Characterizing a User from Large-scale Smartphone-sensed Data

Sha Zhao
Zhejiang University
Hangzhou, Zhejiang, China
szhao@zju.edu.cn

Runhe Huang
Hosei University
Tokyo, Japan
rhuang@hosei.ac.jp

Yifan Zhao
Zhejiang University
Hangzhou, Zhejiang, China
yifanzhao@zju.edu.cn

Shijian Li
Zhejiang University
Hangzhou, Zhejiang, China
shijianli@zju.edu.cn

Zhe Zhao
Zhiling Luo
Zhejiang University
Hangzhou, Zhejiang, China
zhezhao@zju.edu.cn
luozhiling@zju.edu.cn

Gang Pan
Zhejiang University
Hangzhou, Zhejiang, China
gpan@zju.edu.cn

UbiComp/ISWC'17 Adjunct, September 11–15, 2017, Maui, HI, USA

ACM ISBN 978-1-4503-5190-4/17/09.
https://doi.org/10.1145/3123024.3124437

Abstract
Device analyzer can provide a large-scale dataset that captures real-world usage of smart phones [1]. Detailed usage records in smart phones, conveying a partial life log, are important for a deep scientific understanding of human characteristics. In this study, we proposed a feature-based labeling method to characterize users. Eight features from three aspects, i.e., daily mobility, user daily schedule, and social ability, are designed within a time window. Further, we analyze the features' correlation and variation over time. With the features, each user can be attached with a few semantic labels to demonstrate his/her characteristics. This work is a promising step towards drawing "portraits" for users using mobile phone data.

Introduction
Nowadays, roughly 2 billion people worldwide have been covered by smart phones, which are becoming people's essential belongings. Smartphones are equipped with a growing set of powerful sensors, such as accelerometer, GPS, proximity sensor, and camera, which are enabling capture of users' behavior data. Some applications are developed to collect the captured data, such as LDCC for Mobile Data Challenge [2]. A continuous collection of mobile phone data for a long duration will bring detailed records about users' behaviors, such as personal activities, movements, phone usage, and living habits. The detailed records in

mobile phones as a partial life log are important for understanding human characteristics and social phenomena.

Wagner et al developed Device Analyzer [1], a robust data collection tool which is able to reliably collect information on Android smartphone usage from an open community of contributors. The ubiquitous and large-scale mobile phone data provides a good opportunity for researchers to characterize and understand real-life phenomena, including individual traits, as well as human mobility, social network, phone usage patterns [3,4,5].

In this paper, we aim at describing a user with his/her phone-sensed daily data, and try to draw a "portrait" for the user.

The approach

Daily mobility, daily schedule and social ability, as partial life logs, are important cues for drawing the user's "portrait". From the three aspects, eight features are constructed as representatives to characterize a user, which are relatively easy to extract from phone sensed data. We put them into a feature vector within a time window to represent a user. Feature based labels are defined and attached to describe a user's characteristics.

Features extraction

We define eight features for describing users. Table 1 shows the features and their data sources.

1) Daily mobility

A smartphone will observe different sets of WiFi access points (APs) when it is carried by a moving person. A WiFi scanlist of a smartphone is a set of the WiFi APs whose WiFi signal can reach the smartphone. It can roughly indicate the physical location of the phone. Analysis of WiFi scanlists of one's phone can reveal his/her mobility information. Here, we only take one's stay places into consideration, where the user stay for a while, and ignore those places where he/she just passes by.

Table 1. Eight features and their data sources

	Features	Data sources
Daily mobility	***stay places***: how many places a user has stayed for a few minutes	WiFi scan lists
	regularity of visits: how regularly a user moves in a day	
Daily schedule	***get-up time***	Overnight battery charge
	bed time	
	nocturnal phone use: how actively a user use the phone during 12:00 am-6:00 am (6 hours after midnight)	Screen on/ off
Social ability	***social circle***: how many contacts involved in a user's SMSs and call records	SMSs and call records
	contact concentration: how concentratedly a user communicates with intimate contacts	
	contact frequency: how frequently a user uses SMSs and calls	

In order to remove intermittent APs which occur during transition, a WiFi AP will be filtered if it keeps live less than 10 minutes. The UIM clustering algorithm [6] is employed to cluster Wifi scan lists into a set of stay places. Scan lists in the same cluster represent a stay place. Combining the scan time of each Wifi scanlist,

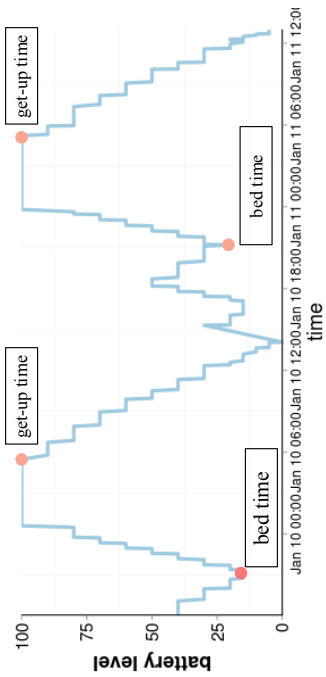


Figure 1. An example of get-up time and bed time

individual mobility record can be created [3]. Formally, an individual's mobility record r is represented by Eq. (1).

$$r = \{(p_1, t_1), (p_2, t_2), \cdots, (p_i, t_i)\} \quad (1)$$

where p_i is the ith stay place, t_i is its scan time.

