인간 이동속도의 분포 추정

이준석*, 송하윤
홍익대학교 컴퓨터공학과
*rolunoa@naver.com, hayoon@hongik.ac.kr

Estimating Distribution of Human Mobility Speed

Jun-Seok Lee*, Ha Yoon Song
Dept of Computer Engineering, Hongik University

요 약

학문적 및 사회적으로 인간의 위치 예측이 높은 가치를 가지고 있다. 위치 예측을 실현시키기 위해 많은 노력을 기울이고 있으며, 이의 기본으로 인간 이동 속도의 분포가 요구된다. 이 논문에서 단순 위치, 장소 데이터 분석이 아닌 위치 데이터에서 파생된 인간의 이동속도를 전체구간, 즉 정지상태의 속도부터 충분히 나올 수 있는 최대속도까지가 아닌 이동 시 나올 수 있는 유효한 속도를 활용하여 분석하였다. 이 논문에서는 이러한 인간 속도의 빈도를 계산하여 속도가 어떠한 확률분포를 따르는지, 그리고 확률 분포들 중 비교적 계산이 용이한 지수 분포와 어느 정도의 유사도를 가지는지 분석할 것이다.

1. 서론

누구나 인간의 위치를 예측하고 싶은 열망을 가지고 있다. 위치 예측을 실현시키기 위해 많은 노력을 기울이고 있고 대다수의 학자들은 위치를 예측하기 전 과거 위치데이터를 분석할 필요를 느껴 현재 위치데이터 분석에 대한 연구들이 활발히 이루어지고 있다. 그 중 가장 기본적인 위치 패턴 분석 분야가 있고 이를 기반 해주는 분야로 위치 클러스터링 분야가 있다. 위치 클러스터링은 간단한 위치데이터(위도, 경도, 시각)만으로 사용자가 어느 특정 위치에 머물러있었는지에 대한 클러스터링 알고리즘이다.

우리는 단순 위치, 장소 데이터 분석이 아닌 위치 데이터에서 파생된 인간의 이동속도를 분석하였다. 인간의 이동속도가 특정 확률 분포와 유사하다면 인간에게서 나올 수 있는 속도를 모델링하여 불규칙처럼 보이는 속도를 어느 정도 신뢰도를 가지고 예측할 수 있을 것이다.[1][2]

이 논문에서는 위치데이터를 이용해 나온 인간의 이동속도가 어떠한 확률분포를 따르는지, 그리고 이러한 확률분포들 중 비교적 계산이 용이한 확률분포 함수인 지수분포와 어느 정도의 유사도를 가지는지 확인해 볼 것이다. 이렇게 인간의 이동속도를 모델링하면 교통공학, 도시공학 분야에서 유용하게 사용할 수 있을 것이고 확률을 통한 GPS 오류 검출 또한 가능케 할 것이다.

2장에서는 이 연구의 배경을 설명할 것이고 3장에서는 데이터수집의 방법, 4장에서는 데이터와 분포의 접합을 논할 것이다.

2. 연구 배경

자연과학분야에서는 바람의 속도, 분자의 운동 속도 등 인간이 예측 할 수 없는 자연현상에 확률 분포에 찾기 위해 많은 실험을 행하고 있다. 여러 가지 이유가 있겠지만 그 중 하나는 불규칙처럼 보이는 움직임, 운동을 모델화하여 다음 움직임, 운동을 예측하기 위함 일 것이다. 이에 입각하여 우리는 인간의 이동속도가 불규칙해보일지라도 많은 양의 위치 데이터를 분석하여 유사한 확률 분포를 찾아낸다면 인간의 움직임을 확률적 모델링을 할 수 있고 다음 속도 또한 예측해 낼 수 있다고 생각하였다.

보편적인 인간의 이동은 출발지에서 특정 목적을 갖고 목적지로 각종 수단(보도, 자전거, 대중교통, 자동차)을 이용해 이동하는 동안 발생한다. 이렇게 발생하는 이동속도는 직관적으로 생각해보면 불규칙으로 느껴질 것이다. 하지만 지금까지 행해진 연구를 보면 인간의 이동속도는 로그정규 분포, 지수 분포, Weibull 분포 등을 비교적 높은 유사도로 따른다 하였다.[3]

우리의 분석은 여기서 조금 더 발전시켜 속도의 전체구간, 즉 정지상태의 속도부터 충분히 나올 수 있는 최대속도까지가 아닌 이동 시 나올 수 있는 유효한 속도를 활용하여 각 종 분포들의 유사도 측정을 할 것이다.

