User profiling from their use of smartphone applications: A survey

Article *in* Pervasive and Mobile Computing · July 2019

DOI: 10.1016/j.pmcj.2019.101052

Sha Zhao
Zhejiang University
23 PUBLICATIONS **124** CITATIONS
SEE PROFILE

Shijian Li
Zhejiang University
101 PUBLICATIONS **3,116** CITATIONS
SEE PROFILE

Julian Andres Ramos Rojas
Carnegie Mellon University
17 PUBLICATIONS **589** CITATIONS
SEE PROFILE

Zhiling Luo
Alibaba Group
34 PUBLICATIONS **127** CITATIONS
SEE PROFILE

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

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

Review

User profiling from their use of smartphone applications: A survey

Sha Zhao [a,b], Shijian Li [a], Julian Ramos [b], Zhiling Luo [a], Ziwen Jiang [a], Anind K. Dey [b], Gang Pan [c,a,*]

[a] Department of Computer Science, Zhejiang University, Zheda Road No. 38, Hangzhou, China
[b] Human–Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, USA
[c] State Key Lab of CAD and CG, Zhejiang University, Zheda Road No. 38, Hangzhou, China

ARTICLE INFO

Article history:
Received 5 January 2019
Received in revised form 8 July 2019
Accepted 11 July 2019
Available online xxxx

Keywords:
User profiling
Smartphone applications
Mobile sensing

ABSTRACT

The number and popularity of smartphone applications is rising dramatically. Users install and use applications depending on their needs and interests. Applications on smartphones convey lots of personal information, providing us a new lens to well profile users. In this paper, we first describe application information for user profiling. Second, we analyze what types of user information can be profiled from smartphone applications. Then, we overview the previous work and summarize the state-of-the-art methods for profiling users from application data. We also describe implications from different perspectives. Finally, the challenges in profiling users from smartphone applications are discussed.

Contents

* Corresponding author.
E-mail address: gpan@zju.edu.cn (G. Pan).

https://doi.org/10.1016/j.pmcj.2019.101052
1574-1192/© 2019 Elsevier B.V. All rights reserved.

1. Introduction

A smartphone is regarded as an essential component of modern life [1]. Nowadays, it is estimated that there are roughly 2 billion smartphone users in the market. Smartphones serve a wide variety of functions, and users can exploit mobile applications to achieve many imaginable purposes [2–4]. The mobile application market has seen explosive growth in recent years, with Apple's app store having about 1.8 million applications and Google's Android market also having around 2.1 million applications as of the first quarter of 2019.[1] Abundant applications provide useful services in many aspects of modern life. Easy to download and often free, applications can be fun and convenient for playing games, getting turn-by-turn directions, and accessing news, books, weather, and more [5].

Applications (abbr. apps) on smartphones can be considered as the entry point to access everyday life services such as communication, shopping, navigation, and entertainment. Smartphone users install and use apps depending on their needs, interests, habits, etc. Since a smartphone is linked to an individual, the apps on it achieve greater personalization. Thus, apps on smartphones can effectively convey lots of personal information. This has the potential to provide us with a new lens to better profile users. User profiling from smartphone apps is a process of analyzing smartphone app data, exploring how apps are correlated with user personal information, and extracting key features to describe or infer users' characteristics. It is a type of context-aware sensing where data used for context awareness is from smartphone apps, according to the concept of context introduced in [6] that the interaction between a user and application can also be considered as context. Many recent studies have sought to profile users based on smartphones' apps in different perspectives, such as inferring demographics and personality [7–17] and detecting users' interests, preferences and habits [18–27].

Profiling smartphone users well is key for improving mobile user experiences. To be specific, it can help us to improve mobile devices, applications, and services. From the view point of advertisement and service providers, they could provide users targeted advertisements, personalized recommendations and smart services, based on users' profiles. From the app developer and designer viewpoint, it could be useful to know users' habits, preferences or interests so that apps could support better recommendations and adaptations to improve the user experience. From the viewpoint of smartphone manufacturers, understanding users' needs could be used to modify specific phone features to make them more valuable for users. From the viewpoint of mobile carriers that sell smartphones, knowing which apps users are interested in, could be used to customize the pre-installed apps in comparison to the current approach of pre-loading phones with a set of default apps. Moreover, it could help user understand themselves objectively so as to improve life quality.

We focus this survey on profiling users from apps on smartphones, and review the related studies in the literature. It is worth noting that this survey focuses on the use of common apps installed on smartphones. Some context-aware apps developed for specific tasks (e.g., BeWell [18]) are out of the scope. In other words, this survey concentrates on the studies that profile users from their use of ordinary apps in daily life. Our survey aims to answer the following questions: what app information can be used to profile users? what types of user information can be profiled from smartphone apps? how to profile users, i.e., the methods? what are the implications? what challenges are there in profiling users from smartphone apps, as well as the suggestions to address them?

This survey is organized as follows: Section 2 describes app information for profiling. Section 3 describes what types of user information can be profiled from smartphone apps. Then, Section 4 summarizes the methods that are used for profiling users, and Section 5 discusses the implications. Section 6 discusses the current challenges and opportunities in this domain. Section 7 concludes this survey.

[1] https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/.

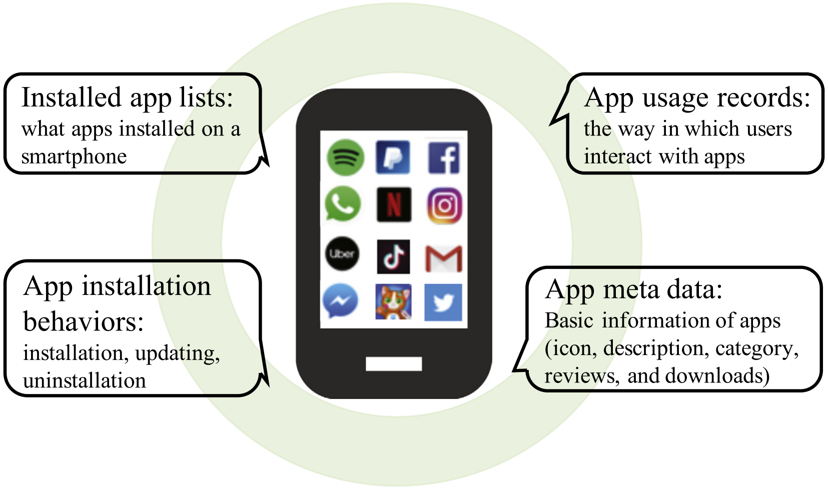

Installed app lists:
what apps installed on a smartphone
App usage records:
the way in which users interact with apps
App installation behaviors:
installation, updating, uninstallation
App meta data:
Basic information of apps (icon, description, category, reviews, and downloads)

Fig. 1. App information for profiling smartphone users.

Table 1
Sample of installed app lists in the dataset used in [19].

User ID	Installation package	App name
004ac0891bb123d89e80ff1182699f2d	com.tencent.minihd.qq	QQ
004ac0891bb123d89e80ff1182699f2d	com.besttone.FortuneStreet.plugin	股票财经
004ac0891bb123d89e80ff1182699f2d	buke.besttone.caipiao.plugin	号百彩票
008f23e4ec85676e5d823abebb2043f1	com.autonavi.xmgd.navigator	高德导航
008f23e4ec85676e5d823abebb2043f1	com.cubic.autohome	汽车之家
008f23e4ec85676e5d823abebb2043f1	com.tencent.qqmusic	QQ 音乐

2. App information for profiling

There is a vast volume of app data that are collected, such as what apps installed on smartphones, how apps used, and the basic information of app. We roughly divide the information regarding apps into four classes: installed app lists, app usage records, app installation behaviors, and app metadata, shown in Fig. 1. We will introduce each type of app data, and discuss some limitations.

2.1. Installed app lists

An installed app list refers to what apps installed on a smartphone. Users decide to install apps depending on their needs, preferences and tastes. For example, someone regularly having workout is likely to install apps that provide services for fitness, and someone who likes playing games may install game apps on his/her smartphone. Thus intuitively, the apps that a user has installed are potentially good indicators of their needs and interests. Moreover, the list of apps installed on a smartphone is relatively accessible. Although the permission for an app to access personal data, such as contacts and location, must be requested from a user at the time of app installation, the list of apps installed on a user's smartphone can be obtained without his/her permission through any app installed on Android devices [28]. Some advertisement tracking libraries have reportedly embedded this feature to collect lists of installed apps [28]. Taken together, apps installed on a smartphone provide a good opportunity for profiling users.

A dataset of installed app lists usually consists of installation records, each of which contains an anonymized user ID, an installation package that can be used to identify an app, and the app name. Table 1 shows a sample of installation records used in [19].

Although many previous studies have shown that installed app lists reflect users' needs or tastes to a certain degree, there still exists a major challenge. It is, whether one user has installed an app may be a weak indicator of whether he/she actually needs the app [10,19,29,29–32]. He/she may simply want to try the app out, and may never use it again or may have uninstalled it. According to the statistics in [33], only 10% of apps were used 80% of the time, suggesting that a lot of apps are downloaded but not used regularly. To alleviate this shortcoming in installed app lists, it is worthwhile to take one user's daily app usage into account, such as usage time and the frequency of app usage, making it more accurate in profiling user characteristics. Besides, although there is a flag that can be used to separate pre-installed apps out, the

Please cite this article as: S. Zhao, S. Li, J. Ramos et al., User profiling from their use of smartphone applications: A survey, Pervasive and Mobile Computing (2019) 101052, https://doi.org/10.1016/j.pmcj.2019.101052.

information whether one app is pre-installed or not is missed in some datasets, such as the dataset used in [19]. The authors manually removed the apps that probably are pre-installed based on their knowledge [19], which could result in deviation from the groundtruth labels.

2.2. App usage records

App usage records report the way in which users interact with apps, such as when an app is launched or killed, how long and how often it is used. Compared with installed app lists, there is temporal context, i.e., time information, accompanied with the app usage records, recording when one user uses which apps. As mentioned above, different individuals might have different interests, which makes it natural for the apps installed on smartphones of different users to be distinct. Even for the same app, its usage can be different across users like frequency and intensity. These differences in app usage make it possible to infer user characteristics.

There are two typical ways to collect app usage records:

- *Event-driven collection* collects an app usage record when an event happens such as user actions (e.g., launching an app, inputting, and clicking), messages received (e.g., incoming calls, and notification), or requests to mobile network. For example, app events were collected once an app was launched [34]. In [35], each time one user accesses an app, the client software captures the event and stores it together with the timestamp. In [36,37], the data was collected when data requests to the mobile network happened.
- *Time sampling scheme* collects app usage records at systematical time intervals, such as sampling hourly. For instance, the recent app task lists used in [13] were collected about every hour. Jesdabodi et al. collected usage sessions by detecting the home screen pause below 2 min [38].

The app usage records used in different studies contain different content. For example, the app usage records collected by Lausanne data collection campaign [34] contained app events on Nokia platforms, such as start, close, foreground and view, which were public in MDC (mobile data challenge)[2] that provided access for research-only purposes to the MDC dataset. The dataset used in [13] contained lists of recent apps used on Android smartphones, ordered from most recent to oldest. Qin et al. [8] analyzed the amount of entries when apps fetch resources from the internet. Yu et al. [36] analyzed app events of start and end, identified from the cellular data accessing traces obtained by Deep Packet Inspection. In some studies, the app usage records were collected by specific apps, such as Device Analyzer [39] and AppSensor [40], while the app usage records used in some other studies were extracted from the raw data collected by the third party, such as the Deep Package Inspector in [36].

