통계적 분석방법을 이용한 Stay Point 추출 연구

Extraction method of Stay Point using a Statistical Analysis

박진관*, 오수열**

(Jin Gwan Park, Oh Soo Lyul)

요 약

최근 모바일 기기가 발전함에 따라 사용자의 위치 수집 및 분석 방법에 대한 연구가 활발히 이루어지고 있다. 사용자의 위치 분석 방법 중 궤적 데이터 마이닝은 사용자의 궤적을 바탕으로 의미 있는 정보를 추출하기 위해 사용된다. 궤적 데이터 마이닝을 수행하기 위해서는 사용자의 GPS 궤적을 분석하여 Stay Point를 추출하는 과정이 선행되어야 한다. 기존의 Stay Point 추출 알고리즘은 두 가지의 임계값이 필요하며 해당 임계값들은 사용자가 임의로 설정함으로써 알고리즘의 신뢰도가 떨어지고, 실내에서 머문 지점과 실외에서 머문지점에 대한 구분이 없기 때문에 Stay Point 위치의 모호성이 높아지는 문제점이 있다.

본 논문에서는 통계적 분석방법을 이용하여 SP를 추출하는 방법을 제안한다. 제안 알고리즘은 가우시안 확률분포를 사용하여 실내와 실외 SP를 각각 추출함으로써 SP 위치의 모호성을 개선하였다. 또한, 사용자가 임의로 설정하는 임계값이 없기 때문에 알고리즘의 신뢰도를 향상시켰다.

■ 중심어 : 궤적 데이터 마이닝 ; GPS 궤적 ; 머문지점 ; 의미기반 ; 가우시안 확률분포

Abstract

Recent researches have been conducted for a user of the position acquisition and analysis since the mobile devices was developed. Trajectory data mining of location analysis method for a user is used to extract the meaningful information based on the user's trajectory. It should be preceded by a process of extracting Stay Point. In order to carry out trajectory data mining by analyzing the user of the GPS Trajectory. The conventional Stay Point extraction algorithm is low confidence because the user to arbitrarily set the threshold values. It does not distinguish between staying indoors and outdoors. Thus, the ambiguity of the position is increased.

In this paper we proposed extraction method of Stay Point using a statistical analysis. We proposed algorithm improves position accuracy by extracting the points that are staying indoors and outdoors using Gaussian distribution. And we also improve reliability of the algorithm since that does not use arbitrarily set threshold.

■ keywords : trajectory data mining ; gps trajectory ; stay point ; semantic ; threshold

Ⅰ. 서 론

최근 들어 모바일 사용이 우리 생활 곳곳에 확산되면서 모바일 기기의 센서 데이터를 통해 사용자의 이동패턴 인식 및 위치 내역 추적이 가능하게 되었다. 이를 통해 위치기반서비스(LBS: Location-Based Service)가 발전하고, 수많은 개인화 서비스들이 제공되고 있다. 하지만, 개인을 포함한 다수에게 의미있는 서비스로 확장하기 위해서는 기존의 GPS 위성으로 검색된 대량의 이동 궤적에서 사용자의 위치를 단순히 연결하는 방식이 아닌 궤적의 특정 부분 에서 사용자의 위치가 내포하는 의미를 알 수 있어야 한다[1]. 수집된 수많은 공간 데이터들은 매우 많은 의미를 내포하고 있는데 중요한 점은, 이러한 공간 데이터를 분석함으로써 사람들을 위한 의미있는 장소를 추출해 낼 수 있다는 것이다[2]. 대부분의 사람들이 일정 공간에 무리지어 머무는 곳이 의미 있는 장소라고 할 수 있는데, 여기서 의미있는 장소란 문화적으로 중요한 장소임을 뜻한다. 예를 들어 뉴욕에 있는 자유의 여신상과 자주찾는 공공장소(쇼핑 몰, 식당, 사무실,

* 학생회원, 목포대학교 컴퓨터공학과
** 종신회원, 목포대학교 컴퓨터공학과

접수일자 : 2016년 11월 30일
수정일자 : 2016년 12월 28일
게재확정일 : 2016년 12월 20일
교신저자 : 오수열 e-mail : syoh@mokpo.c.kr

학교, 공원 등등)가 의미 있는 장소가 될 수 있다. 이러한 정보는 사용자가 주변의 위치를 이해하는 데 도움이 되고, 여행을 할 때 관광명소를 추천 받을 수 있다[3,4]. 사용자가 의미있는 장소를 이동한 기록을 Semantic Location History(SLH)라고 하며(ex. shopping malls → restaurants → cinemas)[5] SLH를 이용하여 사람들의 특성을 파악 및 패턴추출과 궤적 예측을 할 수 있다. 의미있는 장소를 Place of Interest 혹은 Point of Interest(이하 PoI)라고 하며[2] SLH를 생성하기 위해서는 PoI를 추출하는 과정이 선행되어야 한다. PoI를 추출해 내기 위해서는 우선 개개인의 이동 궤적을 분석하여 머무른 지점(Stay point or Stay location 이하 SP)을 알아내야 한다. SP는 개인이 일정 시간 동안 일정 장소에 머무른 곳을 뜻한다[3]. 사람들의 이동 궤적에서 머무른 곳이란 매우 의미하는 바가 크다. 관광 어플리케이션에서 머무른 장소는 관광 명소, 호텔, 공항 등의 장소가 될 것이고, 교통 어플리케이션에서 머무른 장소는 신호등, 로터리, 주차장 등의 장소가 될 것이다[6]. 이처럼 머무른 장소, 즉 SP는 개개인의 의미있는 장소를 추출하고, PoI는 개개인의 SP를 병합한 후 포괄적으로 의미있는 장소를 선별하는 것이다. 아울러, 이러한 정보는 경로 예측, 영업 홍보, 도로 확장, 장소 추천 등의 다양한 분야에서 유용하게 사용 될 수 있다[7,26].

기존의 SP 추출 방법[8]은 수집된 GPS 좌표에서 시간 임계값과 거리 임계값을 만족하면 두 GPS 좌표의 임의의 중간좌표를 SP로 지정한다. 머무른 지점인 SP의 정확도를 결정하는 중요한 요소는 시간 임계값과 거리 임계값이다[9]. 하지만, 기존 SP 추출 연구의 대부분이 임계값을 임의로 지정하고 있다[9-21]. 문제점은, 기존 연구에서 각 임계값을 지정한 이유가 불분명 하다는 것과 각 임계값을 지정할 때 GPS 데이터의 분포를 고려하지 않고 임의로 임계값을 지정함으로서 추출된 SP의 정확도를 장담할 수 없다는 것이다.

본 논문에서는, 기존 연구의 문제점을 해결하기 위해 통계적 분석방법을 이용한 SP 추출방법을 제안한다. 제안 방법은 실내와 실외의 SP를 가우시한 확률분포를 통해 추출하는 것이다.

실외 SP의 경우 GPS 좌표 간의 속도와 방위각을 이용하여 SP를 추출하고 실내 SP의 경우 GPS 좌표 간의 시간과 거리를 이용하여 SP를 추출한다. 실외 SP는 야외에서 사람이 머무른 지역으로서, 기존의 SP 추출 알고리즘으로 추출된 SP 구간내의 평균 속도를 계산하면 사람의 평균 보행속력인 1.3m/s[22] 이하로 측정된다. 하지만 GPS 궤적의 평균 속도가 1.3m/s 이하인 궤적이 존재하면 거의 모든 GPS 좌표가 SP 후보궤적으로 설정되기 때문에 일괄적으로 속도 임계값을 지정할 수가 없다. 그래서 본 논문에서는 Gaussian Mixture Modeling(이하 GMM)을 사용하여 속도 임계값을 지정한다. 사람은 걷거나, 자전거를 타거나, 자동차를 타는 등의 다양한 이동수단을 통해 이동하기 때문에 속도가 일정하지 않다. 그래서 GMM을 사용하여 사람이 머물 가능성이 높은 속도 구간을 추정하여 SP 후보 궤적으로 설정한다. 기존 SP 추출 방법들로 SP를 추출한 후 SP 구간내의 방위각을 계산하면 사람들의 이동 방향이 일정치 않다는 것을 알 수 있다. 즉, SP 궤적내의 GPS 좌표는 무지향성이라는 것이다. 그렇기 때문에 SP 후보 궤적의 방위각을 계산함으로서 사람이 이동하고 있는지, 머물고 있는지를 판단할 수 있다. 방위각 임계값은 SP 후보 궤적의 방위각 편차 값이 전체 GPS 궤적의 방위각 편차보다 크면 실외에서 머문 지점으로 지정된다. 실내 SP의 경우 시간과 거리 임계값을 각각 Outlier detection을 통해서 계산한다. Outlier detection은 유의수준 α를 5%를 각각의 임계값을 지정한다. GPS 포인트간 거리는 Outlier가 아니면 만족하고, 시간은 Outlier면 만족하게 된다. 즉, GPS 포인트간 시간은 크고, 거리는 짧은 구간을 찾는 과정이 실내 SP 추출인 것이다.

본 논문은 통계적 분석방법을 이용하여 SP를 추출한다. 제안 알고리즘에서 지정하는 속도 임계값과 방위각 임계값, 시간 임계값, 거리 임계값은 모두 모집단인 GPS 전체 궤적에서 추출하는 것이다. 제안 방법의 효과를 확인하기 위해 기존 알고리즘과 제안 알고리즘에 따른 SP를 각각 추출한 후, SP가 생성된 위치를 비교하여 SP의 정확도를 검증하였다.

