CCF Transactions on Pervasive Computing and Interaction
https://doi.org/10.1007/s42486-019-00011-4

REGULAR PAPER

Investigating smartphone user differences in their application usage behaviors: an empirical study

Sha Zhao[1] · Feng Xu[1] · Yizhi Xu[1] · Xiaojuan Ma[2] · Zhiling Luo[1] · Shijian Li[1] · Anind Dey[3] · Gang Pan[1]

Abstract

Smartphone applications (Abbr. apps) have become an indispensable part in our everyday lives. Users determine what apps to use depending on their personal needs and interests. Users with different attributes may have different needs, making it natural for their app usage behaviors to be different. The differences in app usage behaviors among users make it possible to infer their attributes. Knowing such differences could help improve mobile user experiences by enhancing smart services and devices. In this paper, we present an empirical study of investigating smartphone user differences on a large-scale dataset of recently used app lists from 106,672 Android users from China. We first investigate the user differences in app usage behaviors with respect to their attributes (gender, age, and income level). We find significant differences in app usage frequency, app usage with time context and functions. We then extract corresponding features from app usage records to infer the attributes of each user, and investigate the predictive ability of individual features and combinations of different individual features. We achieve the accuracy of 83.29% for gender, 69.94% for age (four age ranges) and 71.43% for income level (three income levels) with the best set of features, respectively. Finally, we discuss the implications of our findings and the limitations of this work.

Keywords App usage records · Smartphones · User attribute · User studies

1 Introduction

Smartphones have increasingly become an indispensable part of our daily lives, and there have been around 2.5 billion subscribers in 2019, indicating more than one-third (36%) of the world's population is projected to use a smartphone.[1] Smartphones are undoubtedly much more than a simple communication device as before. Users can achieve many imaginable purposes in daily life through mobile applications on smartphones, such as reading, shopping, navigation and entertainment. 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.[2] Abundant applications (abbr. apps) provide useful services in many aspects of modern life. Easy to download and often free, apps can be fun and convenient for playing games, getting turn-by-turn directions, and

✉ Gang Pan
gpan@zju.edu.cn

Sha Zhao
szhao@zju.edu.cn

Feng Xu
fxuzju@zju.edu.cn

Yizhi Xu
yzxu@zju.edu.cn

Xiaojuan Ma
mxj@cse.ust.hk

Zhiling Luo
luozhiling@zju.edu.cn

Shijian Li
shijianli@zju.edu.cn

Anind Dey
anind@uw.edu

1 Zhejiang University, Hangzhou, China

2 Hong Kong University of Science and Technology, Hong Kong, China

3 University of Washington, Seattle, USA

[1] https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.

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

Published online: 20 August 2019

accessing news, books, weather, and more (Cao and Lin 2017).

Apps on smartphones can be considered as one's entry to access services. Smartphone users install and use apps depending on their needs, preferences, habits, etc. Users with different attributes may have different needs or interests, making it natural for apps installed on smartphones and app usage behaviors of different users to be distinct. For example, young parents are likely to use child rearing related apps more frequently, and users who work in financial sectors are likely to be interested in reading more stock related news. Even for the same app, its usage can be different across users like frequency or intensity of interaction. These differences in smartphone apps enable it possible to infer user personal attributes, such as demographics, interests, or needs. Moreover, a smartphone is usually linked to an individual user, and apps on it have been assumed to achieve greater personalization. Thus, smartphone apps can effectively convey lots of one's personal information. This is providing us a great opportunity to explore smartphone user differences and even infer user attributes, so as to well characterize smartphone users.