There are some limitations in the existing app usage datasets yet. For example, the dataset used in [13] consisted of recently used app lists that were collected once every hour. The low sampling rate could miss information about app usage, and depending on the recently used app lists ignore app usage details, such as how long each app is used, how often it is used, and in which order the lists change. Some understanding processes are longitudinal in nature, but, the duration of the used dataset is not long enough to understand the processes. For instance, it was reported that 4 weeks is a short time to understand how smartphone apps affect users' daily life [41]. A few studies also reported that the used datasets miss the usage records of some apps, such as the app running in background [30], apps used offline or via WiFi network [37,42–44], and SMS [41].

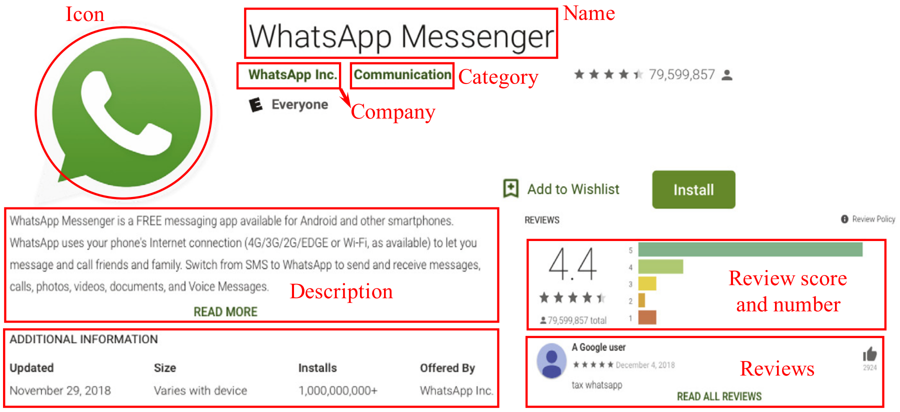
2.3. App installation behaviors

App installation behaviors include the installation, updating, and uninstallation of apps and the corresponding time stamp, i.e., what app is installed/updated/uninstalled and when. As installation behaviors of an app are usually made manually, especially the installation and uninstallation, they can implicitly reflect users' preferences towards the app. It further indicates users' needs or preferences, since users install or uninstall apps depending on their needs or preferences. In particular, the downloading (installation) and updating an app could be aggregated to analyze, because they both reflect the users' interest towards this app.

Each installation behavior record contains the anonymized user ID, time stamp when the behavior happens, app package that can be used to identify an app, and the behavior type, i.e., installation/ updating/ uninstallation. The dataset of app installation behaviors used in [20,45] is a major one in the literature, covering about 0.8 million Android users over one month. It was collected by Wandoujia, a free Android app marketplace in China. The logs of installation behaviors were all automatically recorded by the Wandoujia management app, by which users can manage their apps on Android smartphones.

Despite the fact that users usually install and uninstall apps manually from app stores, many users do not frequently update their apps or even let the OS automatically update apps. It was found that a large number of users (at least in China) do not update their apps from app stores [42]. Such cases could result in some deviations in analysis results. This shortcoming may be alleviated by aggregating app usage records together with app installation behaviors. There are also some limitations in the existing datasets. A few datasets miss the installation behaviors of some apps, for example, only

[2] https://www.idiap.ch/project/mdc/.



WhatsApp Messenger is a FREE messaging app available for Android and other smartphones. WhatsApp uses your phone's Internet connection (4G/3G/2G/EDGE or Wi-Fi, as available) to let you message and call friends and family. Switch from SMS to WhatsApp to send and receive messages, calls, photos, videos, documents, and Voice Messages.

ADDITIONAL INFORMATION

Updated	Size	Installs	Offered By
November 29, 2018	Varies with device	1,000,000,000+	WhatsApp Inc.

Fig. 2. An example (WhatsApp) of app metadata.

Table 2
Example app datasets.

Type	Reference	Platform	#User	Duration	Description	Public
Installed app lists	Seneviratne et al. [48]	Android	231	–	Installed apps	N
	Wagner et al. [39]	Android	17,000	–	Installed apps	Y
App usage records	Kiukkonen et al. [34]	Nokia	200	1 year	App start and close events	Y
	Zhao et al. [13]	Android	106,672	1 month	Recent app task lists	N
App installation behaviors	Li et al. [20]	Android	0.8 M	1 month	Installation, update	N
	Frey et al. [29]	Android	2008	1 month	Installation logs	N
App description	Seneviratne et al. [7]	Android	218	–	App description	N
App categories	Zhao et al. [13]	Android	106,672	–	App categories	N

the apps whose installation behaviors were conducted through Wandoujia app (an Android app store in China) were analyzed in [20,42,43,45] and the datasets miss installation behaviors of certain apps if they are not conducted through the Wandoujia management app. Besides, apps' versions where the data come from were not captured, which may make it difficult to well explain the reason why users install or update the apps.

2.4. App metadata

App metadata refers to the basic information of apps including icon, description that introduces an app function, category that one app belongs to, reviews (comments), and downloads (how many users download the app). We take WhatsApp in Google Play as an example to show the app metadata in Fig. 2. App metadata helps us understand one app and further infer why one user installs or uses it. Particularly, we can extract the app function from its description, which can indicate the activity performed by one user. Thus, app metadata, especially app description and categories, is usually analyzed together with the other three types of app data. App metadata is usually crawled from app store websites, such as Google Play[3] and Wandoujia.[4]

In spite of the fact that app metadata are easily obtained from app store websites, there are some limitations. For example, reviews and rating can be quite sparse and even low-quality of some apps [46] and only very few apps can receive useful feedback from users [47]. The category assigned to the same app may be different across different app stores. Besides, the description of some apps is rather limited, consisting of just a few words.

We list in Table 2 some example datasets for user profiling, including the type of app data, platforms, duration, description, public or not. These datasets were collected from different mobile platforms, including Android, iOS, and Nokia. Among the existing studies, most datasets were collected from Android devices, while the number of iOS devices sampled is much smaller than those associated with other major platforms. Most of the datasets currently used are private, such as the installed app lists in [48], and app usage records [13]. There are only a few app datasets that have been publicly available in the context of programming competitions. For example, app start and close events, together with other phone usage states, were collected from 200 Nokia devices over a one-year period, and made publicly available in a mobile data challenge (MDC) [34]. In 2014, another competition was organized based on a large scale of smartphone usage data (Device Analyzer data) collected from more than 17,000 Android devices globally [39]. In the Device Analyzer

[3] https://play.google.com/store/apps?hl=en.

[4] https://www.wandoujia.com.

Please cite this article as: S. Zhao, S. Li, J. Ramos et al., User profiling from their use of smartphone applications: A survey, Pervasive and Mobile Computing (2019) 101052, https://doi.org/10.1016/j.pmcj.2019.101052.

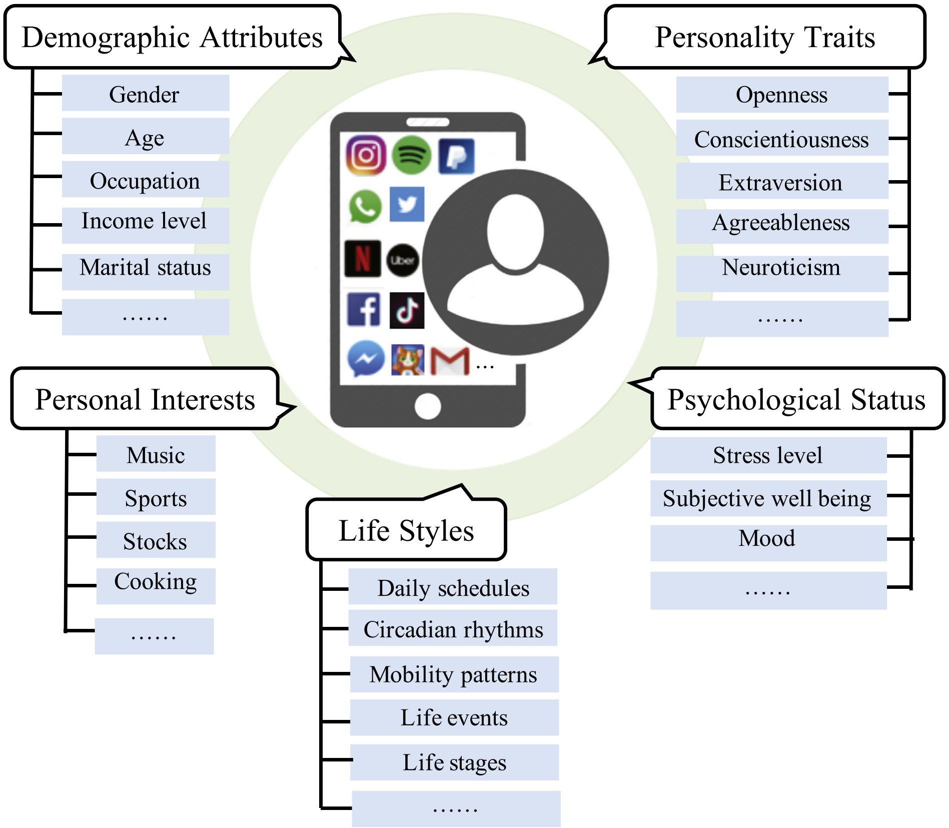


Fig. 3. User information that can be profiled from smartphone apps.

data, the app data was also collected, including the apps installed on the devices, the updates and removals, and the 10 most recently started apps.[5] Besides, there are also some other sensing frameworks Funf,[6] AWARE[7] [49] and Carat project [50] which continuously and unobtrusively gather app data (e.g., running apps, installed apps).

3. User information to profile

Apps on smartphones are a reflection of what users need, what they look like, what they are interested in, what activities they perform, how they live, etc. There have been many types of user information that have been learned from app data, such as gender [7], age [8], income [51], interests [19], personality traits [52], and life stages [29]. This survey roughly divides the types of user information into five categories: demographic attributes, personality traits, psychological status, personal interests, and life styles, shown in Fig. 3.

3.1. Demographic attributes

Demographic attributes specify those intrinsic qualities that belong naturally to a person, usually collected in a census. Typical demographic attributes include age, gender, marital status, race, religion, education, income, and occupation. Demographic attributes are a key factor in marketing products and services, useful to generate more revenue and profit margins. Knowing users' demographics enables a business owner to find the audience who fits the mold of the ideal customer. For example, customers of various ages may have different needs or services, and it is very helpful in marketing to know the primary ages of the most common customers. Specialty fashion retail shops often target a younger female customer, for instance. In this case, an age range like 21 to 34 is often used to depict the likely consumers.

That users with different demographic attributes may have different needs, makes it natural for the apps installed on smartphones of different users to be distinct. For example, young parents are likely to be interested in child rearing

[5] https://deviceanalyzer.cl.cam.ac.uk/collected.htm.

[6] http://www.funf.org/about.html.

[7] https://awareframework.com.

Please cite this article as: S. Zhao, S. Li, J. Ramos et al., User profiling from their use of smartphone applications: A survey, Pervasive and Mobile Computing (2019) 101052, https://doi.org/10.1016/j.pmcj.2019.101052.

related apps, and users who work in financial sectors are likely to be interested in stock related news. Even for the same app, its usage can be different across users on attributes like frequency or duration of interaction. These differences in smartphone apps make it possible to infer personal demographic attributes from apps.

