Knowledge-Based Systems 89 (2015) 181–191

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Measuring cell-id trajectory similarity for mobile phone route classification

Mingqi Lv [a,b], Ling Chen [b,*], Yanbin Shen [b], Gencai Chen [b]

[a] College of Computer Science, Zhejiang University of Technology, Hangzhou 310023, PR China
[b] College of Computer Science, Zhejiang University, Hangzhou 310027, PR China

ARTICLE INFO

Article history:
Received 23 October 2014
Received in revised form 6 July 2015
Accepted 8 July 2015
Available online 17 July 2015

Keywords:
Cell-id trajectory
Similarity measure
Trajectory clustering
Route classification
Mobile phone user

ABSTRACT

Route classification based on trajectory data is one of the most essential issues for many location-aware applications. Most existing methods are based on physical locations of the trajectories. However, obtaining physical locations from mobile phones would incur extra cost (e.g. extra energy cost for using GPS). On the other hand, since every active mobile phone is connected to a nearby cell tower, cell-ids (i.e. identifiers of the connected cell towers) could be easily obtained without any additional hardware or network services. In this paper, a cell-id trajectory is a sequence of cell-ids with no regard to physical locations. We address the problem of route classification based on cell-id trajectory data. Specifically, we propose a novel similarity measure which explores the handoff patterns to capture the similarity between cell-id trajectories with no regard to physical locations. Then, based on the cell-id trajectory similarity measure, a clustering algorithm is used to discover potential route patterns from cell-id trajectories, and a nearest-neighbor classification algorithm is used to match current cell-id trajectories to route patterns. The performance of the proposed method is evaluated upon real-world cell-id trajectory dataset. The experimental results showed that our method outperforms state-of-the-art methods on cell-id trajectory clustering and cell-id route classification.

1. Introduction

Trajectory data mining, one of the most essential tasks in location-aware computing, is often implemented by using GPS trajectory data [1,2]. However, GPS has some significant limitations, making it impractical for mobile phone users: First, GPS is extremely power-hungry, leading to a notable reduction in the battery life of today's mobile phones if used continuously [3]. Second, mobile phones require clear line-of-sight to satellite, but they are usually put in a user's pocket or in a bag, so its operation is highly unstable. Besides, its performance is poor in "urban canyons" near high-rise buildings [4], and it usually suffers from prolonged lock-on periods [5]. Third, not all today's mobile phones are equipped with GPS.

On the other hand, every active mobile phone can receive cellular signal from the cellular network, and the identifier of the cell tower to which the mobile phone connects can be easily obtained. Then, a mobile phone user's trajectory can be collected and represented as a sequence of time-stamped cell-ids (i.e. *cell-id trajectory*). Collecting cell-id trajectory data rather than GPS trajectory data has the following advantages: First, detecting cellular signal consumes about one-fifth to one-sixth of the battery power of detecting GPS signal [3]. Second, GPS signal is available as little as 5% of a typical user's day, while cellular signal coverage is available throughout almost all of a day and does not require the lock-on step [6]. Third, every active mobile phone can detect cellular signal, requiring no additional devices. Thus, cell-id trajectory data could be collected at a lower cost and a larger scale than GPS trajectory data.

Cell-id trajectory data can be collected either server-side or client-side. As for server-side collection, cellular network providers routinely collect their users' Call Detail Records (CDRs), which contain information such as the time a voice call was placed and the identifier of the serving cell tower at that time [7]. Although CDRs could be collected at very large scale, they have the following constraints: First, CDRs are sparse in time because they are collected only when a mobile phone engages in a voice call or text message exchange. Second, CDRs are possessed by cellular network providers, and not available to terminal mobile phone users due to privacy concern. Thus, it is unfeasible to analyze individual users' mobility behaviors based on CDRs.

* Corresponding author. Tel.: +86 13606527774.
E-mail addresses: mingqilv@zjut.edu.cn (M. Lv), lingchen@cs.zju.edu.cn (L. Chen), tcsyb@163.com (Y. Shen), chengc@zju.edu.cn (G. Chen).

http://dx.doi.org/10.1016/j.knosys.2015.07.002
0950-7051/

In contrast, client-side collection which means that the mobile phone users are responsible for collecting their own cell-id trajectory data is free of these problems [8]: First, a mobile phone could record its cell-id trajectory data in a periodic manner (e.g. 1 Hz) or when a handoff (i.e. the connected cell tower transfers from one to another) occurs. Second, the cell-id trajectory data are stored locally on the mobile phones, and the mobile phone users could choose whether or not to reveal to third-party service providers or application developers for location-based services. Hence, it is feasible and convenient to analyze users' mobility behaviors based on their client-side collected cell-id trajectory data. However, unlike cellular network providers, terminal mobile phone users usually could not have access to the physical locations of the serving cell towers.[1] Therefore, the major challenge of analyzing client-side collected cell-id trajectory data is how to process trajectory data with no regard to physical locations.