Ⅱ. 관련 연구

1. 머문지점 추출 연구

수집된 GPS 데이터에서 머문지점을 추출하는 기존 연구는 SMoT[12]와 CB-SMoT[13], Stay Point[3] 등 다양한 방법이 존재한다. 기존 연구에서 머문지점을 추출하기 위해서는 각 알고리즘마다 임계값을 설정해줘야 한다. 하지만 임계값을 설정하는 과정에서 수집된 GPS 데이터의 분포상태를 분석하지 않고 임의로 설정되고 있는 실정이다. 표 1을 보면 각 연구마다 머문지점을 추출하기 위해 각각 다른 임계값을 설정한 것을 알 수 있다.

표 1에서 언급된 연구에서는 머문지점을 추출하기 위해 임계값을 설정하였지만 각 임계값을 설정한 이유는 언급되어 있지 않다. 각 임계값은 머문지점의 정확도를 결정하는 중요한 요소이기 때문에 불분명하게 임계값을 지정하는 기존 알고리즘에서는 추출된 결과값의 정확도를 장담할 수 없게 된다. 하지만 적절한 임계값을 찾기 위해 모든 임계값의 조합으로 실험하는 것은 불가능하기 때문에 사용자가 임의로 임계값을 지정하고 있다[9]. 그림 1은 임계값의 조합을 통해 SP를 추출한 것이다.

표 1. 각 연구에 사용한 임계값 설정

reference	Threshold (Time threshold = t Distance threshold = d)
[9]	d = 150m, t = 30s, 40s, 50s, 60s
[10]	t = Δtc
[11]	d = 20m, t = 30s
[12]	t = Δc
[13]	t = 120s
[14]	d = 200m, t = 30minutes
[15]	d = 200m, t = 20minutes
[16]	t = 30minutes
[17]	d = 5m, t = 30s
[18]	d = θd, t = θt
[19]	d = 30, t = 100s
[20]	t = 4minutes
[21]	d = distThreh, t = timeThreh_min:max

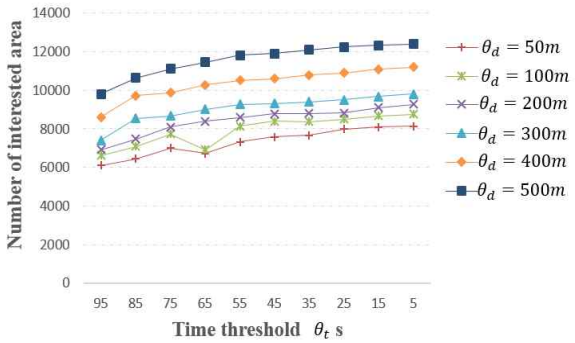

그림 1. 서로 다른 임계값을 통해 SP를 추출한 결과

그림 1은 거리 임계값과 시간 임계값에 따라 SP 추출 개수가 달라지는 것을 보여주고 있다. 하지만 임계값 낮게 지정하면 무분별하게 많은 SP가 생성된다. 즉, 실제로 머물지 않은 지점까지 SP로 추출된다. 또한, 임계값을 높게 지정하면 실제로 머문지점을 추출할 수 없게 된다[9]. 이처럼 적절한 임계값을 설정하는 것이 알고리즘의 정확도를 결정하는 중요한 요소임에도 불구하고 기존 연구[9-21]에서는 임계값을 지정한 이유가 언급되어 있지 않고 임의로 지정함으로서 알고리즘의 신뢰도 및 정확도를 떨어뜨리고 있다.

III. 본 론

본 논문은 통계적 분석방법을 이용하여 GPS 궤적에서 머문지점을 추출하는 연구이다. 본 장에서는 SP 궤적을 분석하여 나온 결과를 토대로 머문 지점을 추출하는 방법에 대해서 기술한다.

본 장에서는 GeoLife 데이터셋을 사용하여 실험을 진행하였다. GeoLife는 GeoLife 프로젝트(Microsoft Research Asia)에 의해서 182명의 user들의 GPS 데이터를 5년동안(from April 2007 to August 2012) 수집해놓은 GPS 궤적 데이터셋이다[8,9]. GeoLife 데이터셋의 경우 90%이상이 매 5초마다 10m 이내로 데이터가 수집되어져 있고, 하루치 궤적들로 구성되어져 있다[18].

1. 기존 SP 추출 알고리즘을 통해 생성된 머문지점 분석

기존 SP 추출 알고리즘을 통해 생성된 머문지점의 궤적을 분석한 결과는 표 1과 같다.

표 2. user 2의 날짜별 SP 궤적 분석

date	10m, 50s		30m, 100s		50m, 150s	
	speed	bearing	speed	bearing	speed	bearing
08.10.23	0.43	77.30	0.63	65.65	0.69	64.56
08.10.24	0.29	75.15	0.52	73.85	0.72	67.26
08.10.25	0.51	82.35	0.94	66.13	1.14	62.10
08.10.27	0.26	74.52	0.52	68.20	0.55	66.98
08.10.28	0.31	71.77	0.55	68.66	0.65	65.28
08.10.29	0.33	75.52	0.52	67.45	0.58	66.10
08.10.30	0.36	76.70	0.58	69.58	0.71	65.83
speed : average speed bearing : average bearing						

표 2는 user2의 각 날짜별 GPS 데이터를 각각 다른 임계값을 적용하여 추출된 SP를 분석한 것이다. SP는 GPS 궤적에서 머문 지점의 부분궤적의 평균 좌표를 통해 추출된다. 표 2를 보면 SP 궤적의 평균 속도는 모두 사람의 평균 보행 속력인 1.3m/s[22] 이하로 측정되는 것을 볼 수 있다. 즉, 사람이 머무는 궤적은 사람의 보행 속력보다 낮게 형성되는 궤적이 되는 것이다. 하지만 임계값이 증가할수록 SP 궤적의 평균 속도가 증가하는 것을 볼 수 있는데 궤적의 범위가 넓어질수록 궤적에서 머문지점이 아닌 부분까지 포함시키기 때문이다. SP 궤적의 특성이 표 2와 같더라도 평균 보행속력인 1.3m/s 이하인 부분궤적을 SP로 판단할 수는 없다.

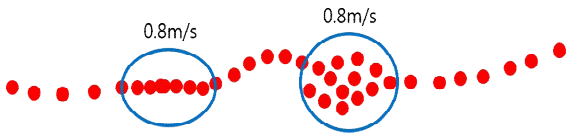

그림 2. 평균 속력이 같은 부분 궤적

그림 2에서 보는 것과 같이 부분궤적의 속력이 같더라도 이동 궤적에 따라 SP로 추출되지 않을 수도 있다. 첫번째 구간은 일정한 방향으로 천천히 걷고 있다고 판단할 수 있

고, 두번째 구간은 일정 공간 내에서 방향성을 띄지 않고 돌아다녔다고 볼 수 있다. 사람의 이동 방위각 변화량에 따라 머물렀는지 유무를 판단할 수 있는데[24], 방위각 변화량이 30도 이상으로 회전하여 일정 간격동안 유지된 경우에는 방향이 변환되었다고 할 수 있다[27]. 위와 같은 이유 때문에 첫 번째 구간은 SP로 추출되지 않는 것이다. 이러한 사실을 뒷받침 할 수 있는 자료가 표 2의 방위각 편차 평균이다. SP 궤적의 경우 방위각 편차 평균이 50도 이상이기 때문에 일정한 방향성을 띄지 않는다는 것을 알 수 있다. 즉, SP 궤적 내의 GPS 포인트들은 무지향성의 특징을 지니고 있다. 그림 2의 경우 속력만을 통해 SP를 추출한다면 첫 번째 구간과 두 번째 구간 모두 SP로 추출될 것이다. 이러한 문제점을 해결하기 위해 방위각 편차 평균을 도입하여 두 번째 구간만을 추출할 수 있도록 해야 한다. 그리고 기존 알고리즘은 모든 GPS 궤적에 동일한 임계값을 적용하여 머문 지점을 추출한다. 하지만 사람의 궤적은 매일 다르기 때문에 동일한 임계값을 통해 머문 지점을 추출하면 실제로 머문지점을 추출할 수가 없게 된다. 표 3은 GPS 전체 궤적의 평균 속도와 방위각 편차 평균을 계산한 것이다.

표 3. user 2의 날짜별 전체 궤적 분석

date	speed	bearing
2008.10.23	0.69	60.59
2008.11,03	0.83	52.25
2008.11.13	0.97	52.01
2008.12.26	2.51	26.03
2008.12.28	1.19	36.32
2009.01.20	16.05	23.71
2009.01.26	0.52	63.70
speed : average speed bearing : average bearing		

표 3을 보면 각 날짜별로 평균 속도와 방위각 편차 평균이 서로 다름을 알 수 있다. 즉, 동일한 사람이라도 매일 이동 궤적과 이동 패턴이 다르다는 것이다. 궤적의 특성이 표 3과 같기 때문에 일괄적인 임계값을 적용하면 평균 속도가 16.05m/s인 2009년 1월 20일의 경우에는, 평균 속도가 0.52m/s 2009년 1월 26일에 비해 SP가 상대적으로 적게 추출된다. 흥미로운 점은 전체 궤적의 평균속도가 높아질수록 방위각 편차 평균이 줄어든다는 것이다. 이 점은 머문지점이 무지향성이라는 사실을 뒷받침 할 수 있다. 즉, 머문 지점의 방위각 편차는 사람이 이동할 때보다 높게 측정된다는 뜻이다. 그림 3은 user2의 2008년 10월 23일 ~ 2009년 3월 22일까지 수집된 궤적의 평균 속력과 방위각 편차 평균의 상관관계를 나타낸 그래프다.