3.2. Personality traits

Personality traits, a relatively permanent individual characteristic, play a central role in describing a person, which refer to the pattern of thoughts, feelings, social adjustments, and behaviors consistently exhibited over time that have a strong influence on one's expectations, self-perceptions, values, and attitudes. The main personality traits considered in psychology are referred to as the Big Five: Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness [53]. Personality traits, besides being important solely from the psychological point of view, influence decision making, academic and job performance, the behaviors in social interactions, physical health and risk of mental disorder.[8] Determining the personality of smartphone users can be used in various ways. For instance, human–computer interaction design principles could be used based on different demographics and personality traits to further improve the adoption of apps. The surface color of apps could be adaptively changed according to users' personality, since personality is linked to user interface preferences [54].

Several recent studies have found that the use of Internet, and forms of social media such as Facebook, are related to personality traits [55]. Smartphones have also become very important tools for social interaction and it is therefore highly likely that the personality of an individual could shape the usage of apps in smartphones. For example, extroverts probably use apps related with social network or communication more often than others, while conscientiousness has negative and significant effect on the adoption of services like photography, media & video, and location-based services, explained by their goal-driven nature and decreased use of leisure services to have fun [10,56]. Thus, it is possible to draw the connection between smartphone app usage and the personality traits of individuals.

3.3. Psychological status

Psychological status refers to a state of psychological well being throughout life, influencing how we behave and make decisions. Given that smartphones now mediate a wide range of daily behaviors from work to entertainment activities, app usage can provide informative cues for understanding individual performance in psychology. For example, LiKamWa et al. [57] found that users use different apps and/or communicate with different sets of people depending on their mood. They found that phone calls and apps grouped by category tended to be the strongest predictors of mood. Gao et al. [58] found that app usage behaviors correlate with subjective well being, especially for females. Subjective well being focuses on how one people evaluates his/her own life, including emotional experiences of pleasure versus pain in response to specific events and cognitive evaluations of what he/she considers a good life [59]. Thus, one's psychological status could be monitored and promoted from his/her use of the apps on smartphones.

3.4. Personal interests

Personal interests refer to the objects, events or processes on which users attend to focus. In addition to demographic attributes and personality traits, one's interests could shape his/her adoption of mobile applications. Knowing users' interests is important for product makers and service providers in marketing. Appropriate products or services could be recommended to users according to their interests. Users with different interests may install different apps on their smartphones. The installed apps are a display of users' interests, especially the niche apps that are popular among a small user group. We refer to niche apps as those generally created to serve a limited function, which is to provide a particular service or connect people sharing a common interest, rather than providing multiple options to suit a broader audience. Hence, this kind of app serves a niche of the user base of smartphone users.

3.5. Life styles

Life styles reflect a way of living of individuals, such as (1) daily schedules (e.g., time to go to bed and to wake up) or everyday routines, (2) circadian rhythms that cyclically happen in a period, (3) mobility patterns, a series of moves that regularly happen, (4) important life events happening in one's life that would change his/her living behaviors, conditions and psychological states in the future (e.g., buying a new house, getting married), and (5) life stages that are developmental phases during life cycle, each of which is with its own biological, psychological, and social characteristics (e.g., teenager, couple with children). With the knowledge of life styles, smart applications or services can be automatically recommended to users, so as to improve their life quality. For example, knowing one user's daily schedule that he/she takes yoga at 8 pm on Saturday, the calendar automatically reminds him/her, and the appropriate app could be highlighted at the right time. For the people who is going to buy a new house, related news about the house could be recommended accordingly.

[8] https://medicalxpress.com/news/2013-01-personality-decision-making-longevity-mental-health.html.

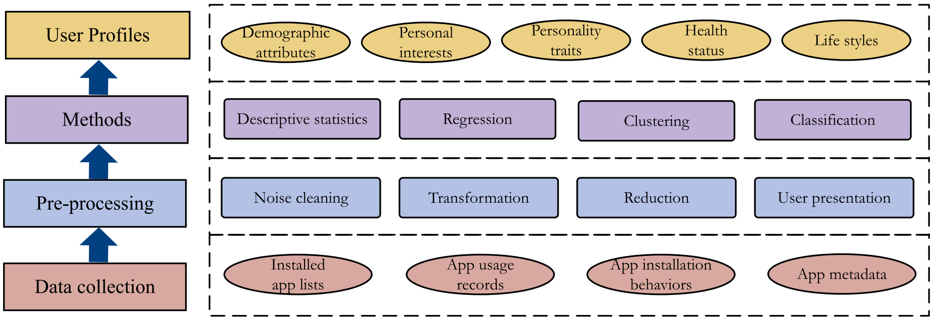


Fig. 4. A generic framework for user profiling from smartphone apps.

Smartphone app usage depends on context information to a certain degree, such as temporal context, spatial context, and environments. For example, users use different types of apps when they wake up or go to bed [38]. At certain locations, users are more likely to exhibit interest in a particular class of apps [36,60–62]. Intuitively, users with different life styles have different needs and interests. In particular, people's attitude and behavior might be associated less with the biological process of aging, but more with individual life events or stages [63]. In other words, life events or stages can bring about changes in interests, behaviors, personality traits, and needs. As mentioned above, users' interests, personality traits, and needs could shape the installation and usage of apps on smartphones. Thus, the differences in life styles could be reflected in installed app lists and app usage records. For example, compared with the singles without children, the young families with a baby probably install and use apps about raising a baby more frequently.

4. User profiling framework

It is not a trivial task to profile users from smartphone apps, consisting of many steps from data collection to the learned user characteristics. We summarize and present a generic framework for learning user information from smartphones' apps, including data collection, pre-processing, methods and user profiles, shown Fig. 4.

We also find there have been some studies using other clues to profile users, from which some solutions could be learned. Thus, before we go into detail on those studies using smartphone apps, we first introduce some user profiling studies from other clues, such as web logs, online social network, and call detail records (CDRs). For example, in [64–67], linguistic or textual differences were analyzed using basic statistics or regression, and user attributes, such as gender, age groups and interests, were inferred from the web logs. In [68–72], the behavior differences in online social networks were analyzed and then the user attributes were inferred by classification methods. There were also behavior differences found from CDRs [16,17,73–76]. User attributes, such as gender, age and marital status, were also inferred by training classifiers [16,17,73], and important locations in daily life, such as home and work place, were detected by clustering methods [75].

As we can see, the user information like demographics, interests and life styles were also inferred from web logs, online social network, and CDRs. To mine the user information, these studies are conducted following similar procedures, firstly collecting data, then performing pre-processing work on data such as noise cleaning and data dimensionality reduction, and building models on the prepared data. Several methods have been applied, such as basic statistics (e.g., average, variance), correlation or regression, classification methods (e.g., SVM, Naive Bayes, Logistic Regression) and clustering methods (e.g., k-Means). The procedures are also needed for user profiling from smartphone apps. It is also found that some methods used in these studies have been applied to profile users from smartphone apps. Thus, the user profiling studies in other domains could give us suggestions and enrich the future directions of using smartphone apps to profile users.

In the next sub-sections, we will discuss the procedures of data collection, pre-processing and modeling for user profiling from smartphone apps.

4.1. Data collection

For app data collection, there are three popular ways. The first one is to collect smartphone app data by designing specific apps, such as AppSensor [40] and AppJoy [30]. Most data collection apps are implemented on Android system which provides the required openness [40]. Among the platforms, it was reported that iOS was the most difficult one for data collection as "jailbreaking" iOS devices were typically required [30,77] before a data acquisition client could be installed. Thus, the number of iOS devices sampled is much smaller than those associated with other major platforms. The collection apps are distributed in app markets like Google Play to recruit voluntary participants. Such a way cannot be widely applied by a variety of crowds because of security and privacy concerns.

The second one is to build complete mobile sensing framework like Funf,[9] AWARE,[10] [49] Carat project [50], and Device Analyzer [78] which continuously and unobtrusively gather mobile phone data. Among the phone sensing data, app data (e.g., running apps, installed apps) is included. The data collected through the mobile sensing frameworks is relatively large-scale, but it usually takes a long time and the duration varies in a wide range across users. The third one is that the data is provided by the cooperated company who collected user data under privacy agreements, and researchers use the data for academic purposes. The internet companies efficiently store and employ customer historical data to analyze the customer characteristics, so as to improve their commercial services accordingly. Researchers use the data on the basis of protecting user privacy. In general, the population size of the provided datasets is much larger, the datasets still have some limitations yet, such as low sampling rate (e.g.,[13]) and the missing of logs of certain apps (e.g., [36] and [45]).

4.2. Pre-processing

The raw app data is often incomplete, inconsistent, or is likely to contain some errors, which can produce misleading results or make knowledge discovery more difficult. Commonly required as a preliminary step, data pre-processing is performed on raw app data to prepare it for further processing. It usually includes noise cleaning, app data transformation, and data dimensionality reduction.

(1) Noise cleaning removes the noise from the app data that is irrelevant or meaningless to research goals. There are several approaches to remove noisy data, and here we summarize three ones, including observation-based, distribution-based, and clustering-based approach. First, the observation-based approach removes noisy data based on researchers' observation and experiences, especially for incomplete and abnormal data samples that are not difficult to identify. For example, the data of the users who did not complete the questionnaires were removed [48,79], and only the users with complete questionnaires were taken into consideration. Pre-installed apps were manually identified and removed from users' installed app lists, since they cannot reflect users' needs and interests and would bias the correlation between users and apps [9,19]. Although the observation-based approach is fairly simple to conduct, it is difficult to find the noisy data samples that are hidden in the normal data samples. Second, the distribution-based approach is based on the distribution of a statistical dataset, such as normal distribution and long-tail distribution, to detect the noisy data samples that are far away from most data samples in the dataset. For example, Malmi et al. [51] focused on the apps frequently used by users and discarded the ones used by less than ten users to remove all personally identifiable information. In [13], the users who used their smartphones less frequently were taken as outliers and removed, and the ones that used fewer than 5 or more than 80 apps in total over the month of data collection were also removed. In spite of the simplicity to implement, it is not easy for the distribution-based approach to determine the boundary to separate out the noisy data samples. Third, clustering-based approach identifies noisy data samples as a by-product of the clustering process, where small clusters, that are far away from other major clusters, are taken as sets of noisy samples. Although it can detect more complex outliers, the clustering-based approach is not so frequently used in this field compared with the other two approaches, since it is sensitive to the choice of clustering algorithms and has difficulties in deciding which clusters should be identified as outliers [80].

(2) App data transformation converts the raw data into understandable, unified, and ready-to-use form. Some app datasets were collected in a complex form which make it difficult to use for profiling. For example, in [13], the source data of lists of recent tasks was used to discover smartphone user groups, each list consisting of up to 10 package names that can be used to identify an app. It was unknown which app was used in each hour slot and how much it was used in this source data, and the number of apps in different lists could be different. It was difficult to obviously describe users' app usage behaviors. Thus, the authors converted the source data into an understandable format, by detecting which app was used and computing how much it was used in each hour slot. In [8], in order to detect the app usage records from the internet entries when apps fetch resources from the Internet, Qin et al. used Regular Expression Matching with different key words ("football" refers to a category of "sports"), and matched the corresponding Internet resource to a certain app category. Thus, each Internet entry was converted to a user's request to an app category. Yu et al. [36] identified app usage records from the raw data which was cellular data accessing traces obtained by Deep Packet Inspection.