The ability to draw connections between users' personal information and behavioral aspects derived though smartphone data could lead to designing and applying machine learning methods to understand user characteristics as well as individual difference and similarity among users. Such knowledge could be used in various ways in the context of devices, services, and mobile applications that are tailored to the individual needs and preferences of a user. In particular, it can be leveraged to enhance the personalization of applications, such as personalized web search, personalized recommendation, targeted advertising, and smart environments. Services can be recommended to users according to their needs, interests or habits. Advertisements can be actively pushed to targeted users. Devices in a smart environment can adjust adaptively according to users' interests, preferences and needs. Mobile app developers, mobile phone manufacturers and mobile carriers could design and customize mobile apps, phones and pre-installed apps to improve user experiences. Besides, dimensions about user characteristics discovered from smartphone apps could be used to improve current user models. From the view points of users themselves, they objectively understand their behaviors derived from smartphone apps so that they could curb poor behavior patterns to improve life quality.

There are some types of data related to smartphone apps that have been explored (Zhao et al. 2019a), such as installed app lists (apps installed on a smartphone, e.g., Zhao et al. 2017a; Seneviratne et al. 2015), app installation behaviors (installation, updating and uninstallation) (e.g., Liu et al. 2018; Li et al. 2015a) and app usage records (e.g., Zhao et al. 2016; Yu et al. 2018). Among the three types of data, app usage records are a better reflection of what activity users perform, what they truly needs or what they look like. There is a major limitation of installed app lists is that, whether one user has installed an app may be a weak indicator of whether he/she actually needs the app (Frey et al. 2017; Xu et al. 2016b, a; Zhao et al. 2017a). 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 Rivron et al. (2016), only 10% of apps were used 80% of the time, suggesting that a lot of apps are downloaded but not used regularly. For the app installation behaviors, especially the updating, 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 (Li and Lu 2017). Compared with installed app lists and app installation behaviors, 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.

In this paper, we present an empirical study of investigating user differences with respect to different user attributes (gender, age, and income level) from their app usage records. We try to answer the following research questions that guide the remainder of this paper:

- *RQ1* Can different user attributes affect app usage behaviors? What differences of the users with different attributes can be explored from their app usage behaviors?
- *RQ2* Can user attributes be reliably inferred from their app usage behaviors? Which feature(s) are the most predictive? What is the best combination of features for building the attribute prediction model? Are the most predictive features the same for inferring different user attributes?

In order to answer these research questions, we conduct experiments based on a large-scale dataset of app usage records collected from 106,672 users from multiple provinces in China. For each smartphone, the dataset contains hourly updates on the ten most recently used apps, for the month of September 2015. We first investigate the user differences in app usage with respect to different user attributes (gender, age and income). We then extract features from app usage records and build classifiers to infer user attributes. Our key findings are as follows.

1. Different user attributes affect users' app usage behaviors on smartphones. Users with different attributes (e.g., females and males, low income level and high income level) have distinctive differences in app usage behaviors in terms of the usage frequency, the usage along with time context and app functions.
2. Features extracted from app usage behaviors can be used to infer the gender, age and income level of its user. We

investigated the predictive ability of individual features and the combination of different individual features for gender, age, and income level. We found, for different attributes, the most powerful individual features are different. We achieved the accuracy of 83.29% for gender with the combination of all the features, 69.94% for age (four age ranges) using the individual feature of app usage with time context, and 71.43% for income levels (three income levels) using the individual feature of app usage with time context.

2 Related work

Recently, a growing number of analyses have sought to understand users based on various cues, such as word use (Fast and Funder 2008; Wei et al. 2017), audio signals (Mairesse et al. 2007), web search logs (Herring and Paolillo 2006; De Bock and Van den Poel 2010), and social network sites (Li et al. 2015b). Compared with these cues, apps on smartphones are inclined to be more personalized, since a smartphone is not only possessed by the same user but also going everywhere with the owner. It promotes emerging research on profiling users with smartphone apps.