Based on mobility records, the following features are extracted:

- *stay places.* It indicates how many stay places a person has.
- *regularity of visits.* It is measured by the proportion of the total duration of a user staying at stay places and the total duration of the whole Wifi logs.

2) Daily schedule

For daily schedule, we focus on: when a user gets up or goes to bed, how actively the user uses phone during 12:00 am-06:00 am after midnight.

- *get-up time / bed time*. The charging cycle of smartphone batteries can be identified by the sensed data of the Device Analyzer. Here we assume that the user starts to charge at the bed time and stop charging when he/she gets up, if a charging activity happens at night for longer than 4 hours. Detailedly, *get-up time* and *bed time* can be defined if a charge cycle satisfies the following conditions: 1) the battery level is up to 100 when the charge ends; 2) the charge cycle is more than 4 hours; 3) the charge begins within the night time from 8:00pm to 5:00am; 4) the charge ends within 4:00am to 1:00pm of the second day. Figure 1 shows an example of *get-up time* and *bed time*.
- *nocturnal phone use.* It measures how actively a user uses the phone during the 6 hours after midnight. We define the interval between the screen turning on and the screen turning off as a *session*. According to the statistics in [7], Americans spend about 2.7 hours in total per day on their phones. The duration of some sessions in the dataset is unbelievably long, e.g. 6 hours and more. It seems impossible for a person to continuously use smartphone for so long time. It most likely is caused by the data collection APP. In this paper, the sessions which last for more than 2 hours are filtered. The total duration of all the sessions are computed. The proportion of the total duration of sessions and 6 hours (12:00 am – 06:00 am) is taken to extract this feature.

3) Social ability

Mobile phones could be considered as the most important tool to connect people in the world. Call detail records (CDRs) provide much information on a person's social behaviors, for example, how frequently a user uses a phone to contact with others, how many contact persons are involved in a person's CDRs. Here, we focus on using information about a user's SMSs and call records to measure his/her social ability. Three features are defined.

- *social circle*. It is measured by the number of all the contact persons involved in a user's SMSs and call records.
- *contact concentration*. It measures how concentratedly a user communicates with his/her intimate contacts. A user's contacts are ranked according to the number of SMSs and call records. The top 20% contact persons are defined as *intimate contacts*. A user' *contact concentration* is measured by the proportion of the number of SMSs and call records with *intimate contacts* and his/her total SMSs and call records.

Table 2. Sixteen semantic labels for two ends of eight feature distribution.

Features	Labels (higher value)	Labels (lower value)
get-up time	Late-riser	Early-bird
bed time	Night-bird	Early-to-bed
nocturnal phone use	Frequent-nocturnal-use	Rare-nocturnal-use
social circle	Broad-circle	Narrow-circle
contact concentration	Concentrated -contact	Spreadout-contact
contact frequency	Frequent-contact	Rare-contact
stay places	More-places	Fewer-places
regularity	Regular-visit	Irregular-visit

- *contact frequency*. It measures how frequently a user use SMSs and calls. It is computed by the number of all the SMSs and call records divided by the value of *social circle*.

User representation

In order to describe the temporality of a user's features, time window is introduced and used to describe a user's features in a given period. Intuitively, we can represent each user as a vector with the eight features above-mentioned within a time window. Formally, given a time window T, each user is defined in Eq. (2).

$$u_T = (f_1^T, f_2^T, \cdots, f_8^T) \quad (2)$$

where f_i^T is a feature in the time window T.

In order to better characterize a user, given that the value of the eight features varies in a wide range, each feature is normalized to a standard normal distribution by Eq. (3).

$$x' = \frac{x-\mu}{\sigma} \quad (3)$$

where μ is the average value of this feature of all the users, and σ is the standard deviation.

User labeling

The characteristic of a person is a special quality or trait that makes him/her different from others. Thus, one's characteristics usually are not very close to the average of people. We assume that users' features follow the Gaussian distribution. For a user, we focus on those features far away from mean of the feature with more than one standard deviation (std), i.e. lying outside the interval of (mean $\pm$std), which may be good recognizable traits for the user. For each feature, a pair of semantic labels is used for two ends of the feature distribution, lying outside of the interval of (mean $\pm$std). Thus, sixteen semantic labels are defined to describe users, shown in Table 2.