(3) Dimensionality reduction refers to reduce the number of variables input to models to minimize the data amount that is needed for research goals [81]. Dimensionality reduction improves computation efficiency, especially for the datasets where there exists large number of variables, data redundancy and data sparsity. For example, considering the high number of apps in the datasets, apps were divided into relatively small number of app categories based on their function similarity [13,48,51,52]. Moreover, compared to apps, app categories have an inherent semantic meaning, e.g., News, Games, and Banking, that allow us to reason more easily about app usage than using the name of the apps alone. In [19], if all the apps were used to represent users, user vectors would be dramatically long and very sparse. Thus, important apps for a given user attribute were selected by the method of information gain, since some apps are redundant and not all the apps are useful for describing a user attribute. In this way, the data dimensionality was dramatically reduced, and the computation efficiency was greatly improved. Seneviratne et al. [7] selected the top-50 words with relatively high information gain from all the app description to represent each user rather than all the words.

[9] http://www.funf.org/about.html.
[10] https://awareframework.com.

PMCJ: 101052

Table 3
Literature survey in profiling users from smartphone apps.

	Reference	Data	#User	Duration	User information	Method	Result
Descriptive statistics	Andone et al. [82]	AppUsage records	30,677	28 days	Differences in gender and age	Statistics	-
	Peltonen et al. [83]	AppUsage records	3,293	1 year	Country differences	Correlation	-
	Lim et al. [47]	Installation behaviors	4,824	2 months	Country differences	Correlation	-
	Murnane et al. [84]	AppUsage records	20	40 days	Circadian rhythms	Correlation	-
	Welke et al. [85]	AppUsage records	46,726	2 years	User differentiating	Hamming-distance	-
	Frey et al. [79]	App installation behaviors	2 008	1 month	life events	Keyword-based classifier	65%
	De Reuver et al. [41]	App usage records	233	28 days	Everyday routines	Correlation	-
Regression	Unal et al. [86]	AppUsage records	285	-	Big-Five personality	Regression Analysis	-
	Xu et al. [56]	App installation behaviors	22	1 month	Big-five personality	Linear regression	60%
	LiKamWa et al. [57]	AppUsage records	32	2 months	Mood	Linear regression	93%
	Gao et al. [58]	AppUsage records	106	-	Subjective well being	Linear regression	62%
Clustering	Zhao et al [13]	AppUsage records	106,762	30 days	User groups	k-means+ Meanshift	382 user groups
	Lee et al. [87]	AppUsage records	180	-	User groups	GMM	10 user groups
	Jesdabodi et al. [38]	AppUsage records	24	3 months	Users' current activity identifying	k-means	13 activities
	Amoretti et al. [60]	AppUsage records	100	2 months	Mobility	k-means	-
Classification	Seneviratne et al. [7]	Installed App lists	218	-	Gender	SVM	70%
	Seneviratne et al. [48]		231	-	Religion, country...	SVM	Precision > 90%
	Zhao et al. [19]	Installed App lists	100,281	-	12 predefined traits	SVM	EER: 16.4%
	Ferdous et al. [88]	AppUsage records	28	6 weeks	5 stress levels	SVM	75%

	Malmi et al. [51]	UsedApp lists	3 760	1 month	Gender;
					Age(18–32 *vs.* 33–100)
					Race
					. . .
						Logistic regression	82%
							77%
							73%
							. . .

	Chittaranjan et al. [52,89]	AppUsage records (MDC)	83	18 months	Big-Five personality	C4.5	76%
	Qin et al. [8]	AppUsage records	32,660	4 months	Gender 5 age groups	Bayes-based classifier	81% 74%
	Zhao et al. [90]	AppUsage records	10,000	3 months	Gender 3 Income levels	Neural networks	82% 64%
	Frey et al. [29]	InstalledApp lists	1,453	-	Life stages	Random forest	85%
	Mo et al. [16] Brdar et al. [73] Ying et al. [17]	AppUsage records (MDC)	83	1 year	Gender, Marital status, Age group, Job type	RandomForest, SVM, GBDT, KNN, multi-level classifier	88% 61% 81%

4.3. Methods

In the literature, there are numerous methods utilized for user profiling from smartphone apps. The choice or propose of methods depends on the research problem to be solved. User profiling can either be for an individual or a group of users, and this property has to be considered when the suitable method is chosen. For example, a simple clustering method, which depends on the similarity of users in app usage behaviors, is inadequate for extracting one's key characteristics from the perspective of an individual. Therefore, it is impossible to obtain meaningful results with inappropriate methods.

It is not easy to make a thorough analytical comparison of the existing methods, because the performance of the given method also depends heavily on the available data sources. Therefore, some of the models would outperform the others based on a given dataset; but for a different dataset, their performance would deteriorate. Hence, the appropriate method should be chosen by examining beforehand the given dataset and understanding the effect of various parameters.

We summarize and roughly divide the methods into four categories: descriptive statistics, regression, clustering and classification. In Table 3, we present prominent empirical user profiling approaches often used in the literature. The table includes data type and population size, user information, methods and results, where MDC for mobile data challenge, SVM for Support Vector Machine, GBDT for Gradient Boosting Decision Tree, kNN for k-Nearest Neighbor, GMM for Gaussian Mixture Model, and EER for equal error rate. We also highlight the advantages and disadvantages of the methods.

As we can see from Table 3, the study scale can vary in a large range from 20 users to more than 100,000 users. The data duration also varies a lot, from 28 days to around 2 years. Among the four types of app data, app usage records have been used the most for user profiling (17 out of 28 tabulated studies). Installed app lists and app installation behaviors have been explored in four studies, respectively. App metadata, especially app categories and app description, has been used in most of the studies, since it can be available from public app store websites. Classification methods are the most used in the studies. Almost half of (13 out of 28 in total) the tabulated studies utilized classification methods to profile users in an individual level.

4.3.1. User profiling with descriptive statistics

Descriptive statistics are a fairly simple approach for user profiling from smartphone apps. These methods use only basic statistical assumptions such as average, variance and correlation, to depict physical laws or patterns of user characteristics from smartphone app data. The modeling process is to discover the physical phenomenon in a statistical way. Many laws of nature are of statistical origin, indicating the relationship between users and smartphone apps.

For example, Peltonen et al. [83] carried out a large-scale analysis of geographic, cultural, and demographic factors in app usage, based on a dataset of 3293 countries from 44 countries. It was demonstrated that, the county of the participant has almost twice the information gain than any demographic factor in explaining app usage. Lim et al. [47] investigated the country differences in terms of their app installation behaviors, such as downloading and visiting app stores to look for apps, based on a dataset of 4824 participants from top 15 GDP countries. It was found that, for example, users from USA likely to download medical apps, and the users from the United Kingdom and Canada are more likely to be influenced by price. Welke et al. [85] explored used app lists to represent users, and measured the difference between users by computing the Hamming distance. It was found 99.67% of the users were unique in their used app lists. By analyzing the same app dataset as Welke, Andone et al. [82] investigated how age and gender affect app usage. It was found that females use communication and social apps longer than males. Communication and social apps are used heavily by teenagers. Games, Media and Video apps usage decline as age increases. Similarly, basic descriptive statistics and contextual descriptive statistics in terms of time of day and location were made in [40,91].

Murnane et al. [84] analyzed 20 participants' app usage over 40 days and investigated the correlation between app usage (app launches or switches, and duration) and circadian rhythms (chronotype traits, alertness levels, and sleep behaviors). In particular, the productivity and entertainment app usage have daily and weekly rhythms, differ in amount and timing for different chronotypes, align with trends in alertness performance, and correlate with adequate and inadequate sleep. Similarly, correlation between common places and app usage (e.g., [35]) and the correlation between subjective well being and app usage were examined [58]. Frey et al. [79] found that users' app installation behaviors could be used for predicting upcoming life events: first car, first job, marriage, first apartment and first child. Analyzing the app installation behaviors of 2008 users in one year, they inferred users' life events with an average accuracy of 64.5% by proposing a keyword-based classifier.

4.3.2. User profiling with regression

Regression is used for modeling the relationship between apps and users. It is a technique that builds a continuous function between a dependent variable (user information) and one or more independent variables (apps). Regression methods help understand how the apps on smartphones are correlated with types of user information. Among the regression methods, linear regression was the first to be studied. Linear regression models have many applications that cover two broad categories:

(1) Quantifying the strength of the relationship between apps and users, and in particular determining which apps may have no linear relationship with the given user characteristic at all, or identifying which subsets of apps may contain redundant information about the user characteristic. For example, multiple regression analyses were performed for each of the five personality traits with certain app usage, in order to determine whether the app usage could be indicators

for inferring personality traits [52]. Unal et al. [32,86] also investigated the relationship between apps and individual differences in personality traits.

(2) Fitting a predictive model to the given user characteristic using app data. For example, Xu et al. [56] used Linear regression method to detect one's personality with an average accuracy of around 60% by evaluating his/her history of app installations and update activities. LiKamWa et al. [57] learned the mood of a smartphone user by analyzing communication history and app usage patterns, and statistically inferred a user's daily mood average with 93% accuracy after a two-month training period using linear regression. App usage, phone calls, email messages, SMS, web browsing histories and location changes over two months were collected from 32 users. They found that phone calls and apps grouped by category tended to be the strongest predictors of mood. Gao et al. [58] also predicted 106 users' SWB (Subjective Well Being) using app usage records together with other smartphone usage behaviors and achieved an accuracy of 62%. They found that smartphone users with higher SWB scores tend to use more communication apps, play more games, and read more, but take fewer photos.

4.3.3. User profiling with clustering

There have been some studies profiling users' characteristics from the perspective of user groups. In the real-world, users with similar characteristics may install similar apps, or use apps in similar ways, attempting to aggregate into a group. In a feature space where features are derived from smartphone apps, users with similar characteristics form a cluster. In the case of clustering, it is to group a given collection of unlabeled patterns into meaningful clusters based on similarity [92]. Thus, the task of profiling user in a group level is smoothly transformed into a clustering problem. It can be solved by segmenting users into clusters, with the most similar users being grouped into the same one cluster. In other words, whereas a user in a certain group should be as similar as possible to all the other users in the same group, it should likewise be as distinct as possible from users in different groups.

Clustering identifies commonalities in the data and then infers the knowledge from such commonalities. Among the clustering methods, k-means is one of the most popular methods. For example, Jesdabodi et al. [38] clustered the app usage behaviors of 24 users during 3 months to identify their current activities by k-means, and detected 13 different kinds of activities, such as gaming and browsing. Amoretti et al. [60] discovered dynamic user groups using k-means based on user preferences. Although k-means runs fast, it is difficult to decide the number of clusters. In order to solve the problem, Zhao et al. [13] proposed a two-step method where they first made use of k-means with a pre-specified number of clusters and then clustered the centroids found using MeanShift, and discovered 382 different kinds of smartphone users by clustering the app usage behaviors of 106,672 Android users. The users in each cluster have distinct habits. For instance, the users in a cluster with 3814 users frequently wake up their smartphones but rarely unlock the screen and enter the main interface, just to check the time or to see if there are any notifications. Some other clustering methods have been used. For example, Lee et al. [87] used GMM to cluster 180 users into 10 groups based on their app usage sequences. In addition to the difficulty in deciding the number of clusters, it is not easy to determine an appropriate metric to measure the performance and interpret the clustering results.