There have been some studies using smartphone apps to infer user personal information. For example, demographic attributes (e.g., gender, region and marital status), interests, personality traits and life stages have been learned from app lists installed on smartphones, app installation behaviors (installation, updating and uninstallation) and app usage behaviors (Chittaranjan et al. 2011, 2013; Frey et al. 2015, 2017; Jesdabodi and Maalej 2015; Malmi and Weber 2016; Qin et al. 2016; Rivron et al. 2016; Seneviratne et al. 2015; Tu et al. 2019; Wang et al. 2015, 2018; Xu et al. 2011, 2016b, a; Zhao et al. 2016, 2017a, b, c, 2018, 2019b, c; Li et al. 2015a; Mo et al. 2012; Brdar et al. 2012; Ying et al. 2012; Andone et al. 2016; Peltonen et al. 2018; Zou et al. 2013; Yu et al. 2018; Ouyang et al. 2018; Wang et al. 2019; Böhmer et al. 2011; Liu et al. 2018). In this section, we will review the related work in three aspects: inferring demographics, explaining personality, and discovering life patterns.

2.1 Inferring demographics

Apps on smartphones were used to infer users' demographic attributes (Seneviratne et al. 2015; Xu et al. 2016b; Zhao et al. 2017a; Qin et al. 2016; Malmi and Weber 2016; Wang et al. 2015). For example, Seneviratne et al. inferred about 200 users' gender from their installed app lists, with an accuracy around 70% (Seneviratne et al. 2015). Qin et al. inferred users' gender and age range by leveraging the differences on app usage behaviors of 32,660 users, with the accuracy of 81.12% and 73.84%, respectively (Qin et al. 2016). Malmi et al. analyzed the used app lists of 3760 Android users, and inferred gender and income using logistic regression with an accuracy of 82.3% and 60.3%, respectively (Malmi and Weber 2016). Zhao et al. mined user attributes (e.g., gender, income, preference) from installed app lists by using SVM (Zhao et al. 2017a), with an average equal error rate of 16%. It was shown that user attributes have a significant impact on the adoption and usage of apps on smartphones. For example, the users with a higher income level use the apps of traveling and online shopping much more frequently (Zhao et al. 2018).

2.2 Explaining personality

The correlation between one's personality and his/her app usage behaviors was analyzed (Chittaranjan et al. 2013; Xu et al. 2016b). For example, Chittaranjan et al. investigated the relationship between app usage behaviors derived from rich smarphone data and self-reported Big-Five personality traits (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience) (Chittaranjan et al. 2013). They analyzed the app usage records on the Nokia 95 smartphone from 83 participants over 8 months, and found that the usage of all the apps, except the use of Maps, Camera, Chat and Game applications significantly explained variance in the traits. The Mail application was more likely to be used by Neurotic and Conscientious participants. Introverts were less likely to use Internet applications on the phone, and conscientious individuals were less likely to use the applications of Audio, Video and Music. They also classified users' Big-Five traits with an accuracy of 75.9% using app usage behaviors. Xu et al. also explained the adoption of thirteen mobile apps by using the Big Five personality traits (Xu et al. 2016b). It was found that conscientiousness has negative and significant effect on the adoption of services like photography, media and video, and location-based services, explained by their goal-driven nature and decreased use of leisure services to have fun.

2.3 Discovering life patterns

User information related with their daily life was detected from smartphones' apps. In particular, life events, such as *first car*, *first job*, *marriage*, *first apartment*, and *first child*, were predicted with an average accuracy of 64.5%, based on the analysis of app installation behaviors of more than 2000 users (Frey et al. 2015). The proposed method was based on app name keywords to predict life events, extracting manually up to four of the most frequent keywords from the names of apps related to a specific life event and installed by participants with the life event. The life event *first child* reached the highest accuracy of

Table 1 Sample of lists of recent app tasks in the dataset

User ID	Time	The list of recent app tasks
0000751aecb005a2	2015/9/1 9:09	com.android.calendar, com.tencent.mobileqq, com.moji.mjweather
0000751aecb005a2	2015/9/1 10:09	com.miui.home, com.android.incallui, com.android.calendar, com.moji.mjweather

93.5%, which was understandable because of the fact that there were many useful apps especially for the life event. Frey et al. (2017) investigated the relationship between one user's installed app lists on smartphones and her current life stage, such as teenager living with parents and couple without children. The app adoption rate of each life stage, and the comparison between any two neighboring life stages were analyzed. They found that the app adoption rate in different life stage is different. Zhao et al. discovered different kinds of smartphone users by analyzing the app usage in different time periods from 106,672 Android users (Zhao et al. 2016). The users in each cluster have distinct habits. For example, 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.