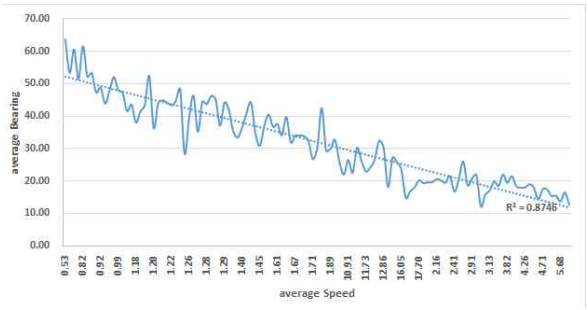

그림 3. 평균 속력과 방위각 편차 평균의 상관관계

그림 3은 GPS 궤적의 평균 속력에 따른 방위각 편차 평균의 상관관계를 도식화 한 것이다. x축은 궤적의 평균 속력, y축은 궤적의 방위각 편차 평균이다. 평균 속력과 방위각 편차 평균의 상관관계는 -0.5158로 음의 상관관계를 가진다. 즉, 머문지점의 방위각 편차 평균이 이동할 때 보다 상대적으로 높게 측정되는 것을 알 수 있다. 또한 결정계수인 R^2의 값이 0.8746로서 매우 높은 상관관계를 나타냄을 알 수 있다.

이처럼, SP 궤적의 특성은 첫째, 사람의 평균 보행 속도보다 낮은 구간이며 둘째, 무지향성을 띈다는 것이다. 이 두 가지의 특성을 고려하면 기존에 사용되는 임계값을 사용하지 않고 SP를 추출할 수 있다.

2. 통계적 분석방법을 이용한 SP 추출

본 논문에서는 기존의 연구에서 사용되는 임계값들을 사용하지 않고 GPS 데이터 분포를 분석하여 SP를 추출한다. 표 2에서 분석한 결과를 토대로 통계적 분석방법 통해 SP를 추출하면 기존 알고리즘에서 추출되는 SP와 동일한 장소의 SP를 추출할 수 있다.

기본적으로 SP는 2가지의 경우에 생성된다. 첫 번째는 객체가 시간 임계값을 초과하여 머무를 때 생성된다. 대표적으로는, 사람이 빌딩안으로 들어가서 GPS 신호가 끊긴 후 일정 시간이 지난 후 사람이 다시 빌딩 밖으로 나올 때 생기는 시간 간격으로 인해서 SP가 생성되는 것이다. 두 번째는 객체가 일정 공간을 일정 시간동안 머무를 때 생성된다. 사람들이 목적을 가지고 특정 야외 공간에서 머무를 때 SP가 생성되는 것이다. 기존의 GPS 좌표와 비교해 볼 때, SP는 각 user들이 방문한 쇼핑 몰, 레스토랑, 공원, 도서관등과 같은 개개인의 의미있는 장소를 추출할 수 있다[23]. 하지만 기존 연구에서는 실내와 실외에서 각각 생성되는 SP를 구분하지 않고 추출하기 때문에 그림 3처럼 어느 지점에서 머문것인지 판단할 수 없게 된다.

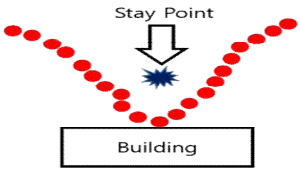

그림 4. 실내에서 머물렀을 때 생성되는 Stay Point

그림 4를 보면 유저가 실내로 들어가 일정 시간을 보낸 후 다시 입구로 나와서 SP가 생성된 경우이다. SP는 임계값을 만족하는 구간의 평균 좌표이기 때문에 그림 4처럼 실제로 머문 지점이 아닌 궤적의 중간 좌표에 생성된다. 그렇기 때문에 생성된 SP가 실외에서 머문것인지, 실내에서 머문것인지 판단할 수 없게 된다. 즉, 실내와 실외의 머문지점을 구분하지 않으면 SP의 모호성이 높아지게 된다. 이와 같은 이유 때문에 실내와 실외를 구분해서 머문지점을 추출하는 알고리즘이 필요하다.

가. 실외에서 머문 지점 추출

표 1에서 분석한 바에 따르면 사람이 머문 지점은 평균 보행 속도인 1.3m/s 보다 낮은 구간이며 무지향성인 특성을 지니고 있다. 이러한 SP의 특성을 이용하면 실외에서 머무른 지점을 추출할 수 있다. 제안하는 실외 SP 추출 알고리즘은 다음과 같다.

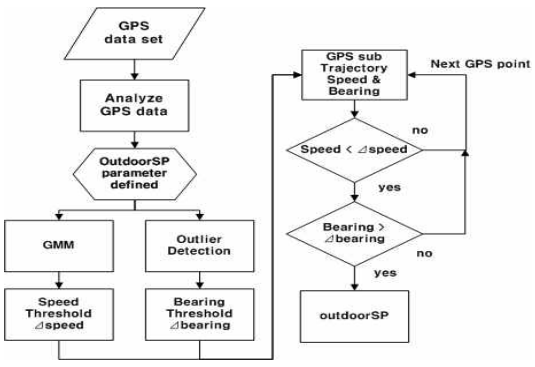

그림 5. 실외 SP 추출 알고리즘

그림 5는 수집된 GPS 데이터에서 실외에서 머무른 지점을 추출하는 과정을 보여준다. GPS 데이터를 분석하여 속도 임계값(ΔSpeed)과 방위각 편차 임계값(ΔBearing)을 지정한다. 각 임계값을 지정하는 방법은 다음과 같다.

$$\Delta Speed = \sum_{0}^{end-1} \left[\frac{p_{i+1}.dist + p_i.dist}{p_{i+1}.time - p_i.time} \right] / end - 1 \quad (1)$$

$$Bearing = acos \frac{\sin(p_{i+1}.lat) - \sin(p_i.lat) \times \cos(d)}{\sin(d) \times \cos(p_i.lat)} \quad (2)$$

$$\Delta Bearing = \sum_{0}^{end-1} [\,|Bearing_{i+1} - Bearing_i|\,] / end - 1 \quad (3)$$

식 1은 수집된 GPS 데이터의 평균 속도를 계산하는 것이고, 식 2는 방위각 편차 평균을 구하는 식이다. 식 1은 수집된 GPS 데이터의 평균 속도를 계산하는 것이고, 식 2와 3은 방위각 편차 평균을 구하는 식이다. 식 1은 GPS 부분 궤적에서 속도평균을 계산하는 식이며, 식 2는 GPS 포인트간 방위각을 계산하는 공식이다. 그리고 식 3은 방위각 편차 평균을 계산하는 식을 나타낸다. 각 임계값을 통해 실외 SP를 추출할 수 있지만 표 3에서 분석한 결과처럼 수집된 GPS 궤적의 평균 속도가 사람의 평균 보행속도인 1.3m/s 보다 낮게 형성될 수 있기 때문에 식 1의 ΔSpeed를 그대로 사용할 수는 없다. 표 3을 보면 평균 속도가 1.3m/s 보다 낮은 날짜가 있기 때문에 1.3m/s를 적용하면 거의 대부분의 GPS 좌표가 SP로 추출될 것이다. 그렇기 때문에 전체 궤적 속도를 분석하여 적절한 임계값을 지정해야 한다. 그 후, 순차적으로 GPS 궤적의 속도를 계산하여 ΔSpeed 미만인 구간을 SP 후보 궤적으로 추출한다. 추출된 SP 후보 궤적의 방위각 편차 평균이 ΔBearing을 초과하면 해당 궤적의 평균 좌표를 SP로 지정한다.

(1) 속도 임계값 설정

속도 임계값 ΔSpeed를 구하기 위해서 GPS 전체 궤적의 속도 분포를 분석한다. GeoLife는 각 유저별, 각 날짜별로 데이터가 나누어져 있기 때문에 각 날짜별로 궤적의 속도 분포를 관찰할 수 있다. 그림 6은 유저별, 날짜별로 각 GPS 포인트간 속도를 히스토그램으로 나타낸 것이다. 육안으로 보기에도 분포가 단일인 unimodal 형태가 있는 반면에 분포가 복수인 multimodal 형태의 데이터도 존재한다. 그렇기 때문에 속도임계값을 일괄적인 값으로 지정하면 제대로 된 SP를 추출할 수가 없게 된다.

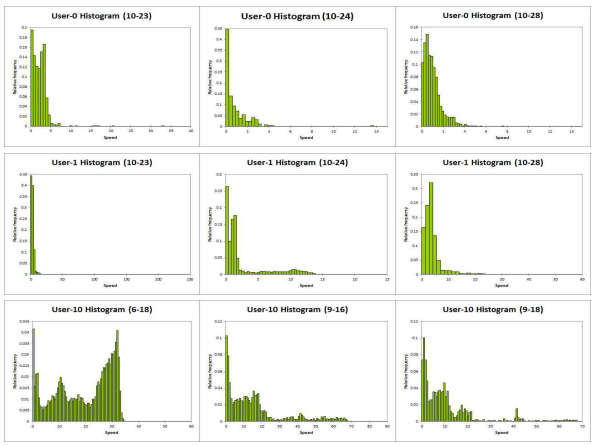

그림 6. 유저에 따른 각 날짜별 속도 분포

각각의 히스토그램을 보면 0번 유저의 10월 28일처럼 속도의 분포가 단일인 데이터가 있는 반면에, 10번 유저의 속도 데이터처럼 복수개의 분포를 띄는 데이터도 존재함을 알 수 있다.

