Journal of Computer and System Sciences 82 (2016) 566–581

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Identifying locations from geospatial trajectories

Alasdair Thomason *, Nathan Griffiths, Victor Sanchez

Department of Computer Science, University of Warwick, Coventry, CV4 7AL, United Kingdom

ARTICLE INFO

Article history:
Received 5 March 2015
Received in revised form 23 June 2015
Accepted 28 October 2015
Available online 30 November 2015

Keywords:
Geospatial
Location extraction
Significant locations
Trajectory
Visits

ABSTRACT

Harnessing the latent knowledge present in geospatial trajectories allows for the potential to revolutionise our understanding of behaviour. This paper discusses one component of such analysis, namely the extraction of significant locations. Specifically, we: (i) present the Gradient-based Visit Extractor (GVE) algorithm capable of extracting periods of low mobility from geospatial data, while maintaining resilience to noise, and addressing the drawbacks of existing techniques, (ii) provide a comprehensive analysis of the properties of these visits and consequent locations, extracted through clustering, and (iii) demonstrate the applicability of GVE to the problem of visit extraction with respect to representative use-cases.

1. Introduction

The recent rise in the prevalence of location-aware hardware has brought with it an increase in the availability of geospatial trajectories, along with the consequent foundation for reasoning about the actions of users that this affords. Fundamentally, such trajectories are sets of spatio-temporal datapoints that relate the whereabouts of an individual, animal or other entity to specific times. The extraction of meaningful locations from geospatial data provides a basis for modelling a user's interactions with their environment, an important part of many location-aware applications. Such applications may include location prediction, typically relying on extracted locations to discretise the set of possible predictive outcomes, or context-aware services such as recommender systems or digital assistants that use location to provide a greater level of personalisation.

This paper considers location extraction as a two-step process that first extracts periods of low mobility from geospatial data, referred to as *visits*, and then clusters these visits together to form *locations*. Specifically, the paper provides a detailed discussion of existing techniques for *visit extraction* and *visit clustering*, and presents the Gradient-based Visit Extraction (GVE) algorithm for the purpose of visit extraction. This algorithm provides resilience to noise and overcomes several drawbacks of previous works: the algorithm does not place a minimum bound on visit duration, has no assumption of evenly spaced observations and operates in real-time without imposing a delay on trajectory points being considered. Additionally, we provide a thorough exploration of the parameter space for GVE through a comprehensive analysis of the properties of visits and locations produced, and demonstrate the applicability of GVE to the task of visit extraction with respect to representative use-cases and a comparison to the current state-of-the-art.

The remainder of this section discusses visit extraction and proposes two representative use-cases for extracted visits and locations. Section 2 then goes on to discuss related work, including approaches to visit extraction and clustering. The GVE algorithm is presented and discussed in Section 3, and discussion of the methodology employed to analyse the algorithm

* Corresponding author.
E-mail addresses: ali@dcs.warwick.ac.uk (A. Thomason), nathan@dcs.warwick.ac.uk (N. Griffiths), vsanchez@dcs.warwick.ac.uk (V. Sanchez).

http://dx.doi.org/10.1016/j.jcss.2015.10.005

and the visits it produces can be found in Section 4. Finally, Section 5 details the results of the analysis and demonstrates the applicability of GVE to the task of visit extraction, with Section 6 providing a conclusion and summary of the findings and contributions made.

1.1. Visit extraction

The extraction of locations meaningful to users is achieved by analysing the datapoints found within trajectories and identifying the regions where the user has spent time. Although a variety of techniques have been used in literature to extract locations from data, they are typically used as a precursor step to performing another activity, such as location prediction, and have not been investigated or evaluated in depth. While it is possible to perform location extraction by using a single clustering algorithm, the process is typically performed in two distinct clustering steps [1–3,7,24,28]. The first step, *visit extraction*, is responsible for partitioning a temporally ordered dataset into periods of low mobility, referred to as either *stops*, *stays* or *visits*, where during each period the user is expected to have remained in one geographic location. For clarity, in this paper, we refer to such periods as *visits*. Visits of no duration (i.e. either the visit consists of a single point, or all points that make up a visit were recorded at the same time instance) are classed as noise since they represent the user travelling and not stationary. The second step in the process, *visit clustering*, summarises the extracted visits and performs clustering to determine which visits belong to the same *location*.

Utilising a two-step approach for location extraction has several advantages, namely:

- Visit extraction can be performed in linear time, summarising vast portions of the dataset, thus reducing the complexity for the clustering step.
- By considering their temporal nature, individual data points that occur when an entity is moving are ignored. In traditional clustering, if several points were to be recorded in close proximity, but on different occasions, for instance along a road, an erroneous location would be identified.
- Extracted visits consist of contiguous points and thus characterise a period of time in which the user remained at the location, providing a basis for modelling historic time spent.

The disadvantages of the two-step process relate to the location clusters extracted at the visit clustering step. In order to reduce the complexity of this stage, extracted visits are typically summarised into a single point (e.g. centroid), and consequently the shape of overall locations extracted are not likely to be represented. Depending upon the goal of location clustering, this could be problematic.

1.2. Use-cases

The uses for extracted locations and visits are varied as they provide a foundation for modelling behaviour. However, for the remainder of this paper we consider two representative examples, which apply equally well to trajectories from various sources. The first use-case, referred to as *accurate visits*, considers the visits as a source of context, aiming to accurately characterise a user's physical location at any point in time. In this case, clustered locations primarily serve to group the visits together to model transitions correctly. The second use-case, *location properties*, is less focused on visits, but rather, considers the accurate identification of the properties (i.e. shape and position) of locations. Accurately identified locations are essential to certain services, such as creating geofences, where the visits are of less importance. It is important to note, however, that although the *location properties* use-case does not strictly require the accurate extraction of visits, the runtime of visit clustering algorithms is severely detrimented as the number of visits increases.

2. Related work

Treating location extraction from trajectories as a two-step process is a technique that has been employed in literature [1–3,7,24,28]. One of the earliest examples is an investigation conducted by Ashbrook and Starner into identifying significant locations, from a dataset of GPS points, with the aim of predicting user movement [3,4]. From the collected data, Ashbrook and Starner observed that the data loggers used did not function well indoors, as a GPS signal was rarely available, and therefore treated periods of missing data as visits. Once extracted, these visits were clustered together using the k-means clustering algorithm, selecting an appropriate value for k by performing the clustering multiple times on different parameters and observing the results. The particular algorithms employed here have their own respective drawbacks, and it is the focus of the remainder of this section to explore these drawbacks along with alternative approaches and related work. Towards understanding existing approaches to visit extraction and clustering, Section 2.1 begins with a discussion on the collection of data over which visit extraction can be performed. An overview of visit extraction techniques follows in Section 2.2, and visit clustering in Section 2.3. Finally, Section 2.4 explores work that has been conducted into enriching extracted locations with additional meaning and context information.