Experiments and analyses

We conduct experiments on feature extraction, user representation, and user labeling using the Device Analyzer data. In the experiments, each mobile phone involved in the dataset is assumed to be used by the same individual during data collection.

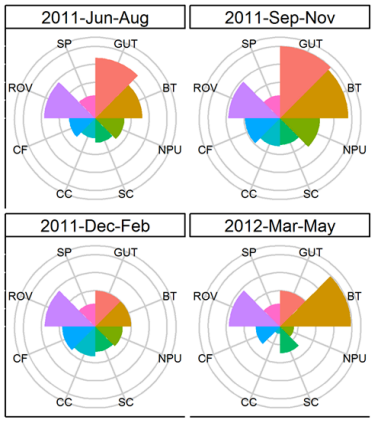
Data preprocessing

Based on our observation, in data collection process a user's data could be missing for some time, when there is no data collected. This time is considered as invalid duration, which may have impact on our experiment results. It is necessary for us to detect and filter the invalid duration when we take samples for our experiments. Battery state[1], such as level, scale and temperature, is collected by timing sampling mechanism, i.e., battery state collection depends on devices rather than user behaviors. A duration when battery state is not collected is considered as an *invalid duration*. According to the collection of battery state, we filtered invalid durations for our experiments. A month is considered as *valid month*, in which there are less than 10 days with *invalid duration*. 497 users were selected for our experiments, whose duration is more than six consecutive *valid months*.

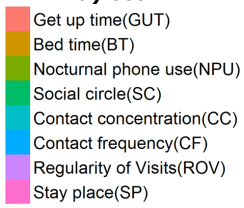
Correlation between features

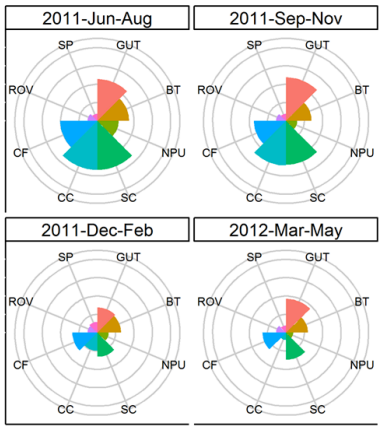
Figure 2 shows the correlation matrix of eight features. It is found that three pairs of features have correlation (r >0.3). 1) *bed time* has a correlation with *nocturnal phone use* (r =0.460). As users goes for sleep later, they spend more time on phone usage. Before users

[1] https://deviceanalyzer.cl.cam.ac.uk/keyValuePairs.htm



a) User1

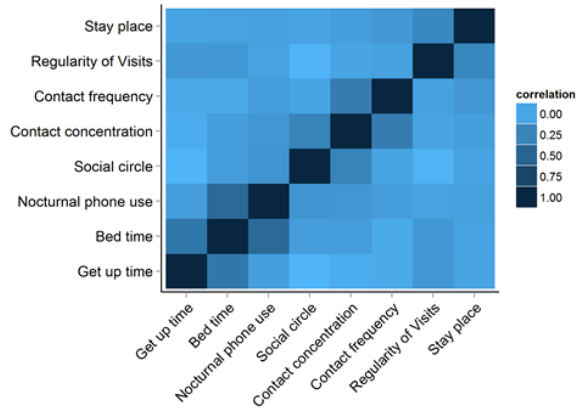




b) User2

Figure 3. Two examples for three months-based user representation

get to sleep, they may spend more time to use phones for entertainment, reading when they lie in bed. 2) *get-up time* has a correlation with *bed time* (r =0.350). The correlation shows that users who get up later tend to go to bed later for enough sleep time. 3) *contact frequency* has a correlation with *contact concentration* (r =0.306). The correlation shows that a user with a broader *social circle* tends to keep in touch with a smaller intimate circle. This may reveal the total of social time is limited.



Figrue 2. Correlation matrix between eight features

User representation within a time window

We represented each user as a feature vector within a time window using 8 extracted features. Here, the value of time window T was set as three months. Two users were taken for example, shown in Figure 3. A user is represented by eight features within four different time windows. For each radar map, bigger the related feature value, larger the area of a sector. It can be seen that, User1 gets up later in the first two time windows while he/she gets up earlier in the other two time windows. The regularity of visits keeps roughly stable in the four time windows. The value of the three features about social ability is always lower, which shows that User1 has few social activities related with SMSs and call records. User2's three features about social ability vary obviously as time goes. His/ her *get-up time* and *bed time* also changes within different time windows.