표 4. 10번 유저의 이동 수단

Start Time	End Time	Transportation Mode
2008-06-18 0:34	2008-06-18 0:39	walk
2008-06-18 0:40	2008-06-18 0:53	taxi
2008-06-18 1:10	2008-06-18 3:23	train
2008-06-18 3:44	2008-06-18 3:52	walk
2008-06-18 3:58	2008-06-18 4:12	bus
2008-06-18 4:46	2008-06-18 4:54	subway
2008-06-18 5:13	2008-06-18 5:15	bus
2008-06-18 12:10	2008-06-18 12:20	walk
2008-06-18 12:33	2008-06-18 13:23	bus
2008-06-18 13:28	2008-06-18 13:32	bus
2008-06-18 13:32	2008-06-18 13:42	walk

표 4는 GeoLIfe 데이터셋에서 제공하는 유저들의 이동 수단을 기록해놓은 자료 중 일부이다. 표 4에서 보는 것과 같이 10번 유저의 6월 18일 데이터의 경우에는 걷고, 택시 타고, 지하철 타고, 버스를 타는 등의 다양한 행동 패턴을 보였기 때문에 위와 같은 속도 분포가 나온 것이다. 이렇듯 모든 유저별, 날짜별로 행동패턴이 다르기 때문에 일괄적인 임계값을 적용하여 SP를 추출하면 신뢰성이 매우 떨어지게 된다. 이처럼 다양한 분포를 띄는 데이터들을 단일 가우시안을 통해 표현할 수는 없다. 그래서 GMM을 이용하여 각각의 데이터의 다양한 분포형태를 분석하여 가장 알맞은 임계값을 지정할 수 있다. 위에서 언급한 각각의 임계값을 적정하게 근사해 낼 수 있어야 하는데, 데이터들의 분포 특성을 파악 하는 것이 매우 중요하다. 데이터의 분포 특성을 알기 위해서는 적절한 확률밀도함수를 이용하여 데이터 분포에 대한 모델을 만드는 것을 확률 모델이라 한다. 가장 대표적으로 사용되는 확률 모델로 가우시안 확률모델이 있다. 가우시안 확률분포는 가장 널리 사용되는 분포이기는 하지만, 데이터들이 평균을 중심으로 하나의 그룹으로 뭉쳐있는 unimodal 형태만을 표현 가능한 단점이 있다. 그림 6을 보면, 육안으로 보기에도 분포가 하나인 그래프가 있는 반면에 2개 이상의 분포가 존재하는 그래프도 있다. 그렇기 때문에 가우시안 분포만으로는 데이터의 분포 특성을 정확하게 반영할 수가 없다. 그래서 복수개의 가우시안 분포를 분석할 수 있는 가우시안 혼합 모델(GMM)을 사용한다[25].

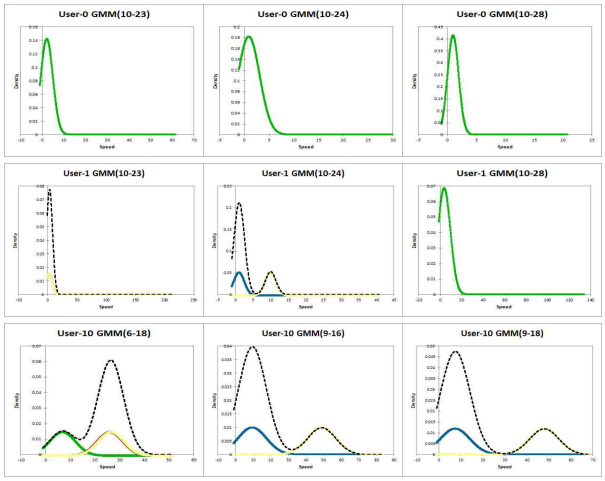

그림 7. 각 유저의 날짜별 GMM을 이용한 속도 분포

그림 7을 보면 분포가 단일인 unimodal 분포를 보이는 데이터가 있는 반면에, 복수개의 분포를 보이는 multimodal 분포를 띄는 데이터도 있다. unimodal 분포의 경우에는 평균값이나 중앙값을 임계값으로 정해도 무방하지만, multimodal 형태의 경우에는 평균값이나 중앙값으로는 데이터의 특성을 설명하지 못하므로 임계값으로 지정할 수가 없다.

표 5. 10번 유저 6월 18일 가우시안 평균값들

Gaussian	1	2	3	4	5
Mean (Speed)	6.964	26.164	26.363	26.561	26.588

표 5는 10번 유저의 6월 18일 속도 데이터를 GMM을 통해 생성된 각 가우시안 평균값들이다. 속도 임계값은 가장 낮은 속도를 유지하고 있는 구간을 지정해야 하므로 첫 번째 가우시안의 평균을 속도 임계값으로 지정하고, 속도 임계값보다 낮은 구간을 SP 후보 구간으로 지정한다.

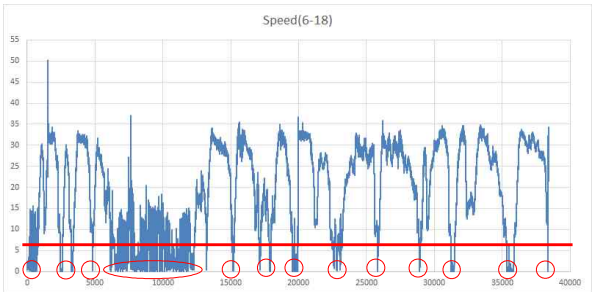

그림 8. 속도 임계값보다 낮은 구간 탐색

그림 8은 10번 유저의 6월 18일 GPS 데이터를 시간의 순서

대로 속도를 추출한 그래프다. x축은 시간이고 y축은 속도를 나타낸다. GMM을 통해 생성된 첫 번째 가우시안의 평균값인 6.964는 위 그래프의 붉은색 가로선이다. 그래프에서 보는 것과 같이 임계값보다 낮게 형성된 속도 구간이 유저가 야외에서 머물렀을 가능성이 있는 후보 구간이 된다.

(2) 방위각 임계값 설정

위에서 알아본 것과 같이, 사람이 머문지점은 속도가 상대적으로 낮은 구간과 그 구간에서 무지향성을 띄는 움직임을 보이는 특징을 보였다. 무지향성을 추론하기 위해서는 GPS 궤적을 수식(2)를 이용하여 유저가 이동한 방위각을 알 수 있다. 그리고 생성된 방위각의 편차를 분석하여 임계값을 지정할 수 있다.

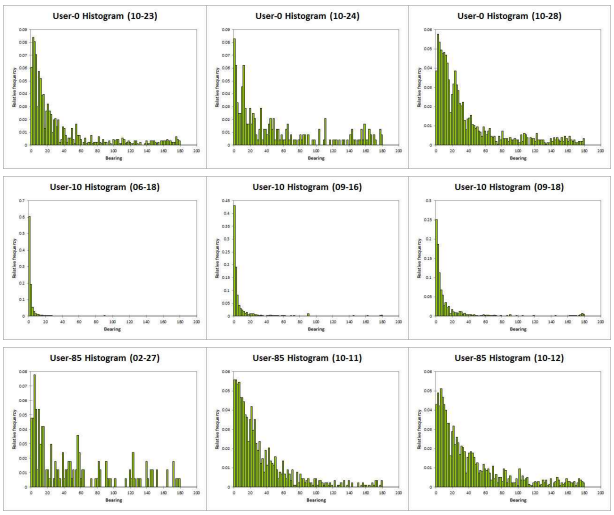

그림 9. 각 유저의 날짜별 방위각 편차 분포

표 6. 각 유저들의 이동수단

user	date	Transportation Mode
10	6-18	walk, taxi, train, bus, subway
	9-16	walk, taxi, train, bus, subway
	9-18	walk, subway, bus, train
85	2-27	walk
	10-11	walk
	10-12	walk

표 6에서 보는 것과 같이 10번 유저의 경우 다양한 이동 수단으로 이동하였기 때문에 방위각 편차가 일정함을 볼 수 있다. 대부분의 운송수단이 목적지를 대상으로 직선으로 빠르게 이동하기 때문이다. 85번 유저의 경우에는 모두 걸어서 이동했기 때문에 방위각 편차 분포가 10번 유저에 비해서 매우 불규칙함을 알 수 있다. 표 3에서 보는 것과 같이 속도가 상대적으로 느려지면 방위각 편차가 커진다.

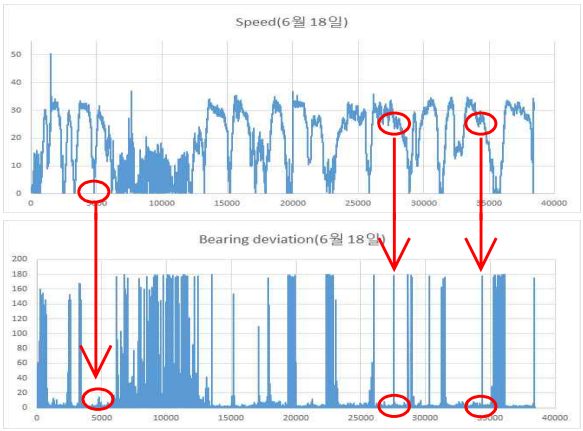

그림 10. 시간의 순서에 따른 속도와 방위각 편차

그림 10은 10번 유저의 6월 18일에 수집된 GPS 궤적에서 속도와 방위각 편차를 시간의 순서대로 나열한 그래프다. x 축은 시간이고, y 축은 각각 속도와 방위각 편차를 나타낸다. 왼쪽부터 표시해놓은 부분을 보면, 속도가 0으로 떨어지는 구간이 있다. 만약 10번 유저가 그 구간에서 머물렀다면 방위각 편차가 커야 함에도 불구하고 매우 낮음을 확인 할 수 있다. 그리고 나머지 표시해놓은 구간을 보면, 방위각 편차가 매우 높지만 속도 또한 높음을 알 수 있다. 즉, 속도가 낮은 구간에서 머물렀다고 할 수 없고, 방위각 편차가 높은 구간에서 머물렀다고 판단할 수 없는 것이다. 버스와 같은 이동 수단을 타고 이동할 때보다 걷기처럼 상대적으로 느리게 이동할수록 방위각 편차의 폭이 커짐을 알 수 있다. 그래서 적절한 각도편차 임계값을 지정하면 야외에서 머문지점을 추출할 수 있는 것이다.