To summarize, app usage records provide us with a great opportunity for inferring user personal attributes. Based on the large-scale dataset of app usage records from 106,672 Android users from China for the month of September of 2015, we first investigate gender, age, and income differences in app usage in terms of usage frequency, and usage along with time context and app functions. Then, we extract features from such differences for user representation, and investigate the predictive ability of individual features and combinations of different features for different user attributes. Finally, we discuss the implications and limitations of our study. We will describe our dataset and how we analyze it.

3 Dataset overview

The dataset we use to identify user groups contains lists of recent apps used on Android smartphones, provided by a mobile Internet company in China. It contains 106,762 unique smartphones and 77,685 unique apps from Sept. 1st, 2015 to Sept. 30th, 2015. The data was collected approximately every hour using the function *ActivityManager.getRecentTasks()*. It returns a list of the tasks that the user recently launched, ordered from most recent to oldest. The dataset consists of 52,872,129 usage records in total. A sample of the dataset is shown in Table 1, with each record consisting of a:

- User ID: the unique identity of the sampled smartphone. Each user ID is anonymized for security and privacy reasons before the data is collected.
- Time stamp: the time when the list of tasks was collected.
- List of recent tasks: each list consists of up to ten package names that can be used to identify an app.

In order to get a high understanding of the apps used by uses in the dataset, we categorize all the apps into 29 *categories* (Zhao et al. 2016). In addition, we crawl the *description* of the apps in the dataset from appstore websites.

3.1 Pre-processing

To give a sense of the richness of the dataset, about 60,000 users have 30 days of data from Sept. 1st to Sept. 30th, and about 90% of users have more than 20 days. 25% of the days have 24 records (i.e., complete data collection), and about 80% have more than 15 records. Each record can contain 1–10 data points, and about 30% of the records consist of ten apps.

In order to detect which apps were used in each hour slot, we perform a comparison of two consecutive lists on the raw dataset (Zhao et al. 2016). By doing so, it can be known the usage frequency of each app. Table 2 shows a sample of one user's app usage records extracted from the raw data, where each row represents the apps used in the corresponding hour slot on Sept. 1st, 2015. As we can see from Table 2, the user sample has nine records on Sept. 1st, for instance, from 8 to 9 a.m., he/she used the apps of "com.miui.home" (MiUI Home), "com.tencent.qqmusic" (QQ Music), and "com.tencent.mm" (WeChat). For all the users, there are 25,208,134 app usage records in total, and per user has 243.2 records in average (standard deviation: 141.5) in the 30 days. It means that per user has around eight app usage records 1 day, i.e., 8 active hours for app usage.

According to our observation, we find that there are a few users that have very few records. We calculate the cumulative distribution function of users in terms of the number of app usage records, shown in Fig. 1. The x-axis is the number of app usage records, and the y-axis is the cumulative distribution function of the users with the exact number of app usage records. It can be seen that, the curve grows quickly from 100 to 500, indicating that most

Table 2 Sample of one user's app usage records

No.	User ID	Date	Hours	Used apps
1	0000256bcd	2015-09-01	08–09	com.miui.home, com.tencent.qqmusic, com.tencent.mm
2	0000256bcd	2015-09-01	11–12	com.android.contacts, com.android.incallui, com.moji.mjweather
3	0000256bcd	2015-09-01	13–14	com.android.article.news, com.tencent.mm, com.tencent.qqmusic
4	0000256bcd	2015-09-01	14–15	com.miui.home, com.tencent.mm
5	0000256bcd	2015-09-01	15–16	com.android.incallui
7	0000256bcd	2015-09-01	17–18	com.tencent.mm, com.autonavi.minimap
8	0000256bcd	2015-09-01	20–21	com.tencent.mm, com.taobao.taobao, com.android.AlipayGphone
9	0000256bcd	2015-09-01	22–23	com.moji.mjweather, com.android.article.news, com.tencent.mm