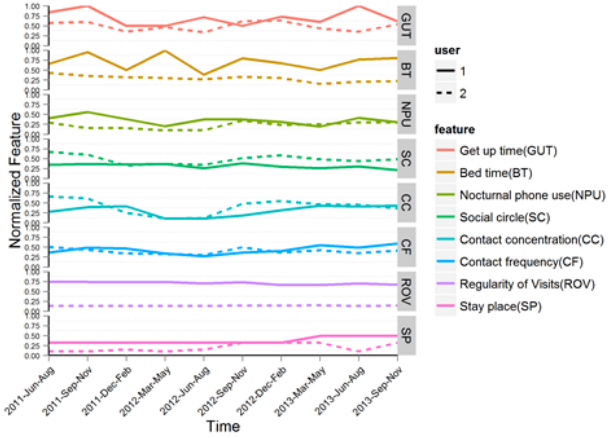
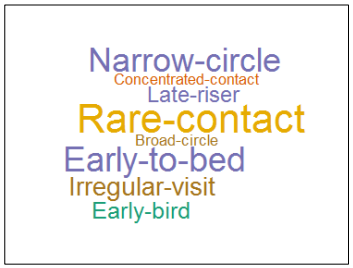


Figure 4. The variation of diffentent features for User1 and User2

In Figure 4, the variation of different features over nine time windows is shown for User1and User2. The trends of two users' curves about *get-up time* and *bed time* is roughly the same, which illustrates the stronger correlation between the two features. Similarly, the strongger correlation between *nocturnal phone use* and *bed time* is also illustrated by the roughly same trends of the curves. The *regularity of visits* and *staly places* of two users keep roughly stable in the nine time windows, which reflects individual mobility pattern to some extent. For two users, the trends of the curves about

a) Labels for User3



b) Labels for User4

c) Labels for User5

Figure 5. Three examples of user labeling

social circle, contact concentration and contatct frequency is about the same.

User labeling

For user labeling, the value of time window T was set as one week. Each user being represented within a time window of one week can reflect the variation of eight features in a more fined-granularity, which is helpful for attaching proper labels to each user. A user is attached a semantic label if one feature lies outside of the interval of (mean $\pm$std). The frequency for each feature when its value lies outside of the interval of (mean $\pm$ std) is considered as the weight of labels. For each label, bigger the frequency value, larger the label is. We took three users for example to show the results.

From Figure 5, it can be seen that each user has a few large labels, which illustrate the user's significant characteristics. For User3, the larger labels show that he/she visits places regularly and rarely contacts with others. Interestingly, he/she does not have obvious bias on *contact concentration* because there is little difference between the labels of Concentrated-contact and Spreadout-contact.

For User4, he/she has three labels significantly bigger than others: Narrow-circle, Rare-contact and Early-to-bed. It reveals that he/she has fewer phone contacts involved in SMSs and call records, contact with others rarely, and goes to bed early during most weeks. This user also has a larger label of Irregular-visit, which shows that he/she visits more irregularly. For *get-up time*, sometimes he/she gets up earlier while sometimes he/she gets up later.

User5 has only one label significantly bigger than others: Regular-visit, which means higher regularity of visits during most weeks. He also has a smaller label of Irregular-visit, which illustrates that he/she moves irregularly in some weeks He/she has a narrow social circle (Narrow-circle) and prefers to keep in touch with in a smaller intimate circle (Concentrated-contact).

Conclusions

Based on the large-scale dataset collected by Device Analyzer, we present a framework to describe users using feature-based semantic labels. The semantic labels make up the user's "portrait". Eight features are defined from three aspects: daily mobility, daily schedule and social ability. With the extracted features, each user could be represented as a feature vector. The experiments were carried out and the results are analysed. The framework is easy to extend for more features. This work is a promising step towards drawing users' characteristics using mobile phone data.

References

[1] Wagner, D. T., Rice, A., et al. Device Analyzer: Understanding smartphone usage. MOBIQUITOUS 2013.

[2] Kiukkonen, N., Blom, J., et al. Towards rich mobile phone datasets: Lausanne data collection campaign. Proc. ICPS, Berlin(2010).

[3] Zhao, S., Zhao, Z., et al. Discovering People's Life Patterns from Anonymized WiFi Scanlists. Proc. UIC 2014, pp. 276-283.

[4] Zhao, S., Pan, G., et al. Mining user attributes using large-scale app lists of smartphones. IEEE Systems Journal 11, no. 1 (2017): 315-323.

[5] Zhao, S., Ramos, J., et al. Discovering different kinds of smartphone users through their application usage behaviors. Proc. UbiComp 2016, pp. 498-509.

[6] Vu, L., Do, Q., et al. Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/bluetooth trace. PerCom 2011, IEEE Press (2011), 54-62.

[7] Shin, C. and Dey, A. K. Automatically detecting problematic use of smartphones. UbiComp 2013, ACM Press (2013), 335-344.