3. 데이터 수집

우리 연구에서 쓰인 위치 데이터는 8명의 지원자들로부터 2013년 5월부터 2014년 2월까지의 기간 동안 수집된 것이다. 자신의 집에 있는 상태를 제외하고 외출하고 있는 동안의 모든 정지 상태, 이동 상태 위치를 시각과 함께 기록된 것이다. 데이터 수집을 위해 사용된 기기는 GPS통신과 기록이 가능한 스마트폰과 Gamin 시리즈이다. GPS신

호가 원활하다면 초당 한 개의 위치 데이터가 기록되도록 설정 하였다. 이렇게 약 10개월간 수집된 순수 위치 데이터의 수는 총 2,218,020개이다.

수집된 데이터는 위도, 경도, 시각을 포함하고 있고 한국시각 기준으로 일 단위로 분리한 뒤 Haversine 공식을 이용하여 속도를 추출하였다. GPS신호에 문제가 없으면 데이터 수집에 문제가 없지만 GPS신호가 잡히지 않으면 위치가 기록되지 않는다. 이때 속도의 빈도를 왜곡할 수 있지만 실내에서의 이동은 우리 연구에서 다루고 있지 않다 . 하지만 터널, 지하철에서는 우리 연구에서 필요한 속도가 발생하지만 GPS신호가 잡히지 않기 때문에 위치 데이터로 기록 되지 않는다. 이러한 점은 속도의 빈도가 왜곡되어 모델링하는데 있어 신뢰도가 떨어지는 문제가 발생 시킨다. 하지만 우리는 데이터 수가 많으면 이러한 문제를 무시 할 수 있다고 가정하고 접합을 진행하였다.

4. 데이터 접합

우리는 인간의 최소 이동속도를 2m/s라 정의하였고 각 상황마다 최대로 나올 수 있는 속도를 30m/s(108km/h), 50m/s(180km/h), 70m/s(252km/h), 100m/s(360km/h)라 정의하여 분석하였다. 또한 속도 정밀도에 상관없이 확률분포에 대한 유사도를 측정하기 위해 속도의 단위를 0.5m/s, 0.1m/s, 0.01m/s 으로 나누어 데이터 접합을 하였다.

데이터 접합의 방식은 다음과 같다. 위치데이터에서 나온 속도와 빈도수를 통해 각 속도에 대한 확률을 계산한 뒤 MLE(Maximum Likelihood Estimation)을 이용해 각각의 확률 밀도 함수 매개변수를 구한다. 그리고 확률 분포와 데이터 누적 분포의 비교를 Kolmogorov-Smirnov Test(K-S test)를 통하여 유사도의 측도가 되는 Statistic을 계산해 분포들의 순위를 결정한다.[4][5]

4.1 0.5m/s단위

<표 1>는 0.5m/s 단위 기준의 2.0m/s ~ 30.0m/s 구간에 존재하는 속도와 속도에서 나온 표본개수를 확률 분포에 접합한 결과이다.

<표 1>를 보면 분포에 대한 Statistic값이 낮을수록 높은 순위의 유사한 분포임을 알 수 있다. 여기서 주목해야 할 점은 21위에 위치하고 있는 Phased Bi-Exponential 분포이다. 이 분포의 매개변수를 보면 gamma1의 값이 2이고 gamma2의 값은 25.0이다. 또한 24위에 위치하고 있는 지수 분포보다 높은 유사도를 보인다. 여기서 우리는 2.0m/s - 24.5m/s 구간과 24.5m/s - 30.0m/s 구간에서 각각 다른 lambda값을 가지는 지수 분포를 따른다 해석하였다. 그리하여 2.0m/s - 24.5m/s 구간을 잘라내어 부분 접합을 행하였다. <표 2>가 그 결과이다.

