을 맵(map)과 리듀스(reduce) 함수를 설계한다. 먼저 캐노피 군집화 의사 코드는 (그림 7)과 같다.

```
Input : Vectors
map(key, value):
        for canopy in canopyList
          if dist(canopy, value) < T2
            return
          else
            canopyList.add(new Canopy(vector,
            canopyId++, measure));
reduce(key, value):
        for canopy in canopyList
          for finalCanopy in finalCanopyList
            if dist(finalCanopy,canopy) < T2
              return
            else
              finalCanopies.add(finalCanopy));
        //removes duplicate for the same canopy center
        remove(finalCanopies)
        write(finalCanopyList)
```

(그림 7) 캐노피 군집화 의사 코드

```
Input : Vectors
map(key, value):
        for canopy in canopyList
          if kmeansCluster== null
            kmeansCluster= canopy
            nearestDistance = dist(canopy, value)
          else
            if dist(canopy, value) < nearestDistance
              nearest = canopy
              nearestDistance = dist;
reduce(key, value) :
        for kmeansCluster in kmeansClusterList
          kmeansCluster.center =
          sum(value) /
        kmeansCluster.size()
          kmeansCluster.setValue(index)
        write(KmeansClusterList)
```

(그림 8) K-평균 군집화 의사 코드

맵 함수에서는 군집의 중심들로부터 거리가 T2 이상 떨어진 GPS 위치 데이터 값(value)을 찾는다. 그리고 캐노피 군집 중심 후보들을 저장한 캐노피 컬렉션(collection)에 추가한다. 리듀스 함수에서는 맵 함수에서 얻은 캐노피 군집 중심 후보들 간의 거리를 계산하여 T2 영역 밖에 존재할 경우 최종 캐노피 컬렉션에 추가한다. 그리고 중복되는 캐노피 중심 후보를 제거한다. 그 다음 캐노피 군집화를 통해 얻은 군집의 수와 초기 중심 값들을 이용하여, K-평균 군집화를 수행한다. K-평균 군집화 의사코드는 (그림 8)과 같다. 맵 함수에서는 캐노피 군집들의 중심 값들과 GPS 위치 데이터 값과 거리를 계산하여 가장 가까운 거리에 있는 군집을 구한다. 그리고 리듀스 함수에서는 각 군집들에 포함된 모든 GPS 위치 데이터 값에 평균을 구한 후, 그 값을 군집들의 새로운 중심으로 갱신한다. 군집들의 중심 위치 변화가 임계치보다 적어질 때까지 위의 과정을 반복한다.

4.2 이동 궤적 변환 작업

이 절에서는 각 사용자의 과거 이동 궤적을 주요 장소들의 시퀀스로 변환하기 위한 맵과 리듀스 함수를 설계한다. 궤적 변환 작업은 맵 함수에서는 입력으로 들어온 각 GPS 위치 데이터로부터 위도와 경도를 추출하여, 해당 GPS 위치 데이터 대신 이 위치가 속한 군집 번호, 즉 주요 장소(POI)의 식별자로 변환한다. 리듀스 함수에서는 <사용자 번호, GPS 측정 날짜와 시간, 해당 장소 식별자> 형태의 벡터들의 시퀀스를 생성한다.

4.3 이동 패턴 모델 학습 작업

이 절에서는 주요 장소들의 시퀀스 형태로 변환된 각 사용자별 이동 궤적 데이터들로부터 사용자 고유의 이동 패턴 모델을 학습하는 작업을 맵과 리듀스 함수로 설계한다. 앞서 설명한대로 본 연구에서는 사용자별 주중, 그리고 주말동안 이동 패턴은 각각 독립적인 하나의 은닉 마코프 모델(HMM)로 표현한다고 가정한다. 따라서 각 사용자의 이동 궤적들을 담은 하나의 훈련 데이터 집합으로부터, 그 사용자의 고유한 이동 패턴을 잘 표현할 수 있는 은닉 마코프 모델을 효율적으로 학습할 수 있도록 맵 함수와 리듀스 함수를 (그림 9)와 같이 정의하였다. 맵 함수에서는 우선 모델을 초기화한 후, 이동 궤적 변환 작업을 통해 얻은 이동 궤적 시퀀스를 이용하여 모델의 초기 상태 확률 값과 상태 전이 확률, 관측 확률을 구한다. 리듀스 함수에서는 맵 함수를 통해 얻어진 은닉 마코프 모델들의 초기 상태 확률 값과 상태 전이 확률, 관측 확률의 평균을 구하여, 최종 은닉 마코프 모델을 생성한다.

```
Input:ClusterID Sequence
map(key, value):
        hmmModel = initialize(Model);
        α = forward(value, hmmModel)
        β = backward(value, hmmModel)
        for q in S
          I(q) = α1(q) * β1(q)
        for t = 1 to value.length()
          for q in S
            O(q){value_t}
              = O(q){value_t} + α_t(q) * β_t(q)
          t = t + 1
          for t = 1 to value.length()-1
            for q in S
              for r in S
                T(q){r} = T(q){r}
                          + α_t(q) * A_q(r) * B_r(x_t+1) * β_t+1(r)
          t = t + 1
```

```
Input:HMM Model
reduce(key, value):
        finalHmmModel = initial(Model)
          for hmmModel in HmmModel
            Sum(finalHmmModel.I, hmmModel.I)
            Sum(finalHmmModel.O, hmmModel.O)
            Sum(finalHmmModel.T, hmmModel.T)

        finalHmmModel.I = finalHmmModel.I / HmmModel.length
        finalHmmModel.O = finalHmmModel.O / HmmModel.length
        finalHmmModel.T = finalHmmModel.T / HmmModel.length
        write(ModelID, finalHmmModel)
```

