Computer Networks 64 (2014) 296–307

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Estimating human trajectories and hotspots through mobile phone data

Sahar Hoteit [a,*], Stefano Secci [a], Stanislav Sobolevsky [b], Carlo Ratti [b], Guy Pujolle [a]

[a] Sorbonne Universities, UPMC Univ. Paris 06, UMR 7606, LIP6, F-75005 Paris, France
[b] MIT Senseable City Laboratory, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 23 September 2013
Received in revised form 2 January 2014
Accepted 9 February 2014
Available online 20 February 2014

Keywords:
Mobility patterns
Interpolation methods
Trajectory estimation
Radius of gyration
Hotspot estimation

ABSTRACT

Nowadays, the huge worldwide mobile-phone penetration is increasingly turning the mobile network into a gigantic ubiquitous sensing platform, enabling large-scale analysis and applications. Recently, mobile data-based research reached important conclusions about various aspects of human mobility patterns. But how accurately do these conclusions reflect the reality? To evaluate the difference between reality and approximation methods, we study in this paper the error between real human trajectory and the one obtained through mobile phone data using different interpolation methods (linear, cubic, nearest interpolations) taking into consideration mobility parameters. Moreover, we evaluate the error between real and estimated load using the proposed interpolation methods. From extensive evaluations based on real cellular network activity data of the state of Massachusetts, we show that, with respect to human trajectories, the linear interpolation offers the best estimation for sedentary people while the cubic one for commuters. Another important experimental finding is that trajectory estimation methods show different error regimes whether used within or outside the "territory" of the user defined by the radius of gyration. Regarding the load estimation error, we show that by using linear and cubic interpolation methods, we can find the positions of the most crowded regions ("hotspots") with a median error lower than 7%.

1. Introduction

Human mobility and behavior pattern analysis has long been a prominent research topic for social scientists, urban planners, geographers, transportation and telecommunication researchers, but the pertinence of results has thus far been limited by the availability of quality data and suitable data mining techniques. Nowadays, the huge worldwide mobile-phone penetration is increasingly turning the mobile network into a gigantic ubiquitous sensing platform, enabling large-scale analysis and applications.

In recent years, mobile data-based research reaches important conclusions about various aspects of human characteristics, such as human mobility and calling patterns [1–3], virus spreading [4,5], social networks [6–8], content consumption cartography [9], urban and transport planning [10,11], network design [12].

Nevertheless, in such user displacement sampling data, a high uncertainty is related to users movements, since available samples strongly depend on the user-network interaction frequency. For instance, Call Data Records alone do not provide a sufficiently fine granularity and accuracy, exhibiting a vast uncertainty about the periods when the user is not active, i.e., not communicating. This represents an issue for applications or analyses assuming ubiquitous and continuous user-tracking capability.

* Corresponding author. Tel.: +33 1 44 27 88 43.
E-mail addresses: sahar.hoteit@upmc.fr (S. Hoteit), stefano.secci@upmc.fr (S. Secci), stanly@mit.edu (S. Sobolevsky), ratti@mit.edu (C. Ratti), guy.pujolle@upmc.fr (G. Pujolle).

http://dx.doi.org/10.1016/j.comnet.2014.02.011
1389-1286/

Some modeling techniques have been proposed in the literature to predict user movement between two places.

Authors in [13,14] infer the top-k routes traversing a given location sequence within a specified travel time from uncertain trajectories; they use check-in datasets from mobile social applications.[1] Their proposed methods permit to identify the most popular travel routes in a city, but they do not allow constructing time-sensitive routes.

Authors in [15] propose a space–time prism approach, where the prism represents reachable positions as a space–time cube, given user's origin and destination points – i.e., the assumption of knowing the location of a user at one time and then again at another time fits well mobile phone data in which we only know users' position during their communication events – as well as time budget and maximum speed. Spatial prisms so allow evaluating of binary statements, such as the potential of encounter between two moving users. However, the maximum speed cannot be set for all users in general, which limits the model applicability.