본 논문에서 제안하는 실외 SP추출 알고리즘의 경우, 전체 궤적에서 속도 임계값보다 낮은 부분 궤적을 지정한 후에, 그 구간의 방위각 편차의 평균을 계산하여 방위각 임계값보다 높으면 야외에서 머물렀다고 인정되는 것이다. 그래서 속도 임계값은 GMM을 사용하여 첫 번째 가우시안의 평균값을 임계값으로 지정하고, 방위각 임계값은 Z-Score를 사용하여 유의수준 α를 5%로 지정한다. 즉, 95%의 신뢰구간을 지정하고 5%를 Outlier로 지정하는 것이다. 속도 임계값을 만족하는 부분궤적의 방위각 편차 평균의 값이 유의수준 α를 초과하면 유저는 해당 부분궤적에서 머물렀다고 판단할 수 있다.

그림 11은 유저 85의 10월 12일 날짜의 속도와 방위각 편차를 나타낸다. 표 6을 보면 유저 85의 10월 12일은 오로지 걸어서 이동하였다. 그래서 전 구간에 걸쳐서 방위각 편차가 유저 10의 6월 18일 보다 상대적으로 매우 불규칙함을 알 수 있다. 이렇듯 일괄적인 임계값을 적용하면 85번 유저의 10월 12일은 전 구간에 걸쳐 SP가 추출될 가능성이 매우 높다. 그래서 적

절한 임계값을 추출하기 위해 Z-Score을 사용한다.

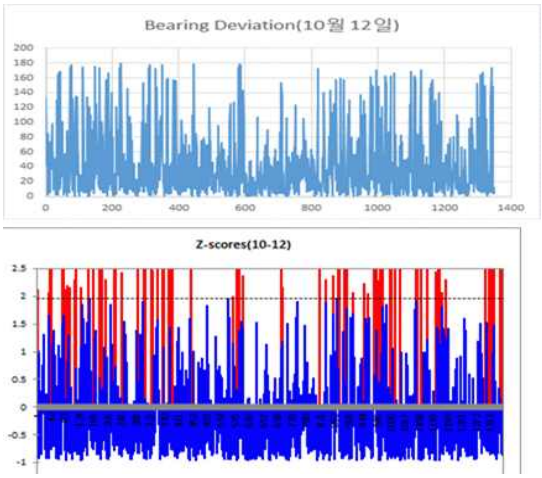

그림 11. 방위각 편차의 Outlier 검출

본 논문에서는 Outlier detection을 수행하기 위해 표준정규분포를 사용하고 유의수준 α는 0.05로 설정한 후, 유의수준을 초과하면 Outlier로 판단한다. 그림 11은 Z-Score를 통해 Outlier를 검출한 것이다. 방위각 편차가 Z-score 값 1.96(Standard Deviation * 1.96)를 초과하면 Outlier로 판단한다. Outlier로 검출된 방위각 편차 구간은 유저가 야외에서 머물렀을 가능성이 있는 후보 구간으로 지정된다.

(3) 속도 임계값과 방위각 임계값을 통한 OutdoorSP 추출

속도 임계값은 GMM을 통해 첫 번째 가우시안의 평균값으로 지정하고, 방위각 임계값은 Z-Score를 통해 유의수준 0.05 값을 임계값으로 지정한다.

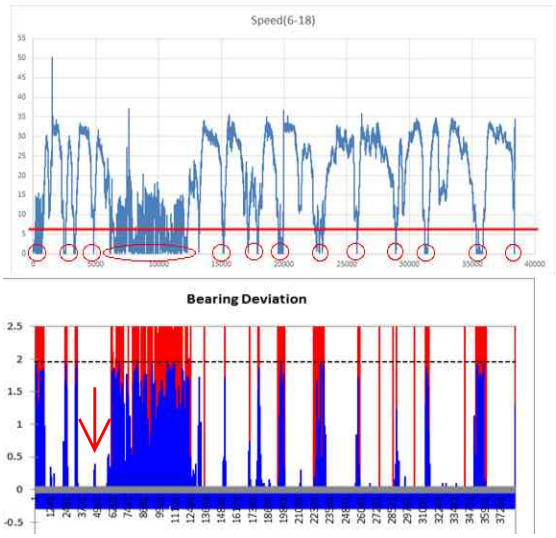

그림 12. 속도와 방위각을 이용한 Outdoor SP 추출

그림 12를 보면 속도 임계값을 만족하는 부분궤적을 먼저 선정하고, 그 구간의 방위각 편차 평균이 방위각 임계값을 초과하면 해당 부분궤적은 OutdoorSP로 지정되는 것이다. 그림 12를 보면 속도 그래프에서 총 13개의 후보 구간이 존재한다. 지정된 후보 구간에서 방위각 편차 평균이 방위각 임계값을 초과 하는 구간은 12개 구간이다. 그림 12에서 화살표로 지정된 구간은 속도 임계값은 만족하지만 방위각 임계값은 만족하지 않으므로 제외된 구간이다. 해당 구간은 상대적으로 느리게 이동했지만 일정한 방향으로 이동을 했기 때문에 머물렀다고 판단할 수 없는 것이다. 위와 같은 방법으로 유저의 OutdoorSP를 추정할 수 있다.

나. 실내에서 머문 지점 추출

사람들이 실내에서 머문 지점을 추출하기 위해서는 수집된 GPS 데이터에서 시간 임계값(ΔTime), 거리 임계값(ΔDistance)를 추출해야 한다. 그 후, GPS 포인트간 시간이 ΔTime을 초과하고 거리가 ΔDistance 미만이면 실내에서 머무른 지점으로 지정한다. 왜냐하면, 유저가 실내로 들어가서 GPS 신호가 끊긴 후, 다시 실외로 나와서 GPS 신호가 연결되는 상황이 실내에서 머무른 곳이 되기 때문이다[23]. 일정 간격으로 수집되는 시간 외의 시간은 유저가 실내로 들어갔거나, GPS 신호가 잡히지 않는 음영지역을 지나가거나, 혹은 GPS 신호자체가 lost 된 경우이다. 유저가 건물 내부로 들어갔다가 다시 입구로 나오면 GPS 포인트 간격의 거리는 평균 이동거리의 범위 내에 있을 것이다. 시간 임계값을 만족하더라도 GPS 포인트간 거리가 일정 거리를 초과하면 같은 장소에서 머물렀다고 볼 수 없다.

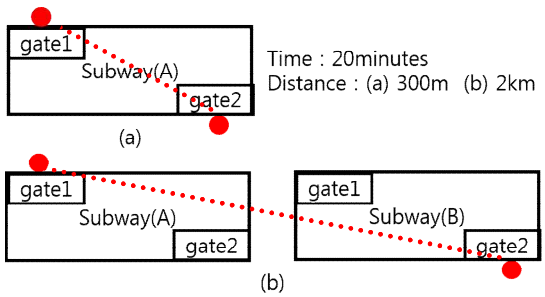

그림 13. 실내 SP 판단

그림 13의 상황은 지하철 내부에서 GPS 신호가 20분 정도 끊긴 후 다시 연결되는 상황을 가정한 것이다. 그림 13의 (a)의 상황은 Subway(A) 지점의 gate1에 들어가서 20분을 머문 후 300m 거리의 gate2로 나온 것이다. (a)의 상황은 GPS 좌표가 같은 건물에서 수집됐기 때문에 실내에서 머문지점으로 지정될 수 있다. (b)의 상황은 Subway(A) 지점의 gate1 으로 들

어가서 20분 을 머문 후 2km 거리의 Subway(B) 지점의 gate2로 나온 것이다. (b) 경우 시간 임계값은 만족하지만 GPS 좌표가 다른 건물에서 수집됐기 때문에 실내에서 머문지점으로 지정될 수 없다. 위와 같은 이유 때문에 적절한 거리 임계값을 지정하지 않으면 (b)와 같은 상황도 SP로 지정될 수 있다. 수집된 GPS 데이터에서 시간 임계값(ΔTime)과 거리 임계값(ΔDistance)을 계산해야 한다. 대부분의 경우 1~5초 사이 간격으로 GPS가 수집되지만 GPS를 수집하는 기기의 설정 상태에 따라서 신호를 수집하는 시간 간격이 각각 다르기 때문에[18] 일괄적인 임계값을 지정하면 실내에서 머무른 지점이 아닌 곳이 SP로 지정될 수 있다. 그래서 각 데이터마다 적절한 시간 임계값과 거리 임계값을 지정해야만 신뢰성 높은 SP가 추출될 수 있다. 하지만 시간과 거리 분포를 확인 후 적절한 임계값을 매번 지정할 수는 없기 때문에, Z-Score를 통해 ΔTime과 거리 임계값 ΔDistance를 지정한다.

본 논문에서 제안하는 실내 SP 추출과정은 그림 14와 같다.

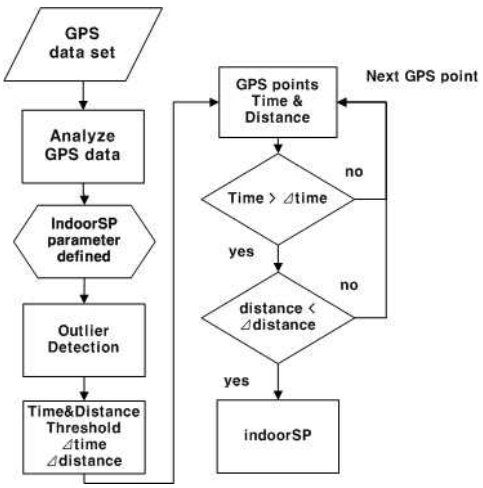

그림 14. 실내 SP 추출 과정

수집된 GPS 데이터에서 시간 임계값(ΔTime)과 거리 임계값(ΔDistance)을 계산해야 한다. GPS 데이터를 모집단이라고 하면, 거리 임계값 ΔDistance와 시간 임계값 ΔTime은 모집단의 시간과 거리에 대한 평균과 표준편차를 통해 표준정규분포를 생성한다. 그 후, 유의수준 α를 5%로 지정한 후 거리 임계값 ΔDistance는 신뢰구간인 0~95% 구간에 값이 존재하면 만족하고, 시간 임계값 ΔTime은 유의수준 5% 구간에 값이 존재하면 만족하게 된다. 지정된 거리 임계값과 시간 임계값을 통해 실내 SP를 추출하게 되는 것이다.