(그림 9) 이동 패턴 모델 학습 의사 코드

4.4 이동 경로 후보군 생성 작업

이 절에서는 현재 이동 중인 한 사용자의 현재까지 이동 궤적에 다음 방문 가능한 장소들을 붙여, 가능한 모든 이동 경로 후보군을 생성하는 작업을 맵과 리듀스 함수로 설계한다. 이동 경로 후보군 생성 작업은 맵 함수에서는 입력 파라미터로 들어온 현재까지 이동 궤적의 시퀀스 끝에, 방문 가능한 장소들의 식별자를 추가하여 이동 경로 후보 시퀀스를 만든다. 리듀스 함수에서는 맵 함수를 통해 얻어진 이동 경로 후보 시퀀스 가운데 중복된 시퀀스를 제거한다.

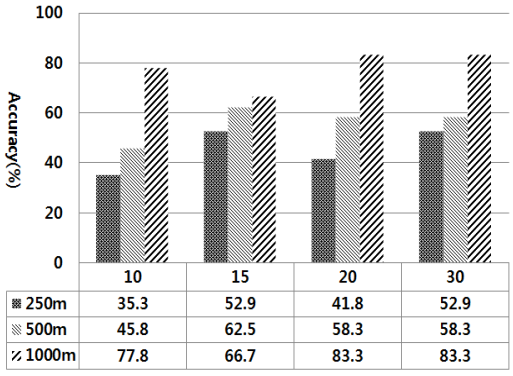
4.5 다음 장소 예측 작업

이 절에서는 앞서 설명한 방식대로 생성한 사용자의 이동 경로 후보군과 이동 패턴 모델을 이용하여, 현재 이동 중인 한 사용자의 다음 방문 장소를 예측하는 작업을 맵과 리듀스 함수로 설계한다. 다음 장소 예측 작업은 맵 함수에서는 이동 경로 후보군 생성 작업을 통해 얻은 이동 경로 후보들의 각각에 대한 발생 확률을 구한다. 리듀스 함수는 입력된 각각의 방문 가능한 장소에 해당하는 이동 경로 후보에 발생 확률을 반복 비교함으로써 발생 확률이 가장 높은 확률을 가지는 이동 경로 후보의 마지막 장소를 다음 장소로 예측한다.

5. 실험 및 평가

여러 휴대용 기기 사용자들의 이동 궤적을 기록한 대용량의 GPS 데이터 집합을 이용하여, 본 논문에서 제안한 분산 데이터 마이닝 시스템의 성능을 분석하기 위한 실험을 수행하였다. 연구용으로 공개되어 있는 대용량의 GeoLife 벤치마크 데이터 집합[6]을 사용하여 실험을 수행하였다. GeoLife 데이터 집합은 182명의 사용자들을 대상으로 2007년~2012년까지 약 5년 동안 수집한 GPS 데이터 집합으로서, 총 50,176 시간 동안 매 1~5초 혹은 매 10미터 간격으로 GPS 수신 가능한 휴대폰으로 수집한 17,621개의 이동 궤적 데이터들을 포함하고 있다.

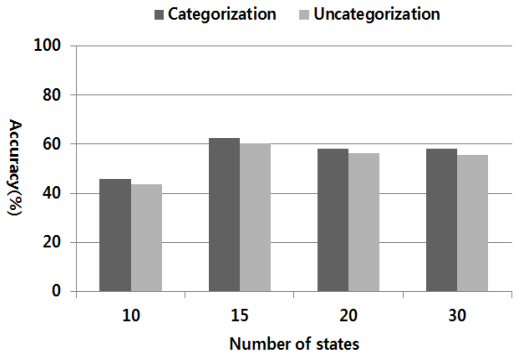
실험은 크게 두 가지로 진행하였다. 첫 번째 실험은 본 논문에서 제안하는 시스템의 다음 방문 장소 예측 성능을 분석하기 위한 목적으로 수행하였다. 이 예측 성능은 주요 장소를 추출하기 위한 군집화의 파라미터인 반경(radius)과 이동 패턴을 학습하기 위한 은닉 마코프 모델(HMM)의 상태 수(number of states)에 따라 달라질 수 있다. 따라서 본 실험에서는 이 파라미터들을 변경하면서 시스템의 예측 성능이 어떻게 변화하는지 분석하여 보았다. 이 실험을 위해 GeoLife 데이터 집합 중 10명의 사용자들을 대상으로 각각 주중과 주말의 이동 패턴을 독립적인 은닉 마코프 모델로 학습하였고, 이 모델들을 토대로 예측 성능 분석 실험을 수행하였다.