<표 1> 2.0m/s ~ 30.0m/s에서의 확률 분포 접합 결과 순위

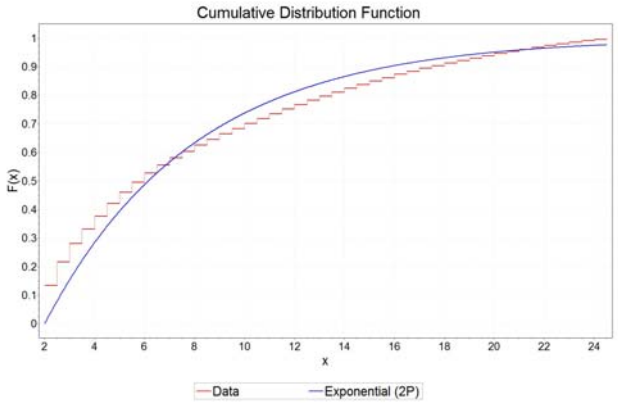
2.0 ~ 30.0(0.5m/s)			
순위	분포이름	Statistic	매개변수
1	Kumaraswamy	0.07121	α=0.485, α2=1.529, a=2.0, b=32.083
2	Frechet(3P)	0.09069	α=1.303, β=3.699, ɣ=2.0
3	Weibull(3P)	0.09098	α=0.714, β=5.016, ɣ=2.0
...	...	...	...
21	Phased Bi-Exponential	0.12996	ɣ1=2, λ1=0.169, ɣ2=25.0, λ2=0.072
...	...	...	...
24	Exponential	0.13688	ɣ=2, λ1=0.154

<표 2> 2.0m/s ~ 24.5m/s에서의 확률 분포 접합 결과 순위

2.0 ~ 24.5(0.5m/s)			
순위	분포이름	Statistic	매개변수
1	Gamma(3P)	0.07187	α=0.623, β=9.324, ɣ=2.0
2	Johnson SB	0.08058	ɣ=0.844, δ=0.564, λ=23.573, ξ=1.566
3	Frechet(3P)	0.09290	α1=0.1.373, β=3.710, ɣ=0.328
...	...	...	...
27	Phased Bi-Exponential	0.13441	ɣ1=2, λ1=0.172, ɣ2=21.0, λ2=0.018
28	Exponential	0.13666	ɣ=2, λ1=0.167

<표 2>의 28위에 위치하고 있는 지수 분포의 Statistic 값을 주목해보자. <표 1>에서의 지수 분포 Statistic값보다 줄어든 것을 확인 할 수 있다. 추가적으로 <표 2>에서 나타난 Phased Bi-Exponential 분포의 매개변수를 이용하여 한 번 더 부분 접합을 시행하게 되면 오히려 지수 분포의 Statistic값이 높아진다. 그리하여 우리는 0.5m/s 단위에서의 속도는 2.0m/s - 24,5m/s 구간에서 0.13666의 오차로 지수 분포를 따른다고 결론 내렸다. <그림 1>은 2.0m/s - 24.5m/s구간 내 에서의 데이터 확률 누적 그래프와 지수 확률 분포의 누적 그래프이다.

같은 방식으로 2.0m/s부터 50m/s, 70m/s, 100m/s까지 초기 접합을 하였을 때는 Phased Bi-Exponential 분포와 지수 분포의 Statistic 차이가 거의 없거나 지수 분포가 유사도가 더 높다는 결과가 나왔다.



<그림 1> 2.0m/s ― 24.5m/s에서의 데이터 확률 누적 그래프와 지수 확률 분포의 누적 그래프

<표 3> 2.0m/s ~ 22.0m/s에서의 확률 분포 접합 결과 순위

2.0 ~ 22.0(0.1m/s)			
순위	분포이름	Statistic	매개변수
1	Johnson SB	0.02850	γ=0.754, δ=0.564, λ=20.714, ξ=1.815
2	Kumaraswamy	0.03032	α1=0.677, α2=1.617, a=2.0, b=23.117
3	Beta	0.04381	α1=0.537, α2=1.301, a=2.0, b=22.259
...	...	...	...
6	Exponential	0.06188	γ=2, λ1=0.169
...	...	...	...
8	Phased Bi-Exponential	0.06796	γ1=2, λ1=0.174, γ2=16.5, λ2=0.662

4.2 0.1m/s단위

<표 3>은 0.5m/s단위에서 행한 접합과 같은 방식으로 0.1m/s 단위의 속도에서 2.0m/s - 30.0m/s구간내의 초기 데이터 접합을 하고 Phased Bi-Exponential 분포의 매개변수를 바탕으로 지수 분포의 Statistic값이 최고로 낮아질 때 까지 부분 접합한 최종 결과이다. 지수 분포의 Statistic값이 최고로 낮아졌을 때의 속도 구간은 2.0m/s - 22.0m/s이다. <그림 2>와 <그림 1>을 비교해보면 0.1m/s단위, 2.0m/s - 22.0m/s구간에서의 접합이 0.5m/s 단위, 2.0m/s - 24.5m/s구간에서의 접합 결과와 비슷한 그래프가 나오는 것을 확인 할 수 있다.