Aiming at this problem, in this paper, we focus on trajectory data mining based on cell-id trajectories with no regard to physical locations for terminal mobile phone users. In particular, a novel similarity measure which could capture the similarity between cell-id trajectories with no regard to physical locations is proposed. Then, based on the cell-id trajectory similarity measure, a clustering algorithm is used to discover potential route patterns from cell-id trajectories, and a nearest-neighbor classification algorithm is used to match current cell-id trajectories to route patterns. We give an application scenario for the proposed method: In the data collection phase, an application running on the background collects the user's daily cell-id trajectory data, without requiring GPS or network connection to reduce battery power and network traffic consumption. In the pattern mining phase, the trajectory clustering method is applied when a certain amount of new data is collected. In the pattern application phase, the route classification could enable various location-aware applications. For example, it could enable the local application to predict the future route of the user in the form of a cell-id sequence. The user could choose to reveal the cell-id sequence to a service provider, and the service provider could then parse the cell-id sequence to pinpoint the physical locations and give corresponding feedbacks to the user (e.g. POI recommendation [9], target advertisement, etc.).

In summary, the primary contributions of this paper are as follows:

(1) We propose a novel similarity measure which explores the handoff patterns to capture the similarity between cell-id trajectories. The similarity measure takes into account the cell tower similarity, the order of cell-ids in the cell-id trajectory and the time duration of the connected cell towers.
(2) We design a trajectory clustering algorithm to discover route patterns and a route classification algorithm to match current cell-id trajectories to route patterns based on the cell-id trajectory similarity measure.
(3) We conducted extensive experiments using real cell-id trajectory data. The results show that our method outperforms state-of-the-art methods on cell-id trajectory clustering and cell-id route classification.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 details our method for cell-id trajectory similarity measure and route classification based on cell-id trajectory clustering. Section 4 presents the experimental results. Section 5 concludes this paper and discusses the future work.

2. Related work

2.1. Cell-id trajectory data mining

There is a growing trend of work attempting to discover mobility patterns from large-scale CDR data. González et al. [10] analyzed CDR data of nearly 100,000 mobile phone users to create models of human mobility patterns. Becker et al. [11] used CDR data to analyze the activity patterns of different groups of people and the mobile phone usage patterns in different parts of a city. Calabrese et al. [12] estimated dynamic city origin–destination flows by using CDR data. Isaacman et al. [13] proposed a clustering and regression method for analyzing CDR data to identify important places. Janecek et al. [14] leveraged anonymized signaling data collected from a cellular network provider to infer vehicle travel times and road congestion. These works analyzed large-scale CDR data in an aggregate level, so the discovered mobility patterns should also be used in an aggregate level (e.g. for urban planning). However, the CDR data are not available to terminal mobile phone users, and the discovered mobility patterns do not contain fine-grained knowledge about individual terminal mobile phone users.

For client-side collected cell-id trajectory data mining, Farrahi and Gatica-Perez [15] used probabilistic topic models to discover location-driven routines (e.g. "going to work late", "going home early", etc.) contained in a massive cell-id trajectory dataset collected from mobile phones. Xiong et al. [16] proposed a novel prediction schema that aims to forecasting a mobile phone user's future locations based on the association behavioral patterns of different mobile phone users. However, these researches focus on analyzing highly abstracted routines (e.g. transition from one place to another), and the detailed routes of the transitions are largely ignored. For route mining based on client-side collected cell-id trajectory data, most existing work requires a cell-id to physical location mapping. For example, Thiagarajan et al. [17] proposed a two-pass HMM (Hidden Markov Model) to map a cell-id trajectory to a sequence of road segments. It requires a training phase based on ground truth GPS locations sampled along with the cell-id trajectory data. Becker et al. [18] explored handoff patterns on given routes, and proposed an algorithm that matches handoff patterns of test drives to routes based on Earth Mover's Distance. It requires that the physical locations of cell towers are known in advance.

2.2. Cell-id trajectory similarity measure

Capturing the similarity between trajectories is crucial for many applications, e.g. trajectory clustering and trajectory classification. Prevalent trajectory clustering methods measure the similarity between trajectories based on the physical locations. For example, Froehlich and Krumm [19] calculated the similarity between GPS trajectories based on a variant of the Hausdorff Distance, and merged similar GPS trajectories to find route patterns. Lee et al. [20] proposed a partition-and-group clustering framework, which partitions trajectories into a set of line segments, and then groups similar line segments into a cluster. Perpendicular distance, parallel distance and angle distance are used to measure the similarity between line segments. Yuan and Raubal [21] extended the traditional Edit Distance by incorporating both spatial and temporal information into the cost functions to measure the similarity between trajectories based on CDRs. Likewise, traditional

[1] The cell tower location information possessed by cellular network providers is not available to the public. Although there are some open databases of cell tower locations created based on war-driving techniques (e.g. OpenCellID), they often have the following constraints: First, the coverage is often limited, especially in rural and developing areas. Second, the mobile phone terminals have to continuously query the server for cell tower locations during client-side collection, resulting in high network traffic. Moreover, mobile phone users are usually used to shut down the network connection for network traffic conservation when not using their phones.