그림 15는 10번 유저의 6월 18일과 9월 16일의 GPS 포인트간 시간과 거리를 시간의 순서로 나타낸 그래프다. 실내 SP의 경우 GPS 포인트간 시간차가 커야하고 거리의 폭은 좁아야 한다. 그래서 위에 원으로 표시한 지점이 실내 SP 후보 지점이 된다. 10번 유저의 경우 다양한 이동수단을 통해 이동하면서 2곳의 실내에서 머물렀을 가능성이 있다.

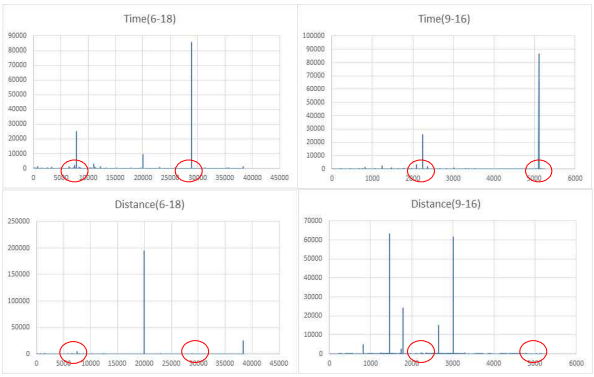

그림 15. 10번 유저의 시간의 순서에 따른 GPS 포인트간 시간과 거리

그림 16은 85번 유저의 10월 16일과 10월 20일의 GPS 포인트간 시간과 거리를 시간의 순서로 나타낸 그래프다.

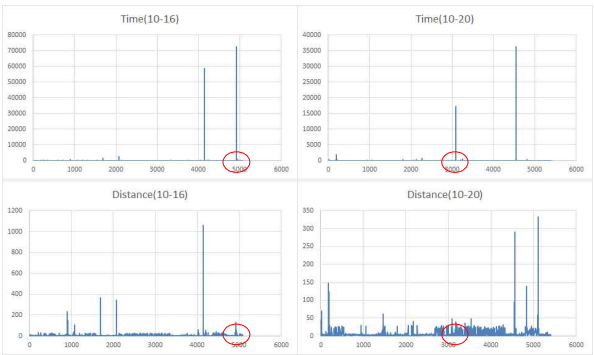

그림 16. 85번 유저의 시간의 순서에 따른 GPS 포인트간 시간과 거리

표 7에서 보는 것과 같이 85번 유저의 경우에는 하루 동안 걸어서 이동하였다. 85번 유저는 10월 16일에 1곳, 10월 20일에 1곳의 실내에서 머물렀을 가능성이 있다. 육안으로 그래프를 보면 GPS 포인트간 시간이 큰 구간과 거리가 상대적으로 짧은 구간이 실내에서 머무른 것으로 판단 할 수 있다. 즉, 실내에서 머문 것으로 판단할 수 있는 조건은 GPS 포인트간 시간이 비정상 적으로 높은 포인트를 찾아야 하며, 해당 GPS 포인트간 거리가 비정상적으로 높지 않아야 한다. 그래서 시간과 거리 각각 비정상적인 값, 즉 극단값을 찾아서 각각의 임계값으로 지정해야 한다. 시간의 경우 극단값 보다 커야 하고, 거리의 경우 극단값 보다 낮아야 한다. 실외 SP 추출에서 사용한 Z-Score를 이용하여 유의수준 α는 0.05로 설정한 후, 유의수준을 초과하면 Outlier로 판단하는 방법을 통해 유저가 실내에서 머문 지점을 추출한다.

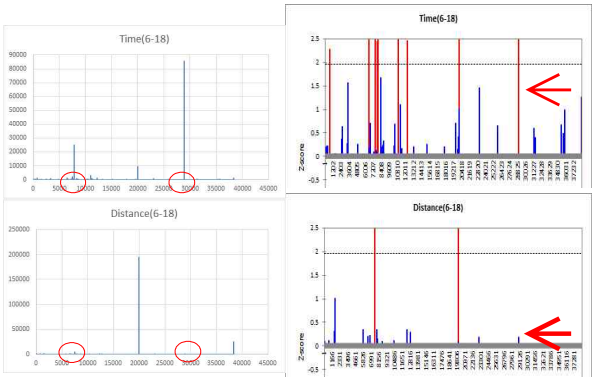

그림 17. 10번 유저의 Outlier detection을 통한 실내 SP 추출

그림 17은 10번 유저의 6월 18일 GPS 궤적에서 실내에서 머문 지점을 예상한 지점을 표시한 것이다. 예상되는 지점은 총 2곳으로 지정하였지만, Z-Score를 통해 Outlier detection을 한 결과 값은 오른쪽 그림과 같다. 예상한 2곳 중 한곳은 시간 임계값은 만족하지만 거리 임계값을 초과하여 실내 SP로 지정되지 못하였다. 나머지 한곳은 시간 임계값을 초과 하고 거리 임계값 미만이기 때문에 실내 SP로 지정된다.

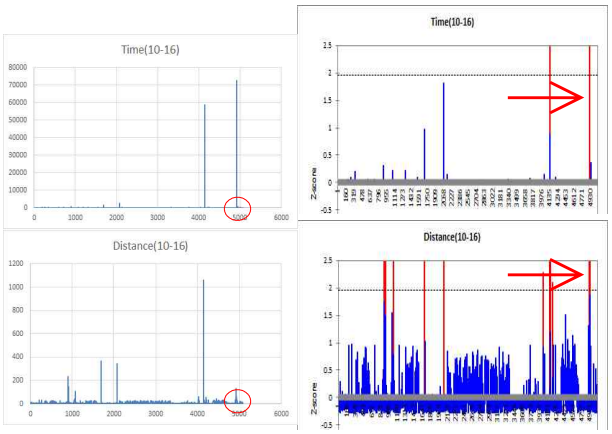

그림 18. 85번 유저의 Outlier detection을 통한 실내 SP 추출

그림 18은 85번 유저의 10월 16일 GPS 궤적에서 실내에서 머문 지점을 예상한 지점을 표시한 것이다. 85번 유저가 실내에서 머물렀을 거라고 예상되는 지점 1곳은, 시간 임계값은 만족하지만 거리 임계값을 초과함으로써 실내 SP로 지정되지 못한다.

Ⅳ. 실험환경 및 결과

1. 실험 환경 및 데이터셋 구성

가. 실험 환경

표 7. 실험환경

OS	Windows 7 64bit
CPU	Intel core i7-2630QM(2GHz)
RAM	8GB
DataSet	GeoLife

나. 데이터셋 구성

본 논문에서 제안한 알고리즘을 검증하기 위해 사용한 데이터셋은 GeoLife 데이터이다. GeoLife 데이터셋의 구성은 표 8과 같다.

표 8. GeoLife 데이터셋 구성

Field 1: Latitude in decimal degrees. Field 2: Longitude in decimal degrees. Field 3: All set to 0 for this dataset. Field 4: Altitude in feet (-777 if not valid). Field 5: Date - number of days (with fractional part) that have passed since 12/30/1899. Field 6: Date as a string. Field 7: Time as a string. Example: 39.906631,116.385564,0,492,40097.5864583,2009-10-11,14:04:30

본 논문에서는 Field 1,2,6,7을 사용하여 SP를 추출하였다.

2. SP 추출 실험 결과

본 장에서는 기존 알고리즘과 제안 알고리즘으로부터 생성된 SP를 비교하는 실험을 진행 하였다. 실험은 총 4가지의 비교를 통해 제안 알고리즘의 효율성을 검증하였다.

가. 추출된 SP 개수 비교

기존 알고리즘과 제안 알고리즘을 이용하여 GeoLife 데이터셋 전체 유저를 분석하여 SP를 추출한 개수를 비교하였다.

표 9는 각 알고리즘을 통해 생성된 SP의 개수이다. 기존 알고리즘의 경우 임계값에 따라 추출되는 SP의 개수가 천차만별임을 알 수 있다. 그림 19는 각각의 임계값을 통해 생성

된 SP를 히스토그램으로 나타낸 것이다.