	10	15	20	30
250m	35.3	52.9	41.8	52.9
500m	45.8	62.5	58.3	58.3
1000m	77.8	66.7	83.3	83.3

(그림 10) 군집 반경과 상태 수에 따른 시스템의 예측 성능

(그림 10)은 이 실험의 결과로서, 반경을 250, 500, 1000로 하였을 때, 또 상태의 수를 10, 15, 20, 30으로 정하였을 때, 각각의 경우에 대한 시스템의 예측 정확도(accuracy)를 백분율로 보여주고 있다. 실험 결과를 보면, 본 시스템의 예측 성능은 반경이 1000 미터이고, 상태 수가 20 혹은 30개일 때, 약 83.3 %이 높은 정확도를 보였다. 하지만, 주요 장소를 추출하기 위한 반경이 작아질수록, 크게는 약 40%이상의 낮은 예측 성능 차이를 보인다. 이러한 결과의 원인은, 주요 장소를 추출할 때 반경이 너무 작아지면 불필요하게 세분화된 많은 수의 후보 장소들을 생성하게 되므로, 반경이 클 때보다 예측 확률이 낮아진 것으로 추정된다. 또한, 은닉 마코프 모델의 상태 수가 감소할수록 크게는 약17%이상의 낮은 성능 차이를 보인 것을 알 수 있다. 이는 상태 수가 너무 작아 주어진 훈련 데이터 집합을 제대로 표현하지 못하기 때문이다.



(그림 11) 그룹화에 따른 예측 성능

두 번째 실험은 사용자마다 주중과 주말의 이동 패턴이 서로 뚜렷이 다르다는 본 시스템의 가정이 어느 정도 타당성이 있는지를 분석하기 위한 목적으로 수행되었다. 이를 위해 GeoLife 데이터 집합에서 각 사용자의 이동 궤적 데이터들을 주중과 주말로 나누어 이동 패턴을 별도의 은닉 마코프 모델로 학습한 본 시스템의 경우(Categorization)와 주중과 주말을 구분하지 않고 통합된 훈련 데이터 집합으로부터 하나의 이동 패턴 모델을 학습한 경우(Uncategorization)의 시스템 예측 성능을 비교 분석하였다. (그림 11)은 실험의 결과를 나타내며, 이 실험에서는 반경은 500m로 설정하는 반면, 상태 수는 10, 15, 20, 30 등으로 변경해 보았다. 실험 결과는 그림에서 보듯이 상태 수에 무관하게 어떤 경우에나 주중과 주말의 이동 패턴을 별도의 모델로 학습한 경우(Categorization)가 그렇지 않은 경우(Uncategorization) 보다 예측 성능이 더 높게 나타났다. 이를 통해, 동일한 사용자도 주중과 주말의 이동 패턴이 서로 뚜렷이 다르다는 본 시스템의 가정이 어느 정도 설득력이 있다는 것을 확인할 수 있었다.

6. 결론

본 논문에서는 휴대용 기기 사용자의 이동 궤적을 기록한 대용량의 GPS 위치 데이터 집합으로부터 각 사용자의 이동 패턴 모델을 학습해내고, 이 모델을 적용하여 각 사용자의 다음 방문 장소를 효율적으로 예측할 수 있는 맵리듀스 기반의 분산 데이터 마이닝 시스템을 제안하였다. 그리고 대용량의 GeoLife 벤치마크 데이터 집합을 이용한 실험을 통해, 본 시스템의 높은 예측 성능을 확인하였다.

참고문헌

[1] W. Mathew, R. Rapos, and B. Martins, "Predicting Future Locations with Hidden Markov Models." Proc. of the ACM Conference on Ubiquitous Computing. ACM, pp. 911-918, 2012.

[2] G. Sébastien, M. O. Killijian, and M. N. del Prado Cortez, "Next Place Prediction Using Mobility Markov Chains." Proc. of the First Workshop on Measurement, Privacy, and Mobility. ACM, 2012.

[3] M. Mikołaj. "Mining Frequent Trajectories of Moving Objects for Location Prediction." Machine Learning and Data Mining in Pattern Recognition. Springer Berlin Heidelberg, pp. 667-680, 2007.

[4] A. Akinori, K. Maruyama, and A Sato, et al, "Pedestrian-Movement Prediction Based on Mixed Markov-Chain Model." Proc. of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2011.

[5] S. Scellato, M. Musolesi, and C Mascolo, et al, "NextPlace: a Spatio-Temporal Prediction Framework for Pervasive Systems." Proc. of the 9th International Conference on Pervasive computing. Springer-Verlag, 2011.

[6] http://research.microsoft.com/en-us/downloads/