Similarly, the authors in [16] propose a probabilistic extension of the space–time approach, applying a non-uniform probability distribution within the space–time prism. A strong assumption made therein is that users move linearly over time. This hypothesis is in a high contrast with the results obtained in [17] that show the tendency of users to stay in the vicinity of their call places. Authors in [17] propose a probabilistic inter-call mobility model, using a finite Gaussian mixture model to determine users' position between their consecutive communication events (call or SMS) using Call Data Records. The model evaluates the density estimation of the spatio-temporal probability distribution of users position between calls, but it does not give an approximation of the fine-grained trajectory between calls. User displacements using GPS traces have been analyzed in [18]; the authors find the displacement behavior show Levy walk properties (i.e., random walk with pause and flight lengths following truncated power laws). While very interesting in order to model inter-contact time distributions and general massive mobility, such random-based approaches cannot give precise approximations between given points on a per-user basis.

The objective of this paper is to assess the pertinence of different conceivable trajectory estimation approaches in terms of error from real available trajectories, via the analysis of real data from the state of Massachusetts. These estimated trajectories are then used to determine cells load in the considered region. By subsampling data-plan smartphone user position samplings, and applying various interpolation methods, we assess the error between real human trajectories and estimated ones. We evaluate simple interpolation method such as linear, nearest and cubic interpolations taking into consideration mobility parameters the network operator may associate with each user.

In particular, we highlight the dependence on the human mobility characteristic, with the user's radius of gyration as user mobility index. Our analysis proves that the linear interpolation shows the best performance for sedentary people (with a small radius of gyration) whereas the cubic one outperforms the others for commuters (having a big radius of gyration). On the other hand, the nearest interpolation presents the smallest error for a set of population movements we identify as "ordinary moves", with long stops. In addition, we experimentally find that interpolations are more accurate when performed within the territory of the user, defined by the user's radius of gyration. Finally we show that the usage of linear and cubic interpolations for modeling human trajectories allows us to determine the hotspot positions with a median error of less than 7%.

The paper is organized as follows. Section 2 presents the dataset used in our study and describes a user ranking with the radius of gyration as mobility pattern parameter. Section 3 presents the different interpolation methods evaluated in this paper. Section 4 summarizes the results of the comparison between the different methods. Section 5 evaluates the load estimation error. Finally, Section 6 draws some perspectives and discusses possible future work.

2. Dataset description

We use a dataset consisting of anonymous cellular phone signaling data collected by Airsage [19], which converts the signaling data into anonymous locations over time for cellular devices. The dataset consists of location estimations – latitude and longitude – for about one million devices from July to October 2009 in the Massachusetts state.

These data are generated each time the device connects to the cellular network including:

- When a call is placed or received (both at the beginning and end of a call).
- When a short message is sent or received.
- When the user connects to the Internet (e.g., to browse the web, or through email synch programs).

The location estimations[2] not only consist of ids of the mobile phone towers that the mobile phones are connected to, but an estimation of their positions generated through triangulation by means of the Airsage's Wireless Signal Extraction technology [19] that aggregates and analyzes wireless signaling data[3] from mobile phones to securely and privately monitor the location and movement of populations in real-time, while guaranteeing acceptable user anonymity and privacy.

In this paper, we select anonymized signaling data of all users during a single day (the observation period is limited to one day because the anonymized user identifiers change for day to another to ensure user privacy).

[1] In recent years, mobile social applications have become so popular that they generate huge volume of social media data, such as check-in records or geo-tagged photos. In a check-in service, users note their locations via a mobile phone to share photos, activities, etc.

[2] Each location measurement is characterized by a position expressed in latitude and longitude and a timestamp.

[3] The location measurements are generated based on signaling events, i.e., when a cellphone communicates with the cellular network's elements through control channel messages.