Information Sciences 430–431 (2018) 127–141

Contents lists available at ScienceDirect

Information Sciences

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

A multi-granularity perspective for spatial profiling of mobile apps

Yunhe Feng[a], Zheng Lu[a], Wenjun Zhou[b,*], Qing Cao[a], Xiaolin Li[c]

[a] Electrical Engineering and Computer Science, University of Tennessee-Knoxville, USA
[b] Business Analytics and Statistics, University of Tennessee-Knoxville, USA
[c] School of Management, Nanjing University, China

ARTICLE INFO

Article history:
Received 4 May 2016
Revised 31 May 2017
Accepted 17 November 2017
Available online 21 November 2017

MSC:
00-01
99-00

Keywords:
Mobile apps
Spatial profiling
Kernel density estimation
Application

ABSTRACT

The prevalence of mobile apps has greatly changed people's lives, generating myriads of data and creating new research opportunities. By exploiting mobile apps' data, existing studies have primarily investigated app users' usage patterns. Few studies have focused on profiling the geospatial distributions of individual apps' usage patterns on a large scale. A major challenge for profiling app usage is the heterogeneity and sparsity of individual apps' data. In this study, we propose a multi-level mixture of kernel density estimation (mlKDE) model for robust profiling of the geospatial distribution of any given app. A major advantage of this model is that it can leverage aggregate information from groups of related apps, and adaptively train the weights for different groups. Using a real-world, large-scale app usage dataset covering more than one thousand apps available on the market, we demonstrate that our model can effectively characterize the distributions of app usage in the real world, and has considerable advantages when compared to the baseline models reported in the literature.

1. Introduction

In recent years, mobile apps have become an integral part of the smartphone user experience [1–3], greatly changing people's lifestyles. For example, more and more people are sending and receiving messages through instant messaging apps (e.g., `WhatsApp`) or social networking apps (e.g., `Twitter`), rather than through Short Message Services (SMS, also known as "text messages") for their personal communications. In particular, it was reported in 2014 that the data traffic generated by traditional voice calls accounts for less than 10% of all mobile data traffic [4]. Such changes coincided with the large and rapidly growing number of smartphone users, who more and more frequently use data-intensive services. Recent advances in technology and developments in infrastructure [4] have played a major role in enabling those data-intensive services. In order to meet the data traffic demands, mobile network operators are constantly upgrading network capacities. App usage data with geospatial information can inform the actions that mobile network operators take in a number of critical operational decisions. Therefore, the abundance of mobile apps' data, which accounts for a large proportion of base stations' traffic load, is ripe for exploration and analysis.

Existing studies explored the geospatial data generated by mobile apps from various perspectives. Many studies exploited the geospatial information captured by mobile devices to track human mobility [5,6], ignoring the data volumes. Other

* Corresponding author.
E-mail address: wzhou4@utk.edu (W. Zhou).

https://doi.org/10.1016/j.ins.2017.11.040

studies identified diverse usage behaviors of smartphone apps and showed the geospatial distribution of aggregated usage data of apps in the same genres (e.g., social networking, game, email) [7–9] instead of individual apps. Antithetically, we assume the perspective of service providers (i.e., app developers and mobile network operators) to focus on the geospatial characteristics of individual apps. The question we attempt to address in the current study is that, given records of app usages, can we profile the geospatial distribution of data volumes caused by the usage of each app in a robust manner? This question is essential for solving a number of real-world resource scheduling and optimization problems, such as bandwidth allocation, load balancing, and network caching.

Despite the exciting possibilities, profiling the distribution of data volumes per app is a challenging problem. First, different apps exhibit a high degree of heterogeneity in functions and characteristics. For example, navigation apps are mainly used on highways or local roads; flashlight apps become active after nightfall; and instant messaging apps are typically used continuously. Second, the profiles have to be dynamically updated over time. The same app has different geospatial distributions at different times. This could be caused by many factors, such as fluctuations in the number of app users, movements of app users and usages of the app's provided functions. Third, the mobile data traffic of an individual app can be quite sparse during short intervals, especially for a large number of less popular apps on the long tail. It is hard to obtain useable distributional patterns by applying existing models to an individual app's data.

To address these challenges, we propose a multi-level mixture model based on kernel densities, where each level is represented by a single kernel density estimation. In particular, we consider the data volumes of mobile apps at three different granularity levels. At the individual app level, we consider an individual app's data volume, which represents the app's resource demand with all users aggregated. At the group level, we aggregate data volumes from a group of related apps to represent certain properties of the app (e.g., based on categories or purposes). At the population level, we aggregate the data volumes from all apps, which represent the overall activeness of all users. After estimating the spatial density of data volumes at each level, we adaptively train the kernel weights, which not only profile the geospatial distribution of an app more robustly, but also provide interpretable insights about the app and its peers, at multiple levels. Finally, the three kernel estimations are combined together, based on their respective weights, to build the multi-level mixture model.

Our contributions can be summarized as follows. First, we provide a multi-granularity solution of modeling individual mobile apps' data to overcome the sparsity of individual level data. Our multi-level mixture model based on kernel densities achieved a high fitting accuracy. Second, our model can handle the heterogeneity of different apps as well as the dynamic profiling for the same app over time. For both different apps and the same app at different time points, our model adaptively assigns weights for the three levels of data. The weights can also be used to analyze the uniqueness of apps. Third, our model is general enough to be applied to other domains with similar formats.

The rest of this paper is organized as follows. In Section 2, we introduce the application scenario and problem formulation. Section 3 discusses our multi-level mixture of kernel density estimation model in detail. In Section 4, various experimental results are presented. In Section 5, we briefly survey the related work regarding the analysis of mobile app usages, studies of smartphone data traffic and spatial modeling. Section 6 presents the applications of our model and its limitations. Finally, Section 7 concludes the paper.

2. Data scenario and problem formulation

In this section, we describe the application scenario, introduce our data, and formulate the problem. We first discuss a general scenario where our model could be applied, then describe the spatial feature of our data, and finally provide a descriptive study about the app categories used in our data.

2.1. The scenario

In this study, we focus on the geospatial distribution of data volumes due to mobile app usages. As the cell towers provide services to nearby users, it is possible to measure various activities through these towers, such as data volumes downloaded by individual apps, and estimate the overall distribution of such measures over the region. Formally, given records that can be written as

$$\langle \texttt{TowerID}, \texttt{TimeStamp}, \texttt{AppID}, \texttt{Volumes} \rangle, \tag{1}$$

we are interested in finding the density distribution of download volumes attributable to each given app over the region. The objective is to identify estimates that are smooth, efficient, robust, and interpretable.

It is straight-forward to slice time into a preferred level of granularity, such as per minute, or per hour. When slicing time into refined, short intervals, the pass-through data volumes will be so low that density estimates at the app level will become unstable. It is hard to summarize the distribution smoothly when the data change abruptly. One possible source of information we could leverage is the hierarchy of app categories. Information at the population level (i.e., all apps) may indicate the overall activeness of users at that time of day. Information at the group level (i.e., all apps in the corresponding category) may indicate the users' interests.

Investigating geospatial distribution of apps' data is important for several reasons. First, understanding geospatial distribution of apps' data can provide valuable insights on the need to increase network capacity based on data traffic demands. Monitoring and analyzing apps' geospatial distribution can assist mobile network operators in improving their customers'