다음에 나오는 <표 4>는 2.0m/s - 50.0m/s에서 데이터 접합을 한 최종 결과이고 70m/s, 100m/s까지의 초기 접합은 지수 분포가 더 높은 유사도로 나타났다.

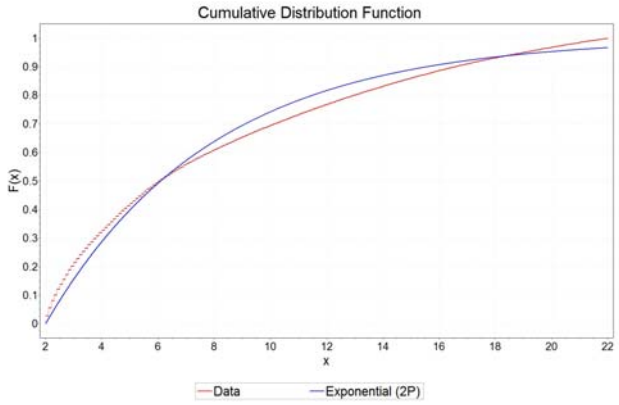
<표 4>를 통해 우리는 2.0m/s - 30.0m/s구간에서 접합을 하던, 2.0m/s - 50.0m/s구간에서 접합을 하던 지수 분포와 높은 유사도를 나타내는 구간은 2.0m/s부터 약 22.0m/s까지의 속도임을 확인 할 수 있다.

4.3 0.01m/s단위

이번 절에서는 기존 0.5m/s, 0.1m/s의 속도단위보다 정밀도가 더 높은 0.01m/s단위에서의 접합을 하였다. 첫 번째로 2.0m/s - 30.0m/s구간에서의 초기 접합결과에서의 Phased Bi-Exponential분포의 매개변수를 바탕으로 지수 분포가 더 높은 Statistic값을 가질 때까지 재귀적으로 부분 접합을 하였다. <표 5>는 최종적으로 지수 분포의 Statistic값이 가장 낮을 때의 결과이고 그 구간은 2.00m/s - 18.63m/s이다.

<표 4> 2.0m/s ~ 22.7m/s에서의 확률 분포 접합 결과 순위

2.0 ~ 22.7(0.1m/s)			
순위	분포이름	Statistic	매개변수
1	Kumaraswamy	0.02012	α1=0.325, α2=1.617, a=2.0, b=23.597
2	Johnson SB	0.03022	γ=0.768, δ=0.567, λ=21.463, ξ=1.799
3	Beta	0.0440	α1=0.536, α2=1.319, a=2.0, b=22.967
...	...	...	...
6	Phased Bi-Exponential	0.06233	γ1=2, λ1=0.170, γ2=17.2, λ2=0.633
7	Exponential	0.06280	γ=2, λ1=0.165



<그림 2> 2.0m/s ― 22.0m/s에서의 데이터 확률 누적 그래프와 지수 확률 분포의 누적 그래프

<표 6>과 <그림 3>은 2.00m/s ~ 50.00m/s 구간내의 초기 접합을 통해 얻은 자료를 바탕으로 지수 분포의 유사도가 최고로 높아지는 구간의 부분 접합결과와 그래프이다.