4.3.4. User profiling with classification

Clustering methods profile users from the perspective of user groups. Many studies have attempted to infer individual information from his/her apps, by identifying which category of a given user information each individual belongs to, such as detecting one's gender from the available categories (female and male) through his/her apps. A type of user information are usually divided into multiple categories, such as personality traits (Big-five: Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness). Actually, inferring user attribute is to know which category the individual belongs to for a given user attribute. From the viewpoint of machine learning, it is a multi-class problem. Thus, inferring user information from smartphone apps could be solved by the method of classification.

A classification method identifies a set of categories to which each individual belongs, on the basis of a training set of samples with their groundtruth labels. For inferring user information from smartphone apps, there have been many classification methods, which are roughly divided into the following classes:

(1) *SVM (Support Vector Machine)* shows good generalization performance for solving classification problems [93], which maps the input points into a high-dimensional feature space and finds a separating hyperplane to maximize the margin between classes. It has been widely used in inferring user information from smartphone apps. For example, SVM was used to infer user demographic attributes from installed app lists in [7,19,48], and predict stress levels from app usage records in [88]. Although SVM performs well, it takes relatively long time for training, especially for large-scale datasets.

(2) *Bayesian models* are statistical models, the idea of which is to use the joint probabilities of apps and categories to estimate the probabilities of categories given a user. For example, Qin et al. [8] used a Bayesian model to infer users' gender and age group from app usage records with the accuracy of 81.12% and 73.84%, respectively. Some Bayesian models assumes that the features used are independent from each other in terms of impacting a user's profiles [5]. Naive Bayes makes the computation efficient, however, it ignores the close correlation among features, for example, the relationship between app usage and time that app usage behaviors have been shown to exhibit specific temporal pattern [38].

(3) *Decision tree* is a flowchart-like tree structure, where each internal node denotes a test on one user, each branch represents an outcome of the test, and each leaf node holds a class label [94]. Chittaranjan et al. [52,89] trained C4.5 classifiers to infer users' Big-five personality, where C4.5 is a popular decision tree based algorithm. Although it is simple to understand and interpret the results, decision trees have a large memory requirement problem when processing big data.

(4) *kNN (k-Nearest Neighbors)* is a simple regression algorithm that compares a given user with *k* training user samples which are similar to her. These *k* training user samples are the *k*-nearest neighbors (based on a distance measure) of the given user. Brdar et al. profiled users' demographic attributes with training kNN classifiers [73]. It is computation expensive for finding the nearest neighbors.

(5) *Logistic regression* models the probability of a type of user information as a linear function of a set of variable factors derived from app data. For example, Malmi et al. [51] analyzed the used app lists of 3760 Android users, and used Logistic regression to infer the demographic attributes. Logistic regression is easy to implement and efficient to train, however, it cannot solve non-linear problems.

(6) *Neural networks* have been becoming the most widely used methods nowadays, which are able to model non-linear behaviors and are highly extensible. For the task of inferring user attributes, neural networks stack fully-connected neural layers to learn non-linear, hidden, and implicit features from app data, and then apply activation functions (e.g., softmax, sigmoid and tanh) for classification. For example, a DNN (Deep Neural Network) works on individual features (e.g., app-, category-, and app usage sequence-based features) and combinations of different features to infer users' gender and income levels (three levels) from their app usage behaviors, achieving the accuracy of 81.6% and 63.6%, respectively [90]. Although neural networks can achieve a relatively high performance, the training phase needs much more data than the other classification methods to tune the large number of parameters. Compared to linear methods, neural network models are difficult to interpret and it is challenging to identify which features are the most important and how they are related to the users' profile being modeled.

(7) *Ensemble learning* combines several machine learning techniques into one classification model in order to decrease variance, bias or improve prediction results. There are many types of ensembles, such as bagging and boosting. For example, *Random forest* combines random decision trees with bagging, having each tree in the ensemble vote with equal weight, to achieve high classification performance. Frey et al. [29] used Random forest to predict 1453 users' life stages from installed app lists and achieved an accuracy of 85%. *GBDT (Gradient Boosting Decision Tree)* iteratively constructs an ensemble of weak decision tree learners through boosting, to fit a sequence of the weak decision trees to weighted versions of the data. During the training process, feature selection is inherently performed, so that it is relatively easy to interpret the results. Zhao et al. [90] used GBDT to infer users' gender and income levels from their app usage records, respectively, and obtained the highest accuracy for both gender and income levels (82.5% and 69.7%), compared with the methods of SVM, Logistic regress, and DNN. The top 10 important apps selected by GBDT for distinguishing gender and income levels were analyzed.

5. Implications

So far we have discussed the user profiling studies from smartphone apps in the literature. Different types of user information have been learned through the use of smartphone apps, including demographic attributes, personality traits, psychological status, personal interests, and life styles. In this section, we will discuss some implications and suggestions that can be used in *enhancing commercial services to help improve profits*; *designing mobile apps and smartphones to improve user experience*; and *developing mobile context-aware tools to improve users' life quality*.

5.1. Commercial services

Targeted advertisement and smart services: *Advertisement and service providers* can target the most receptive users with certain characteristics with knowing users' interests, needs and tastes. Advertisements can be actively pushed to specific users that are probably interested in, depending on users' personal interests and life events. For example, the most popular movies right now including new reviews are pushed to users who like watching movies. Vaccines, parenting skills, and school districts could be delivered to young parents, while the advertisement about car sales could be pushed to users who are going to buy cars. Similarly, personalized services are provided to users according to their demands and preferences. For example, knowing users' upcoming life events like buying a new car, a new apartment and giving birth, relevant firms are able to conduct more effective personalized promotion [79]. Smartphones could be switched to mute mode when users go to bed. In addition to targeted advertising and smart services, recommendation could be more personalized after profiling users. For example, news about stocks and business can be recommended to financial users, while news, forum or applications about cars could be recommended to car lovers.

Commercial user profiling services: There have been many commercial services devoted to profile customers by leveraging their data on corresponding websites, mobile apps, etc. In particular, many *Data Management Platforms* (DMPs) are developed to efficiently store and employ customer data (e.g., cookie IDs, mobile identifiers), such as Adobe Audience

Manager[11] and Lotame,[12] which help marketers or advertisers build customer profiles based on demographic data, past browsing behaviors, location, device usage, etc. User profiling from smartphone apps could be used to in many DMPs, to help build a user profile in terms of smartphone apps.

5.2. Mobile apps and smartphones

Design and popularity of mobile apps: *App developers and designers* can think about their target populations in terms of adaptations and recommendations that their apps could support based on users' interests, needs, and tastes. For the users who are active in using phone and SMS apps during night hours, the home screen could be adapted to highlight phone and SMS apps during night hours [13]. For the users with relatively high income level, apps about traveling could be recommended, since they use the apps related to traveling more frequently than the ones with low income [90]. Different persuasive technologies and/or human–computer interaction design principles could be used based on different demographics and personality traits to further improve the adoption of apps [31]. For example, the surface color of apps could be adaptively changed according to users' personality, since personality is linked to user interface preferences [54].

App store management: *App store operators*, with knowing users' installation behaviors (install, uninstall, or update), could improve the app recommendation quality and prediction performance of app popularity, so as to manage app stores more efficiently. App recommendation system could be improved by identifying the apps that need to be recommended, and identifying users' interests in the respective app category [20,95]. If an app is downloaded by most users in a group with similar interests, then it is likely to be of interest for another user in the same group who has not yet downloaded it. By analyzing users' interests through their installed app lists and installation behaviors, app store operators could predict the app popularity to design appropriate ranking policy for apps.

Design and popularity of smartphones: *Smartphone manufacturers* can build smartphones that are targeted towards improving the experience of the users with specific characteristics by providing features that different users may value than others. For example, the ones who frequently take photos may value an improved camera more than others, and the ones who like playing games may value an improved CPU. *Mobile carriers* that sell smartphones often pre-populate the phones with apps of their choosing. With knowing who are the potential adopters of the apps, they could allow for the customization of what apps are made available for the targeted users. For example, a smartphone pre-installing an app that provides mobile education service should target on less conscientious young people, while a phone with a mobile financial service app pre-installed should set old people with high net income and high conscientious as its target group [10]. For young people, mobile carriers could pre-install popular photography related apps [90].

5.3. Mobile context-aware tools

User characteristics learned from smartphone apps could be used by other mobile context-aware tools to improve users' life quality. For example, *health monitoring tools* could be developed to analyze personal heath learned from smartphone app usage, and present feedback to users. Smartphone apps have the potential to infer users' health and well being, such as stress [88], mood [57], subjective well being [58] and Circadian rhythms [84]. Thus, health monitoring tools would be helpful to improve health by empowering users to curb poor behavior patterns and providing users' suggestions accordingly, such as encouraging more exercise and going to bed early. Besides, some other context-aware tools could apply the user profiles learned from the smartphone app usage, such as tools related to social activity, smart city, transportation and LBS (Location Based Service), and further improve people's life quality from different aspects.

6. Challenges

Analysis of smartphone app data is becoming a very promising way to better understand users. Still, there are some challenges for understanding users from smartphone apps. In this section, we will discuss the challenges in the following aspects: data, user representation and modeling, fusion of heterogeneous data, user privacy issues and uncertain factors that affect user behaviors.

6.1. Challenges in data

We will discuss the challenges in data in terms of *data collection*, *groundtruth collection* and *data pre-processing*.

(1) Data collection. App data collection faces some challenges. First, it is hard to collect *a large-scale dataset* in practice. Previous in-field user studies have made efforts to profile users towards using smartphone apps [12,17,38,48,56,57,84,87–89], but most of these studies were conducted using rather small-scale datasets. As we can see from Table 3, the study scale can vary in a large range from 20 users to more than 100,000 users. The number of users can be as large as 106,672. For app usage records, the data duration also varies a lot, from 28 days to around 2 years. Among all the 26 work, only 6 have more than 10,000 users and only 4 last more than 6 months. Some datasets were collected through a monitoring app

[11] https://www.adobe.com/cn/experience-cloud/topics/data-management-platform-dmp.html.

[12] https://www.lotame.com.

voluntarily installed on the subjects' devices. Such a study cannot be widely applied by a large number of devices because of security and privacy concerns. The limited scale of data could impact the training of models, for example, classifiers.

Second, there is also the issue of *population bias*. Due to the fact that population biases exist across different cultures, it becomes difficult to gather smartphone app data that could accurately and completely reflect or represent the real-world smartphone user characteristics. The datasets in most previous studies were collected from biased populations, typically based on subjects such as college students, questionnaire volunteers, or from the same region. For example, college students were focused on as a vulnerable and valuable population to study, but patterns are likely different for people of other age groups or occupation groups [32,84,86]. The participants in [96] all live in the Greater San Francisco Bay area, the subjects in [97] are all Bruneians and the ones in [41] are all from Dutch, the user samples analyzed in [13,24,36,45,98] are all Chinese users. The subjects were biased in terms of language, age, gender, or income [10,29,33,56,79,83].

Third, some datasets were collected from smartphones based *on only one kind of operation system*, for example, Android system. This brings difficulty to generalize the findings on one OS to other OSes, because users of other OSes may behave differently. Most collection tools were implemented on Android system which provides the required openness [40,45,84]. It is infeasible to set up behavior logging across all smartphone operating systems given that some operating systems such as iOS restrict such functionality in apps [47].