표 9. 각 임계값에 따른 SP 추출 개수 비교

User	기존 알고리즘 (거리 임계값, 시간 임계값)				제안 알고리즘
0~182	(30, 100)	(30, 150)	(30, 200)	(25, 50)	51360
	54162	72322	58788	172566	
	(25, 100)	(20, 50)	(20, 100)	(20, 200)	
	84921	157085	75205	43525	

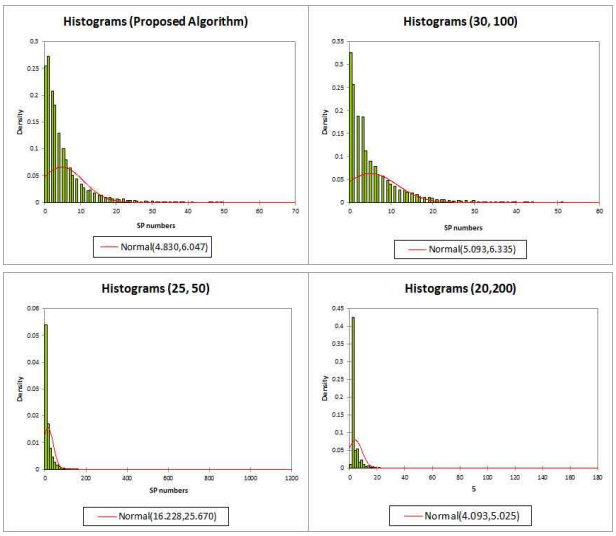

그림 19. 제안 알고리즘과 기존 알고리즘의 머문 지점 추출 개수 비교

그림 19는 각 날짜에 생성된 SP의 개수를 히스토그램으로 나타낸 것이다. 그림 19를 보면 제안 알고리즘과 (30, 100) 임계값을 사용하여 추출한 SP의 분포가 적절함을 알 수 있다. 하지만 (25, 50), (20, 50)을 적용한 결과 값은 분포가 고르지 못하기 때문에 표준편차가 매우 높음을 알 수 있다. 즉, 임계값에 따라 각 날짜마다 생성되는 SP의 개수가 변동됨을 알 수 있다. 또한, (25, 50) 임계값을 적용한 결과 값을 보면 유저가 하루에 1200곳이나 되는 지점에 머물렀다고는 가정할 수 없기 때문에 적절한 임계값을 적용하여 각 날짜마다 적당한 SP의 개수가 추출될 수 있도록 조정해야 한다. 각각의 임계값을 통해 추출된 SP가 적절하게 추출됐는지를 판단하기 위해 분산계수를 계산하면 다음과 같다.

표 10의 분산계수를 통해 각 날짜마다 적절하게 SP가 추출됐는지를 평가하였다. 분산계수가 작을수록 SP가 적절하게 추출됐다고 판단할 수 있다. 하지만 임계값 (20, 200)의 분산계수는 가장 작지만 유저가 하루에 156곳에서 머물렀다고는 보기 어렵기 때문에 해당 임계값을 채택하기는 힘들다. 표 10의 다양한 임계값들 중에서, 제안 알고리즘을 통해 생성된 SP 개수와 분산계수 및 max 값이 비슷하면서 기존 알고리즘에서 가장 대표적인 임계값인 (30, 100)을 통해 생성된 SP를 이용하여 제안 알고리즘의 효율성을 비교 및 검증 하고자 한다.

표 10. 분산계수 비교를 통한 성능 비교

임계값	daily SP's		평균	표준편차	분산계수
	min	max			
(30, 100)	**0**	**58**	**5.093**	**6.335**	**1.24386**
(30, 150)	1	370	6.801	9.954	1.463608
(30, 200)	1	267	5.528	7.508	1.358176
(25, 50)	1	1031	16.228	25.670	1.581833
(25, 100)	1	479	7.986	12.469	1.561357
(20, 50)	1	958	14.772	24.141	1.634240
(20, 100)	1	353	7.072	10.819	1.529835
(20, 200)	1	156	4.093	5.025	1.227752
(35, 150)	1	353	6.581	9.293	1.412095
제안 알고리즘	**0**	**61**	**4.830**	**6.047**	**1.25196**

나. 임계값 변화에 따른 SP 위치 비교

기존 알고리즘을 사용하여 SP를 추출할 경우 임계값에 따라 SP의 위치가 변화한다.

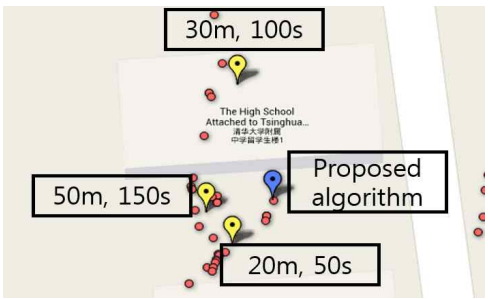

그림 20. 임계값에 따른 SP 위치 변화

그림 20은 3번 유저의 2008년 10월 23일 데이터를 대상으로 각 임계값에 따라 SP를 추출한 것이다. 빨간색 점은 GPS 궤적이며, 노란색 마크는 기존 알고리즘으로 추출된 SP, 파란색 마크는 제안 알고리즘을 통해 추출된 SP이다. 기존 알고리즘을 통해 추출된 SP를 보면 임계값의 변화에 따라 SP의 위치도 변화하는 것을 볼 수 있다. 즉, 임계값에 따라 추출되는 SP의 개수도 변동되고 SP 위치 또한 변화한다. SP 위치의 모호성은 실내와 실외를 구분할 수 없는 이유뿐만 아니라, 임계값에도 영향을 받는다는 것을 그림 20을 통해 알 수 있다. 제안 알고리즘은 임계값을 지정하지 않고 SP를 추출하기 때문에 SP 개수의 변동이나 위치의 변화가 없다.

다. 실제 머문 지점 추출 여부 비교

기존 알고리즘을 사용하여 SP를 추출할 경우 육안으로 봐도 머무른 지점임에도 불구하고 SP로 추출되지 않는 경우가 있다. 사용자가 임의로 설정한 임계값 때문에 유저가 실제로 머문지점을 추출하지 못하기 때문이다. 그림 21은 제안 알고리즘으로는 추출되는 지점이지만 기존 알고리즘으로는 추출되지 않는 지점을 보여주고 있다.

그림 21. 기존 알고리즘으로는 추출되지 않는 장소

그림 21은 2번 유저의 2008년 11월 6일 데이터를 대상으로 기존 알고리즘과 제안 알고리즘을 적용하여 SP를 추출한 것이다. 그림 21에서 빨간 점은 GPS 궤적이며 녹색 마크는 실외에서 머무른 지점이다. 제안 알고리즘을 통해 그림 21의 각 장소에서 유저는 130초, 195초 머문 사실이 있음을 알 수 있다. 물론, 임계값을 조정해서 그림 21의 장소를 추출할 수는 있겠지만 3장에서 언급한 바와 같이 모든 임계값의 조합을 통해 SP를 추출하는 것은 불가능하며 그림 21에서 보는 것과 같이 임계값의 변화에 따라 SP의 위치도 변화하기 때문에 해당 위치를 추출할 수 있음을 장담할 수 없다. 또한, 기존 알고리즘은 실제로 머문 지점이 아닌데도 불구하고 SP로 추출되는 경우가 있다. 그림 22는 제안 알고리즘으로는 SP로 추출되지 않고, 기존 알고리즘으로만 추출되는 장소이다.

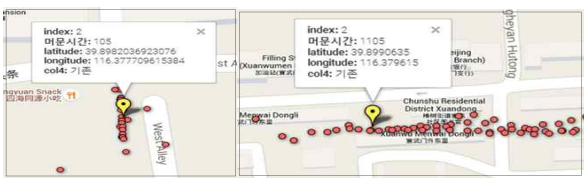

그림 22. 기존 알고리즘으로만 추출되는 SP

유저가 야외에서 머문 지점은 무지향성을 띄어야 한다. 그림 22에서 보는 것과 같이 속도는 만족하지만 무지향성이 아니기 때문에 제안 알고리즘을 통해 SP로 추출되지 않은 장소이다. 즉, 그림 22의 장소에서 유저는 느리고 일정한 방향으로 이동한 것으로 판단할 수 있기 때문에 머물렀다고 판단할 수 없다.

라. 실내 SP 추출 여부 비교

GeoLife 데이터셋에서 2번 유저의 2008년 10월 24일 데이터를 대상으로 기존 알고리즘과 제안 알고리즘을 통해 SP를 추출하는 실험을 진행하였다. 그림 23은 제안 알고리즘으로 생성된 SP 중 실내에서 머문 SP와 기존 알고리즘으로 생성된 SP를 보여주고 있다.

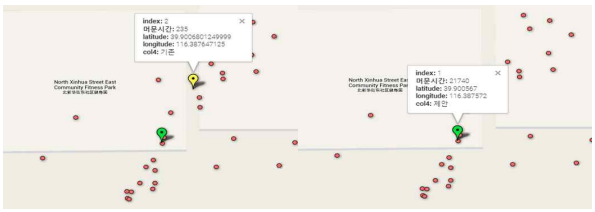

그림 23. 실내 SP 추출

그림 23에서 빨간 점들은 GPS 궤적이며, 녹색 마크는 제안한 방법으로 생성된 실내 SP이다. 그리고 노란색 마크는 기존 알고리즘으로 생성된 SP이다. 기존 알고리즘으로 부터 생성된 SP는 유저가 실외에서 머물렀는지 실내에서 머물렀는지를 판단할 수가 없다. 제안 알고리즘을 통해 생성된 SP를 보면 유저가 Fitness Park에서 21740초 동안 머문 사실이 있음을 보여주고 있다.

마. 구글지도를 이용한 SP 출력

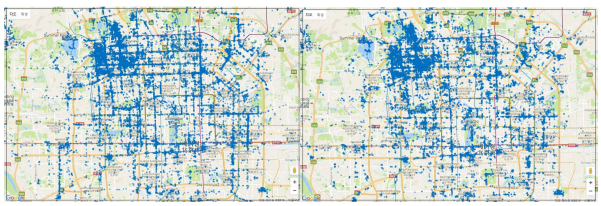

그림 24. 제안 알고리즘과 기존 알고리즘의 SP 추출

그림 24는 Geolife data set의 전체 유저를 대상으로 SP를 추출한 것이다. 왼쪽 그림은 제안 알고리즘을 통해 생성된 SP이고, 오른쪽 그림은 기존 알고리즘에서 임계값 (30, 100)을 사용하여 추출된 SP를 나타낸 것이다. 서로 다른 알고리즘을 통해 SP를 추출하였지만 그림 24에서 보는 것과 같이 비슷한 위치에 SP가 집중되어 있는 것을 볼 수 있다. 즉, 기존 알고리즘과 비교했을 때 제안 알고리즘도 적절하게 유저들이 머문 지점을 추출할 수 있음을 확인할 수 있다.