5. 결론 및 분석

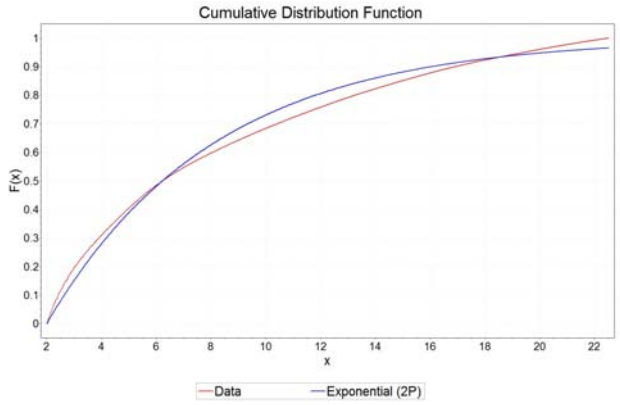
우리는 위 접합 결과들을 통해 인간의 이동속도 분포가 특정 구간에서 비교적 계산이 용이한 지수 분포와 높은 유사도를 가진다는 점을 확인 할 수 있었다. 즉, 2.0m/s부터 대략 20.0m/s ~ 22.0m/s까지의 속도 안에서 특정 속도가 나올 수 있는 확률은 지수 확률을 따른다 말할 수 있다. 다시 말하여 특정 속도 구간에서의 인간 이동속도는 지수 분포로 충분히 근사하며, 이때의 K-S 통계량은 충분히 작은 값을 가지고, 이 상황 에서 지수 분포를 사용할여 인간 이동 속도를 추정할 때 일정 수준 이하의 유의수준 내에 포함된다고 말할 수 있다. 이를 통해 위 구간에서의 속도를 지수 분포로 모델링하면 다음 이동속도

<표 5> 2.00m/s ~ 18.63m/s에서의 확률 분포 접합 결과 순위

2.00 ~ 18.63(0.01m/s)			
순위	분포이름	Statistic	매개변수
1	Johnson SB	0.01092	γ=0.678, δ=0.551, λ=16.992, ξ=1.949
2	Reciprocal	0.02845	a=2.0, b=18.63
3	Kumaraswamy	0.03043	α1=0.747, α2=1.730, a=2.0, b=20.223
...	...	...	...
6	Exponential	0.05234	γ=2, λ1=0.190
...	...	...	...
31	Phased Bi-Exponential	0.08018	γ1=2, λ1=0.185, γ2=13.99, λ2=2.327

<표 6> 2.00m/s ~ 22.52m/s에서의 확률 분포 접합 결과 순위

2.00 ~ 22.52(0.01m/s)			
순위	분포이름	Statistic	매개변수
1	Johnson SB	0.02244	γ=0.755, δ=0.569, λ=21.182, ξ=1.862
2	Reciprocal	0.03178	a=2.0, b=22.52
3	Gen. Gamma	0.03439	k=1.537, α=0.488, β=12.262, γ=2.0
...	...	...	...
8	Exponential	0.04996	γ=2, λ1=0.164
...	...	...	...
12	Phased Bi-Exponential	0.05322	γ1=2, λ1=0.162, γ2=17.74, λ2=2.072



<그림 3> 2.00m/s ~ 22.52m/s에서의 데이터 확률 누적 그래프와 지수 확률 분포의 누적 그래프

를 예측 할 수 있을 것이고, 예측뿐만 아니라 과거 속도를 통하여 GPS 오류검출 분야에도 응용 될 수 있을 것이다. 즉, 과거 위치데이터의 속도를 이용해 지수 분포를 모델링한 뒤 현재 나올 수 있는 최대의 속도를 95%의 신뢰성으로 계산하고 현재 측정된 실제속도와 비교하여 현재속도가 최대속도 이상이라면 오류라 판단하는 등의 알고리즘을 작성할 수 있다. 이와 유사하게 인간의 이동 속도에 고나한 연구가 필요한 분야에서 인간의 속도를 표현하는 확률 분포 모델을 도입할 수 있다면 여러 위치 기반 서비스 분야에서 응용 될 수 있을 것이다.

6. Acknowledgement

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2012R1A2A2A03046473)

참고문헌

[1] Marta C. Gonzalez, Cesar A. Hidalgo, Albert-Laszlo Barabasi "Understanding Individual Human Mobility Pattern" Nature 453, 779 - 782, June 2008.

[2] Hyun-Uk Kim, Ha-Yoon Song "Daily Life Mobility of a Student: From Position Data to Human Mobility Model through Expectation Maximization Clustering", Multimedia, Computer Graphics and Broadcasting Communications in Computer and Information Science Volume 263, 2011, pp 88-97.

[3] Kim, Minkyong, David Kotz, and Songkuk Kim. "Extracting a Mobility Model from Real User Traces." INFOCOM. Vol. 6. 2006.

[4] Scholz, F. W. "Maximum likelihood estimation." Encyclopedia of Statistical Sciences (1985).

[5] Massey Jr, Frank J. "The Kolmogorov-Smirnov test for goodness of fit." Journal of the American statistical Association 46.253 (1951): 68-78.