Fourth, the fact that users' usage records collected data through specific apps installed on their smartphones may impact attention and usage [96,99] due to privacy concerns, resulting in some *deviation from their normal behaviors*. According to the statistics in [100], about 1% of the participants changed their behaviors or switched off the collection app due to privacy concerns during the study. It suggests that it is important to protect users' privacy and explore more unobtrusive apps to track users' behaviors during the procedure of data collection. In addition to the user privacy issues, data collection tools should take *energy constraint* into consideration, since continuous tracking could quickly deplete the devices' battery.

(2) Groundtruth collection. Many approaches requires obtaining the groundtruth of participants, such as training procedure of classifiers and interpreting clustering results. However, it is difficult to collect groundtruth information in practice, let alone a large-scale groundtruth dataset. On one hand, due to privacy concerns, participants are reluctant to provide their groundtruth. On the other hand, participants were asked to do questionnaires to make an estimate based on their past behaviors, which are subjective and inherently prone to bias [37,101].

(3) Data processing. Although smartphone apps reflect users' characteristics, when we conduct analysis on such plenty of datasets, we undoubtedly will face challenges in data processing during procedures. We identify following key research challenges that may be involved, i.e., *data noise*, *data redundancy*, *data sparsity*, and *imbalanced data distribution*. Generally, the raw data will likely be imperfect, containing some noises or redundancy. More specifically, in app usage records, a few apps that are very rarely used by users or the small number of users who use very few apps could be taken as noises, which influences the efficiency of models. The noisy samples could be filtered out according to the statistics [13]. App usage data has some redundant records, which are identical with each other, leading to storage and computation cost. For such cases, redundant records in a very short time period could be merged to one record [98]. In app meta data, reviews and rating can be quite sparse and even low-quality of some apps [46] and only very few apps can receive useful feedback from users [47]. To address the sparsity of app meta data, it would be a good solution to combine other types of app data together, such as installed app lists or app usage records, to compensate for the sparsity of the app metadata. In addition, there also exists imbalanced data distribution problem, such as unbalanced class sizes. For the class with very small size, the weights of samples could be increased in the cost function during training the models [19].

6.2. Challenges in user representation and modeling

Although there have been various models proposed for user profiling from smartphone apps, there are still challenges in many aspects, such as *effectiveness*, *performance metrics*, *interpretability* and *generalizability*. First, although the results of previous studies are quite good, the performance could be further improved by enhancing the *effectiveness* of the models. Currently, techniques used for user representation and modeling are rather straightforward, such as simply using app lists or app usage records as features, without considering the relationship between apps or the correlation between apps and users. More sophisticated machine learning methods or deep learning methods should be used in future research to automatically find out key features that are more predictive but less intuitive, such as key word embedding [102], app usage sequence embedding [87,90], and dynamic user profiling [103]. Second, it is important to determine which performance metrics should be used to measure the effectiveness of models, such as accuracy and execution time. Also, it is necessary to make a good trade-off between different metrics, for example, accuracy and execution time for an algorithm that would be applied for online application scenarios [98]. It is also difficult to determine an appropriate metric to measure the performance of unsupervised models (e.g., clustering methods), which usually depends on the discovered user characteristics.

Third, it is still challenging to interpret some results. In previous studies, various correlation-analysis studies have been made, such as the app usage frequency and users' demographics, app adoption and personality traits. However, not all of the analysis results can be fully interpreted. Many analyses have only correlation instead of causation, and comprehensively interpreting causation is rather difficult [45]. In addition, when we use sophisticated deep learning methods, it is difficult to interpret the discovered features and how the features are related to the user characteristics being modeled. Finally, it is difficult to generalize findings across studies due to several factors [104], such as limited scale of the used datasets, different user populations, different OSes and different experimental environments. Besides, most existing user representation models are task-specific which highly rely on labeled data, making it difficult to generalize to other tasks [105]. It is necessary to learn universal user representations that can capture global patterns of users and be easily applied to different tasks as additional user features [105]. The universal user representations do not highly rely on manually labeled data, instead, they can be learned from heterogeneous and multi-source data.

6.3. Challenges in fusion of heterogeneous data

Here, we focus on not only the fusion of different types of app data, but also the fusion of smartphone app data with the data from other sources.

As mentioned above, each type of smartphone app data can reflect user characteristics to a certain degree. However, only one type of app data, sometimes, is not enough to profile users. For example, in app meta data, reviews and rating can be quite sparse and even low-quality of some apps [46], bringing about difficulty in effectively building user representations. Li et al. found that a large number of users in a Chinese population do not frequently download and update their apps from app stores, implying app installation behaviors are not so strong indicators for profiling users [42]. Whether one user has installed an app may be a weak indicator of whether she actually needs the app. She may simply want to try the app out, and may never use it again or may have uninstalled it [10,19,29,29–32,106]. According to the statistics in [33], only 10% of apps were used 80% of the time, suggesting that a lot of apps are downloaded but not used regularly. To alleviate the shortcomings mentioned above in app meta data, installation behaviors and installed app lists, taking one user's daily app usage into account, such as usage time and the frequency of app usage, could make it more accurate in profiling user characteristics. The usage could help us to sort out the unused apps, and identify which apps one user actually needs. However, the different types of app data are constituted by different formats. For example, app description and ratings are in text, while app usage records are usually presented in temporal sequences and in installed app lists binary values are used to indicate whether one app installed or not. It is still challenging to combine all the app data types together to profile users.

Although smartphone apps provide us a great opportunity for profiling users, depending solely on behavioral signals from smartphone apps does ignore useful data from other sources. Human behaviors can be tracked by various sensing devices (e.g., cameras, laptops, glasses, microphones) [107,108]. The aggregation of these types of data with the smartphone app data will be useful for creating a comprehensive view of users. However, the data from different sources are stored in different structures, such as image, text, video, and audio, making it difficult to effectively fuse them.

Whatever the fusion of different types of smartphone app data, or the fusion of smartphone app data with the data from other sources, how we can collaboratively fuse all these different formats of data for smartphone user profiling remains to be explored. It is a challenging issue to match all the heterogeneous data together, for example, how to discover the relationship between text-based features (e.g., app description) and binary-based features (e.g., installed app lists)? How to combine one type of data with other types of data? What combination of heterogeneous data achieve better performance? Which type of features is more important for modeling? Fortunately, there have been some studies leveraging heterogeneous data to profile users [109–111], which can provide some insights for the fusion here. For example, Li et al. [109] integrated two types of signals observed from social network and user-centric data by taking a probabilistic generative approach to model them jointly. Chen et al. [110,111] proposed a semi-supervised feature selection method to extract significant features from the highly variant, and heterogeneous urban open data.

6.4. Challenges in user privacy

Using personal smartphone app data might trigger privacy concerns. In previous studies, the collection of smartphone app data can be used to mine user attributes and to create user profiles. Thus, such data should be considered as personal data that is sensitive related to privacy concerns. People are understandably sensitive about how app data is captured and used [112], especially the data collection procedure.

The collection of data should need users to grant permission, and be conducted under a privacy policy. But, some data can be obtained without users' permission. For example, it was reported in [28] that the list of apps installed on a user's smartphone can be obtained without his/her permission through any app installed on Android devices. iOS provided lists of installed apps before and removed this functionality in iOS 9. Most collection tools were implemented on Android system which provides the required openness, while the number of iOS device samples is much smaller due to the restrictions of such functionality in apps. On Android 6 and up, users can control which permissions, an app can access after the app is installed, such as the permission to obtain installed app lists on devices.

Although many previous studies used approaches (e.g., cryptography, offline storage, non-disclosure agreement) that help with protecting privacy, they are often insufficient [113]. To alleviate user privacy issues, we first suggest app publishers make the app transparent to corresponding app users, like what information will be collected, how to collect the data, and for what purpose. Users should be given rights to opt-in for providing the data. Second, phone systems like Android OS provide users rights to control which capabilities or information that an installed app can access-known as permissions. Third, app markets like Google Play could reinforce policies for app publishers to undesirable logging activities.

6.5. Challenges in uncertain factors that affect user behaviors

There are many uncertain factors outside the control out of researchers, which could affect app usage patterns and result in deviation. For example, low battery level might dissuade users from using video streaming or other power hungry apps. The introduction of a new technology in smartphones or change of smartphones also impacts the interaction patterns. Sometimes, one smartphone may be used by multiple users (e.g., partners/ kids & parents). In addition, user may change the ways in which they use their smartphones rapidly [78], and therefore may become outdated. In order to alleviate the issues, such as low battery level, change of smartphones and shared ownership of smartphones, we could try to collect rich context information to help with recognize the uncertain activities. On the other hand, if we could collect a relatively large-scale app dataset in terms of population size and duration, such cases mentioned above would not have a significant effect on our analyses. For the behavior change over time, we could build dynamic user profiling model that dynamically updates as time [103].

7. Conclusion and future work

As with the rapid prevalence of smartphone apps in our daily life, there is a vast number of app data collected that can be summarized into four types: installed app lists, app usage records, app installation behaviors and app metadata. Since a smartphone is usually tightly associated with one same user, the apps on smartphones convey lots of personal information, which we roughly divide into five categories: demographic attributes, personality traits, psychological status, personal interests, and life styles. There have been many previous studies to profile users from smartphone apps by exploring various methods, such as descriptive statistics, regression, clustering and classification. User profiling from smartphone applications can be exploited in a wide range of potential application scenarios from the perspective of advertisement and service providers, app developers and designers, etc.

However, it still remains challenging in user profiling from smartphone apps. *(1) Challenges in the data in terms of data collection and data quality*. It is hard to collect relatively large-scale datasets from various crowds of users, as well as the groundtruth information. Data processing procedure is necessary to deal with the issues like data sparsity and imbalanced data distribution. *(2) Challenges in user representation and modeling.* It is urgent to improve the effectiveness, interpretability and generalizability of models, and crucial to determine appropriate metrics to measure performance of models depending on the discovered user characteristics. *(3) Challenges in the fusion of heterogeneous data.* How to effectively aggregate data in different formats remains to be explored, such as discovering the relationship between different types of features and measuring the importance of them. *(4) Challenges in user privacy issues.* As a personal data, smartphone app data triggers user privacy. It is hard to totally protect user privacy. *(5) Challenges in uncertain factors.* There are many factors that could affect users' behaviors, such as behavior changes over time, device changes and shared owner ship, which may be not easy to capture and recognize.

In the future work, efforts can be made towards addressing the challenges mentioned above. For example, in order to protect user privacy, app publishers make app data collection transparent to users and app market reinforces policies for the management of published apps. On the basis of protecting user privacy, we can try to collect data from different sources taking advantage of crowdsourcing techniques, and fuse the data to comprehensively profile users. More sophisticated machine learning or deep learning methods can be proposed or improved to automatically find out key features that are more predictive but less intuitive, to improve the effectiveness of user profiling. Besides, it might be possible to learn more complicated user characteristics such as values and social attributes, so as to obtain a more comprehensive profile of users.

Acknowledgments

This work was supported by National Key R&D Program of China (No. 2018YFC1504006), NSF of China (No. 61772460, 61802342, and 61802340), China Postdoctoral Science Foundation under Grant No. 2017M620246 and 2018T110591, and the Fundamental Research Funds for the Central Universities (No. 181200*172210183). Dr. Gang Pan is the corresponding author.