기존 알고리즘의 경우 모든 임계값들의 조합을 통해 그림 24와 같은 적절한 SP를 추출하는 과정이 쉽지 않다. 제안 알고리즘의 경우에는 GPS 궤적을 분석하여 적절한 SP를 추출하기 때문에 효율성이 높다고 할 수 있다.

V. 결 론

본 논문에서는 기존의 SP 추출 알고리즘에서 제기 되던 임계값의 임의설정에 의한 신뢰도 저하 문제와, SP 위치에 대한 모호성의 문제점을 해결하기 위해 새로운 SP 추출 알고리즘을 제안하였다. 제안 알고리즘은 기존 알고리즘에서 사용하는 임계값들을 사용하지 않고 GPS 궤적의 분포를 분석하여 SP를 추출하는 통계적 분석방법을 사용한다. 그리고 SP 위치에 대한 모호성을 개선하기 위해 실내와 실외의 SP를 각각 추출한다. 실내와 실외의 SP를 각각 추출한 이유는 향후 PoI(Point of Interest) 추출과 궤적 유사도 분석을 할 때 유용하게 사용되기 때문이다. 실내에서 머문 지점은 실내에서 머문 SP 끼리 클러스터링 하고, 실외에서 머문 지점은 실외에서 머문 SP 끼리 클러스터링 한다. 왜냐하면, 머문 지점에 대한 모호성이 높은 상태에서 PoI 추출 및 궤적 유사도 분석을 하게 되면 알고리즘에 대한 신뢰성이 떨어지기 때문이다. 실내에서 생성되는 SP와 실외에서 생성되는 SP의 성향은 각각 다르지만 기존 알고리즘의 경우 실내와 실외를 가리지 않고 한 번에 추출함으로서 SP 위치의 모호성이 높다. 본 논문에서는 실내와 실외에서 머문 지점을 찾기 위해 가우시안 확률분포를 사용하여 각각의 임계값을 지정한 후 실내와 실외에서 머문 지점을 추출한다.

실험 결과, 기존 알고리즘과 제안 알고리즘은 동일한 지점의 SP를 추출할 수 있었지만, 일부 상이한 지점의 SP는 기존 알고리즘이 실제로 머문지점이 아닌 지점을 SP로 추출한 것을 실험결과를 통해 알 수 있었다. 그리고 기존 알고리즘은 유저가 실내에서 머물렀는지 실외에서 머물렀는지를 분간 할 수 없지만, 제안 알고리즘은 그림 23처럼 유저가 명확하게 실내에서 머물렀음을 실험결과로 확인할 수 있다. 또한, 기존 SP 알고리즘은 시간과 거리에 대한 임계값을 조합을 통해 추출하기 때문에 추출하는 과정이 쉽지 않다. 제안 알고리즘은 통계적 분석방법을 통해 임계값을 지정하지 않고 머문 지점을 효과적으로 추출해낼 수 있음을 실험을 통해 확인할 수 있다.

References

[1] 이향진; 최정화; 박영택. 모바일 사용자의 대용량 GPS 데이터로부터 시맨틱 관심지점 발견. 정보과학회논문지: 소프트웨어 및 응용, 2012, 39.3: 175-184.

[2] GUIDOTTI, Riccardo, et al. Retrieving Points of Interest from Human Systematic Movements. In: Software Engineering and Formal Methods. Springer International Publishing, 2014. p. 294-308.

[3] ZIGNANI, Matteo; GAITO, Sabrina. Extracting human mobility patterns from GPS-based traces. In: Wireless Days (WD), 2010 IFIP. IEEE, 2010. p. 1-5.

[4] ZHENG, Yu, et al. Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th international conference on World wide web. ACM, 2009. p. 791-800.

[5] XIAO, Xiangye, et al. Finding similar users using category-based location history. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2010. p. 442-445.

[6] XIU-LI, Zhao; WEI-XIANG, Xu. A clustering based approach for discovering interesting places in a single trajectory. In: Intelligent Computation Technology and Automation, 2009. ICICTA'09. Second International Conference on. IEEE, 2009. p. 429-432.

[7] 장미영(Miyoung-Young Jang), 윤민(Min Yoon), 장재우(Jae-Woo Chang), "위치 기반 서비스 환경에서 이동 객체 궤적데이터 마이닝 기법에 대한 고찰", 한국정보과학회 데이터베이스 소사이어티, 28권 제1호 2012년 4월

[8] LI, Quannan, et al. Mining user similarity based on location history. In: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems. ACM, 2008. p. 34.

[9] Haidong Zhong; Shaozhong Zhang; Yanling Wang; Shifeng Weng; Yonggang Shu, "Mining Users' Similarity from Moving Trajectories for Mobile E-commerce Recommendation", International Journal of Hybrid Information Technology;2014, Vol. 7 Issue 4, p309, July 2014

[10] Zhao Xiu-li, Xu Wei-xiang, "A Clustering-Base

d Approach for Discovering Interesting Places in a Single Trajectory", Intelligent Computation Technology and Automation, 2009. ICICTA '09. Second International Conference on (Volume:3), pp.429 - 432, 10-11 Oct. 2009

[11] Zignani, M, Gaito, S, "Extracting human mobility patterns from GPS-based traces", Wireless Days (WD), 2010 IFIP, pp.1-5, 20-22 Oct. 2010

[12] Luis Otavio Alvares , Jose Antonio Fernandes , De Macedo , Vania Bogorny , Bart Moelans , Bart Kuijpers , Alejandro Vaisman, "A model for enriching trajectories with semantic geographical information", GIS '07 Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems, No. 22, 2007

[13] Andrey Tietbohl , Vania Bogorny , Bart Kuijpers , Luis Otavio Alvares, "A Clustering-based Approach for Discovering Interesting Places in Trajectories", Proceedings of the 2008 ACM symposium on Applied computing. pp. 863-868, 2008.

[14] Quannan Li , Yu Zheng , Xing Xie , Yukun Chen , Wenyu Liu , Wei-ying, "Mining user similarity based on location history", Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems. pp. 34, 2008.

[15] Yu Zheng, Xing Xie, and Wei-Ying Ma, "Mining Interesting Locations and Travel Sequences From GPS Trajectories", WWW '09 Proceedings of the 18th international conference on World wide web, Pages 791-800, 2009.

[16] Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, Wang-Chien Lee, Tz-Chiao Weng, Vincent S. Tseng, "Mining user similarity from semantic trajectories", LBSN '10 Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, Pages 19-26, 2010.

[17] Kyunghan Lee, Seongik Hong, Seong Joon Kim, Injong Rhee and Song Chong, "Demystifying Levy Walk Patterns in Human Walks", Technical Report, CSC, NCSU, 2008.

[18] Xihui Chen, Jun Pang, Ran Xue, "Constructing and comparing user mobility profiles for location-based services", SAC '13 Proceedings of the 28th Annual ACM Symposium on Applied Computing, Pages 261-266, 2013.

[19] Ruipeng, L. U. "New User Similarity Measures Based on Mobility Profiles.", 2013.

[20] 김범수(Beoum-Su Kim), 이연(Yeon Lee), 김경배(Gyeong-Bae Kim), 배해영(Hae-Young Bae), "The Development of Users' Interesting Points Analyses Method and POI Recommendation System for Indoor Location Based Services", 韓國컴퓨터情報學會論文誌 第17卷 第5號 / 81~91 (11pages), 2012

[21] 이향진, 최정화, 박영택, "Semantic Point of Interest Detection from Large-scale GPS Data of Mobile Users", 정보과학회논문지 : 소프트웨어 및 응용 제39권 제3호, 175-184 (10 pages), 2012.3

[22] 이병문, "스마트 기기를 활용한 보행속력에 따른 맞춤보폭의 적용", 한국콘텐츠학회논문지 제13권 제4호, 35-43 (9 pages), 2013.4

[23] ZHENG, Yu, et al. Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th international conference on World wide web. ACM, 2009. p. 791-800.

[24] 김진수, 반성범, "객체 특징을 이용한 실시간 배회 행위 검출", 스마트미디어저널, 2016, vol.5, no.3, pp.93-98

[25] 이귀상, 여정연, "야간 영상 감시를 위한 GMM기반의 배경 차분", 스마트미디어저널, 2015, vol.4, no.3, pp. 50-55

[26] 안병익, 정구임, 최혜림, "다속성 태도 모델과 협업적 필터링 기반 장소 추천 연구", 스마트미디어저널, 2016, vol.5, no.2, pp. 84-89

[27] 민무홍, 김태연, 윤태복, 이지형, "GPS와 각도 변화량을 이용한 경로 요약에 관한 연구", 한국지능시스템학회 20주년 기념 2010년도 추계학술대회 학술발표논문집 제20권 제2호, 159-161 (3 pages),

2010.11

저 자 소 개

박진관(학생회원)

2011년 목포대학교 컴퓨터교육과 공학사
2013년 목포대학교 컴퓨터공학과 공학 석사
2013년~현재 : 목포대학교 컴퓨터공학과 박사과정
<주관심분야 : 궤적 데이터 마이닝, 데이터마이닝 ,무선센서 네트워크, 증강현실, 실내위치측위>

오수열(종신회원)

1981년 전남대학교 재료공학사
1985년 조선대학교 전산학석사
2004년 전남대학교 전산통계박사수료
2005년 호주 타스마니아대학 객원교수
현재 목포대학교 컴퓨터공학과 교수
<주관심분야 : 프로젝트관리, 소프트웨어 품질, 공학교육>