References

[1] R. Pérez-Torres, C. Torres-Huitzil, H. Galeana-Zapién, Power management techniques in smartphone-based mobility sensing systems: A survey, Pervasive Mob. Comput. 31 (2016) 1–21.
[2] N.D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, A.T. Campbell, A survey of mobile phone sensing, IEEE Communications magazine 48 (9).
[3] M. Conti, A. Passarella, S.K. Das, The internet of people (iop): A new wave in pervasive mobile computing, Pervasive Mob. Comput. 41 (2017) 1–27.
[4] Y. Du, H. Ren, G. Pan, S. Li, Tilt & touch: Mobile phone for 3d interaction, in: Proceedings of the 13th International Conference on Ubiquitous Computing, ACM, 2011, pp. 485–486.
[5] H. Cao, M. Lin, Mining smartphone data for app usage prediction and recommendations: A survey, Pervasive Mob. Comput. 37 (2017) 1–22.
[6] A.K. Dey, Understanding and using context, Pers. Ubiquitous Comput. 5 (1) (2001) 4–7.

[7] S. Seneviratne, A. Seneviratne, P. Mohapatra, A. Mahanti, Your installed apps reveal your gender and more!, ACM SIGMOBILE Mob. Comput. Commun. Rev. 18 (3) (2015) 55–61.

[8] Z. Qin, Y. Wang, H. Cheng, Y. Zhou, Z. Sheng, V.C. Leung, Demographic information prediction: a portrait of smartphone application users, IEEE Trans. Emerg. Top. Comput. 6 (3) (2018) 432–444.

[9] Y. Wang, Y. Tang, J. Ma, Z. Qin, Gender prediction based on data streams of smartphone applications, in: Proceedings of International Conference on Big Data Computing and Communications, Springer, 2015, pp. 115–125.

[10] R. Xu, R.M. Frey, A. Ilic, Individual differences and mobile service adoption: An empirical analysis, in: Proceedings of IEEE Second International Conference on Big Data Computing Service and Applications, IEEE, 2016, pp. 234–243.

[11] C. Shin, A.K. Dey, Automatically detecting problematic use of smartphones, in: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2013, pp. 335–344.

[12] C. Shin, J.-H. Hong, A.K. Dey, Understanding and prediction of mobile application usage for smart phones, in: Proceedings of the 2012 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2012, pp. 173–182.

[13] S. Zhao, J. Ramos, J. Tao, Z. Jiang, S. Li, Z. Wu, G. Pan, A.K. Dey, Discovering different kinds of smartphone users through their application usage behaviors, in: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2016, pp. 498–509.

[14] S. Lin, R. Xie, Y. Chen, Y. Xiao, P. Hui, An empirical study of the usage of the swarm app's cross-site sharing feature, in: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, ACM, 2018, pp. 823–832.

[15] A. Al-Molegi, I. Alsmadi, A. Martínez-Ballesté, Regions-of-interest discovering and predicting in smartphone environments, Pervasive Mob. Comput. 47 (2018) 31–53.

[16] K. Mo, B. Tan, E. Zhong, Q. Yang, Report of task 3: your phone understands you, in: 2012 Nokia Mobile Data Challenge Workshop, Citeseer, 2012, pp. 18–19.

[17] J.J.-C. Ying, Y.-J. Chang, C.-M. Huang, V.S. Tseng, Demographic prediction based on users mobile behaviors, in: 2012 Nokia Mobile Data Challenge Workshop, 2012, pp. 1–6.

[18] A. Campbell, T. Choudhury, From smart to cognitive phones, IEEE Pervasive Comput. 11 (3) (2012) 7–11.

[19] S. Zhao, G. Pan, Y. Zhao, J. Tao, J. Chen, S. Li, Z. Wu, Mining user attributes using large-scale app lists of smartphones, IEEE Syst. J. 11 (1) (2017) 315–323.

[20] H. Li, X. Lu, X. Liu, T. Xie, K. Bian, F.X. Lin, Q. Mei, F. Feng, Characterizing smartphone usage patterns from millions of android users, in: Proceedings of the 2015 Internet Measurement Conference, ACM, 2015, pp. 459–472.

[21] M. Gustarini, M.P. Scipioni, M. Fanourakis, K. Wac, Differences in smartphone usage: validating, evaluating, and predicting mobile user intimacy, Pervasive Mob. Comput. 33 (2016) 50–72.

[22] N. Micallef, E. Adi, G. Misra, Investigating login features in smartphone apps, in: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, ACM, 2018, pp. 842–851.

[23] Y. Ouyang, B. Guo, T. Guo, L. Cao, Z. Yu, Modeling and forecasting the popularity evolution of mobile apps: A multivariate hawkes process approach, Proc. ACM Interact. Mob. Wear. Ubiquitous Technol. 2 (4) (2018) 182.

[24] S. Zhao, J. Ramos, J. Tao, Z. Jiang, S. Li, Z. Wu, G. Pan, A.K. Dey, Who are the smartphone users?: Identifying user groups with apps usage behaviors, GetMob.: Mob. Comput. Commun. 21 (2) (2017) 31–34.

[25] I.M. Pires, N.M. Garcia, N. Pombo, F. Flórez-Revuelta, S. Spinsante, M.C. Teixeira, Identification of activities of daily living through data fusion on motion and magnetic sensors embedded on mobile devices, Pervasive Mob. Comput. 47 (2018) 78–93.

[26] D. Ferreira, J. Goncalves, V. Kostakos, L. Barkhuus, A.K. Dey, Contextual experience sampling of mobile application micro-usage, in: Proceedings of the 16th International Conference on Human–Computer Interaction with Mobile Devices & Services, ACM, 2014, pp. 91–100.

[27] S. Zhao, Y. Zhao, Z. Zhao, Z. Luo, R. Huang, S. Li, G. Pan, Characterizing a user from large-scale smartphone-sensed data, in: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Workshop), ACM, 2017, pp. 482–487.

[28] M.C. Grace, W. Zhou, X. Jiang, A.-R. Sadeghi, Unsafe exposure analysis of mobile in-app advertisements, in: Proceedings of the FFifth ACM Conference on Security and Privacy in Wireless and Mobile Networks, ACM, 2012, pp. 101–112.

[29] R.M. Frey, R. Xu, A. Ilic, Mobile app adoption in different life stages: An empirical analysis, Pervasive Mob. Comput. 40 (2017) 512–527.

[30] B. Yan, G. Chen, Appjoy: personalized mobile application discovery, in: Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services, ACM, 2011, pp. 113–126.

[31] R. Xu, R.M. Frey, E. Fleisch, A. Ilic, Understanding the impact of personality traits on mobile app adoption–insights from a large-scale field study, Comput. Hum. Behav. 62 (2016) 244–256.

[32] P. Ünal, T.T. Temizel, P.E. Eren, What installed mobile applications tell about their owners and how they affect users download behavior, Telemat. Inform. 34 (7) (2017) 1153–1165.

[33] V. Rivron, M.I. Khan, S. Charneau, I. Chrisment, Exploring smartphone application usage logs with declared sociological information, in: 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom), IEEE, 2016, pp. 266–273.

[34] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, J. Laurila, Towards rich mobile phone datasets: lausanne data collection campaign, in: Proceedings of the International Conference on Pervasive Services, 2010.

[35] T.M.T. Do, J. Blom, D. Gatica-Perez, Smartphone usage in the wild: a large-scale analysis of applications and context, in: Proceedings of the 13th International Conference on Multimodal Interfaces, ACM, 2011, pp. 353–360.

[36] D. Yu, Y. Li, F. Xu, P. Zhang, V. Kostakos, Smartphone app usage prediction using points of interest, Proc. ACM Interact. Mob. Wear. Ubiquitous Technol. 1 (4) (2018) 174.

[37] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, S. Venkataraman, Identifying diverse usage behaviors of smartphone apps, in: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, ACM, 2011, pp. 329–344.

[38] C. Jesdabodi, W. Maalej, Understanding usage states on mobile devices, in: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2015, pp. 1221–1225.

[39] D.T. Wagner, A. Rice, A.R. Beresford, Device analyzer: Understanding smartphone usage, in: International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services, Springer, 2013, pp. 195–208.

[40] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, G. Bauer, Falling asleep with angry birds, facebook and kindle: a large scale study on mobile application usage, in: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services, ACM, 2011, pp. 47–56.

[41] M. De Reuver, S. Nikou, H. Bouwman, Domestication of smartphones and mobile applications: A quantitative mixed-method study, Mob. Media Commun. 4 (3) (2016) 347–370.

[42] H. Li, X. Lu, Mining device-specific apps usage patterns from large-scale android users, arXiv preprint arXiv:1707.09252.

[43] J. Huang, F. Xu, Y. Lin, Y. Li, On the understanding of interdependency of mobile app usage, in: 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems, IEEE, 2017, pp. 471–475.

[44] Z. Tu, R. Li, Y. Li, G. Wang, D. Wu, P. Hui, L. Su, D. Jin, Your apps give you away: distinguishing mobile users by their app usage fingerprints, Proc. ACM Interact. Mob. Wear. Ubiquitous Technol. 2 (3) (2018) 138.

[45] X. Liu, H. Li, X. Lu, T. Xie, Q. Mei, F. Feng, H. Mei, Understanding diverse usage patterns from large-scale appstore-service profiles, IEEE Trans. Softw. Eng. 44 (4) (2018) 384–411.

[46] B. Fu, J. Lin, L. Li, C. Faloutsos, J. Hong, N. Sadeh, Why people hate your app: Making sense of user feedback in a mobile app store, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2013, pp. 1276–1284.

[47] S.L. Lim, P. Bentley, N. Kanakam, F. Ishikawa, S. Honiden, Investigating country differences in mobile app user behavior and challenges for software engineering, IEEE Transactions on Software Engineering 41 (1) (2014) 40–64.

[48] S. Seneviratne, A. Seneviratne, P. Mohapatra, A. Mahanti, Predicting user traits from a snapshot of apps installed on a smartphone, ACM SIGMOBILE Mob. Comput. Commun. Rev. 18 (2) (2014) 1–8.

[49] D. Ferreira, V. Kostakos, A.K. Dey, Aware: mobile context instrumentation framework, Front. ICT 2 (2015) 6.

[50] A.J. Oliner, A.P. Iyer, I. Stoica, E. Lagerspetz, S. Tarkoma, Carat: Collaborative energy diagnosis for mobile devices, in: Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems, ACM, 2013, p. 10.

[51] E. Malmi, I. Weber, You are what apps you use: Demographic prediction based on user's apps, arXiv preprint arXiv:1603.00059.

[52] G. Chittaranjan, J. Blom, D. Gatica-Perez, Who's who with big-five: Analyzing and classifying personality traits with smartphones, in: 2011 15th Annual International Symposium on Wearable Computers (ISWC), IEEE, 2011, pp. 29–36.

[53] O.P. John, S. Srivastava, The big five trait taxonomy: History, measurement, and theoretical perspectives, Handb. Personal.: Theory Res. 2 (1999) (1999) 102–138.

[54] W.-P. Brinkman, N. Fine, Towards customized emotional design: an explorative study of user personality and user interface skin preferences, in: Proceedings of the 2005 Annual Conference on European Association of Cognitive Ergonomics, University of Athens, 2005, pp. 107–114.

[55] M.D. Back, J.M. Stopfer, S. Vazire, S. Gaddis, S.C. Schmukle, B. Egloff, S.D. Gosling, Facebook profiles reflect actual personality, not self-idealization, Psychol. Sci. 21 (3) (2010) 372–374.

[56] R. Xu, R.M. Frey, D. Vuckovac, A. Ilic, Towards understanding the impact of personality traits on mobile app adoption-a scalable approach, in: ECIS, 2015.

[57] R. LiKamWa, Y. Liu, N.D. Lane, L. Zhong, Moodscope: Building a mood sensor from smartphone usage patterns, in: Proceeding of the 11th Annual International Conference on Mobile SSystems, Applications, and Services, ACM, 2013, pp. 389–402.

[58] Y. Gao, H. Li, T. Zhu, Predicting subjective well-being by smartphone usage behaviors, in: HEALTHINF, 2014, pp. 317–322.

[59] E. Diener, Subjective well-being: The science of happiness and a proposal for a national index., Amer. Psychol. 55 (1) (2000) 34.

[60] M. Amoretti, L. Belli, F. Zanichelli, Utravel: Smart mobility with a novel user profiling and recommendation approach, Pervasive Mob. Comput. 38 (2017) 474–489.

[61] P. Li, H. Lu, N. Kanhabua, S. Zhao, G. Pan, Location inference for non-geotagged tweets in user timelines, IEEE Trans. Knowl. Data Eng. 31 (6) (2018) 1150–1165.

[62] G. Qi, G. Pan, S. Li, Z. Wu, D. Zhang, L. Sun, L.T. Yang, How long a passenger waits for a vacant taxi–large-scale taxi trace mining for smart cities, in: 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, IEEE, 2013, pp. 1029–1036.

[63] J.B. Lansing, L. Kish, Family life cycle as an independent variable, Am. Sociol. Rev. 22 (5) (1957) 512–519.

[64] S.C. Herring, J.C. Paolillo, Gender and genre variation in weblogs, J. Soc. 10 (4) (2006) 439–459.

[65] S. Nowson, J. Oberlander, The identity of bloggers: Openness and gender in personal weblogs, in: AAAI Spring Symposium: Computational Approaches To Analyzing Weblogs, Palo Alto, CA, 2006, pp. 163–167.

[66] S. Rosenthal, K. McKeown, Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, 2011, pp. 763–772.

[67] C.-Y. Teng, H.-H. Chen, Detection of bloggers' interests: using textual, temporal, and interactive features, in: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, 2006, pp. 366–369.

[68] D. Rao, D. Yarowsky, A. Shreevats, M. Gupta, Classifying latent user attributes in twitter, in: Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, ACM, 2010, pp. 37–44.

[69] D. Preoţiuc-Pietro, Y. Liu, D. Hopkins, L. Ungar, Beyond binary labels: political ideology prediction of twitter users, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 729–740.

[70] S. Mukherjee, P.K. Bala, Gender classification of microblog text based on authorial style, Inf. Syst. E-Bus. Manage. 15 (1) (2017) 117–138.

[71] A.A. Morgan-Lopez, A.E. Kim, R.F. Chew, P. Ruddle, Predicting age groups of twitter users based on language and metadata features, PLoS One 12 (8) (2017) e0183537.

[72] R.G. Guimaraes, R.L. Rosa, D. De Gaetano, D.Z. Rodriguez, G. Bressan, Age groups classification in social network using deep learning, IEEE Access 5 (2017) 10805–10816.

[73] S. Brdar, D. Culibrk, V. Crnojevic, Demographic attributes prediction on the real-world mobile data, in: 2012 Nokia Mobile Data Challenge Workshop, 2012.

[74] A. Arai, A. Witayangkurn, H. Kanasugi, T. Horanont, X. Shao, R. Shibasaki, Understanding user attributes from calling behavior: exploring call detail records through field observations, in: Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia, ACM, 2014, pp. 95–104.

[75] A. Arai, R. Shibasaki, Estimation of human mobility patterns and attributes analyzing anonymized mobile phone cdr: Developing real-time census from crowds of greater dhaka, in: AGILE PhD School, 2013.

[76] S. Zhao, Z. Zhao, R. Huang, Z. Luo, S. Li, J. Tao, S. Cheng, J. Fan, G. Pan, Discovering individual life style from anonymized wifi scan lists on smartphones, IEEE Access 7 (2019) 22698–22709.

[77] A. Mahfouz, I. Muslukhov, K. Beznosov, Android users in the wild: Their authentication and usage behavior, Pervasive Mob. Comput. 32 (2016) 50–61.

[78] D.T. Wagner, A. Rice, A.R. Beresford, Device analyzer: Large-scale mobile data collection, ACM SIGMETRICS Perform. Eval. Rev. 41 (4) (2014) 53–56.

[79] R. Frey, R. Xu, A. Ilic, Reality-Mining with Smartphones: Detecting and Predicting Life Events Based on App Installation Behavior, 2015, pp. 1–10.

[80] H. Xiong, G. Pandey, M. Steinbach, V. Kumar, Enhancing data analysis with noise removal, IEEE Trans. Knowl. Data Eng. 18 (3) (2006) 304–319.

[81] S.T. Rowes, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 232.

[82] I. Andone, K. Błaszkiewicz, M. Eibes, B. Trendafilov, C. Montag, A. Markowetz, How age and gender affect smartphone usage, in: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, ACM, 2016, pp. 9–12.

[83] E. Peltonen, E. Lagerspetz, J. Hamberg, A. Mehrotra, M. Musolesi, P. Nurmi, S. Tarkoma, The hidden image of mobile apps: Geographic, demographic, and cultural factors in mobile usage.

[84] E.L. Murnane, S. Abdullah, M. Matthews, M. Kay, J.A. Kientz, T. Choudhury, G. Gay, D. Cosley, Mobile manifestations of alertness: Connecting biological rhythms with patterns of smartphone app use, in: Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services, ACM, 2016, pp. 465–477.

[85] P. Welke, I. Andone, K. Blaszkiewicz, A. Markowetz, Differentiating smartphone users by app usage, in: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2016, pp. 519–523.

[86] P. Ünal, T.T. Temizel, P.E. Eren, Impact of individual differences on the use of mobile phones and applications, in: International Conference on Mobile Web and Information Systems, Springer, 2016, pp. 379–392.

[87] Y. Lee, I. Park, S. Cho, J. Choi, Smartphone user segmentation based on app usage sequence with neural networks, Telemat. Inform. 35 (2) (2018) 329–339.

[88] R. Ferdous, V. Osmani, O. Mayora, Smartphone app usage as a predictor of perceived stress levels at workplace, in: Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2015, pp. 225–228.

[89] G. Chittaranjan, J. Blom, D. Gatica-Perez, Mining large-scale smartphone data for personality studies, Pers. Ubiquitous Comput. 17 (3) (2013) 433–450.

[90] S. Zhao, F. Xu, Z. Luo, S. Li, G. Pan, Demographic attributes prediction through app usage behaviors on smartphones, in: Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Workshop), ACM, 2018, pp. 870–877.

[91] E.T.K. Wong, W.W. Ma, Sharing data and knowledge: Exploring relationships and difference among day, time, gender, place, and smartphone use, in: New Ecology for Education—Communication X Learning, Springer, 2017, pp. 263–275.

[92] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.

[93] C.J. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (2) (1998) 121–167.

[94] J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.

[95] X. Liu, H. Li, X. Lu, T. Xie, Q. Mei, H. Mei, F. Feng, Mining behavioral patterns from millions of android users, arXiv preprint arXiv:1702.05060.

[96] J.P. Carrascal, K. Church, An in-situ study of mobile app & mobile search interactions, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, 2015, pp. 2739–2748.

[97] M. Anshari, Y. Alas, G. Hardaker, J. Jaidin, M. Smith, A.D. Ahad, Smartphone habit and behavior in brunei: Personalization, gender, and generation gap, Comput. Hum. Behav. 64 (2016) 719–727.

[98] S. Zhao, Z. Luo, Z. Jiang, H. Wang, F. Xu, S. Li, J. Yin, G. Pan, Appusage2vec: Modeling smartphone app usage for prediction, in: Proceedings of the 35th IEEE International Conference on Data Engineering, IEEE, 2019.

[99] B. Thornton, A. Faires, M. Robbins, E. Rollins, The mere presence of a cell phone may be distracting, Social Psychology (2014).

[100] M. de Reuver, H. Bouwman, N. Heerschap, H. Verkasalo, Smartphone measurement: Do people use mobile applications as they say they do?, in: ICMB, 2012, p. 2.

[101] S. Butt, J.G. Phillips, Personality and self reported mobile phone use, Comput. Hum. Behav. 24 (2) (2008) 346–360.

[102] M.-Y. Zheng, H.-Y. Chen, H. Chen, Y.-C. Fan, On cleaning and organizing context logs for mobile user profiling, in: 2017 Twelfth International Conference on Digital Information Management (ICDIM), IEEE, 2017, pp. 161–164.

[103] M. Ouanaim, H. Harroud, A. Berrado, M. Boulmalf, Dynamic user profiling approach for services discovery in mobile environments, in: Proceedings of the 6th International Wireless Communications and Mobile Computing Conference, ACM, 2010, pp. 550–554.

[104] K. Church, D. Ferreira, N. Banovic, K. Lyons, Understanding the challenges of mobile phone usage data, in: Roceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services, ACM, 2015, pp. 504–514.

[105] C. Wu, F. Wu, J. Liu, S. He, Y. Huang, X. Xie, Neural demographic prediction using search query, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, ACM, 2019, pp. 654–662.

[106] X. Zou, W. Zhang, S. Li, G. Pan, Prophet: What app you wish to use next, in: Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication (Poster), ACM, 2013, pp. 167–170.

[107] Z. Yu, H. Du, F. Yi, Z. Wang, B. Guo, Ten scientific problems in human behavior understanding, CCF Trans. Pervasive Comput. Interact. (2019) 1–7.

[108] C. Chen, Y. Ding, X. Xie, S. Zhang, Z. Wang, L. Feng, Trajcompressor: an online map-matching-based trajectory compression framework leveraging vehicle heading direction and change, IEEE Transactions on Intelligent Transportation Systems (2019).

[109] R. Li, S. Wang, H. Deng, R. Wang, K.C.-C. Chang, Towards social user profiling: unified and discriminative influence model for inferring home locations, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 1023–1031.

[110] L. Chen, X. Fan, L. Wang, D. Zhang, Z. Yu, J. Li, T.-M.-T. Nguyen, G. Pan, C. Wang, Radar: Road obstacle identification for disaster response leveraging cross-domain urban data, Proc. ACM Interact. Mob. Wear. Ubiquitous Technol. 1 (4) (2018) 130.

[111] L. Chen, D. Zhang, G. Pan, X. Ma, D. Yang, K. Kushlev, W. Zhang, S. Li, Bike sharing station placement leveraging heterogeneous urban open data, in: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2015, pp. 571–575.

[112] D. Ferreira, V. Kostakos, A.R. Beresford, J. Lindqvist, A.K. Dey, Securacy: an empirical investigation of android applications' network usage, privacy and security, in: Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks, ACM, 2015, p. 11.

[113] A. Kapadia, D. Kotz, N. Triandopoulos, Opportunistic sensing: Security challenges for the new paradigm, in: 2009 First International Communication Systems and Networks and Workshops, IEEE, 2009, pp. 1–10.

Please cite this article as: S. Zhao, S. Li, J. Ramos et al., User profiling from their use of smartphone applications: A survey, Pervasive and Mobile Computing (2019) 101052, https://doi.org/10.1016/j.pmcj.2019.101052.