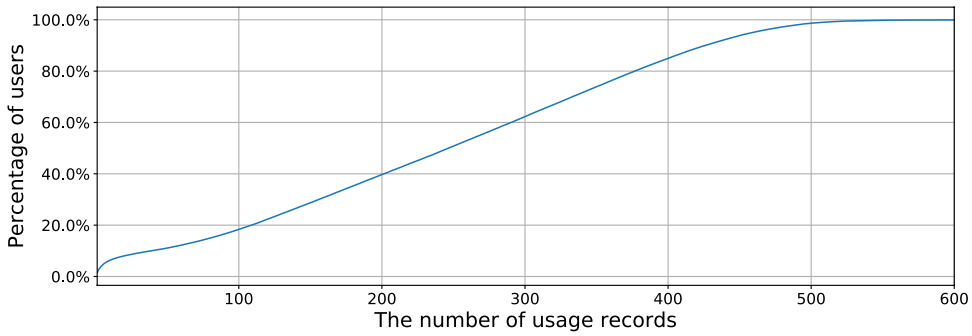


Fig. 1 The cumulative distribution function of users in terms of the number of app usage records

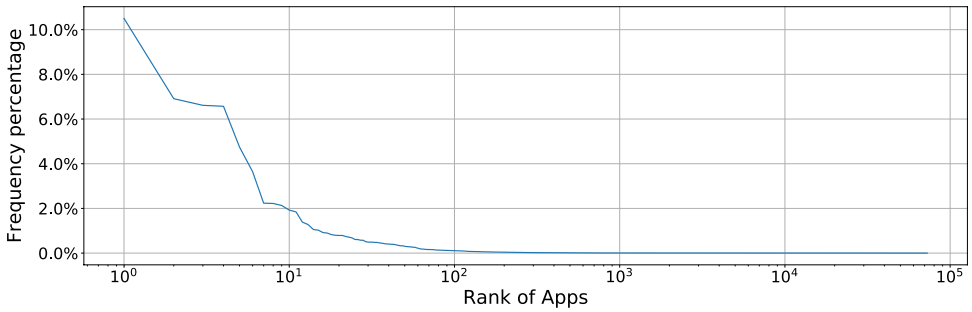


Fig. 2 App distribution in terms of usage frequency percentage

users have 100–500 app usage records in the 30 days. There are very few users (less than 0.06%) having more than 600 app usage records. Around 80% of the users have more than 100 app usage records in the 30 days. We focus our analysis on users who use their smartphones more frequently. We remove those with fewer than 100 app usage records in the 30 days. After filtering, there are 84,810 users remaining and 72,679 unique apps. For these remaining users, there are 24,477,970 app usage records in total, and per user has 288.6 records in average (standard deviation: 113.5) in the 30 days. For these users, each one has 9.6 app usage records in 1 day, i.e., around 10 active hours in which they use apps. The following experiments are conducted on these remaining user samples.

3.2 Basic analysis

Based on the filtered dataset, we do some statistics. The top five most frequently used apps are WeChat, Phone, QQ (an IM client), Contacts, and SMS, which are used for communication and social activities. Per user uses apps for about 1262.5 times and around 43.3 unique apps in 1 month in average. We also calculate the distribution of the apps used by users, as shown in Fig. 2. The x-axis is the logarithm of app rank, and the y-axis is frequency percentage of the corresponding app. We find app frequency to have an unsurprisingly long tail, suggesting there are few apps that were used with very high frequency but most apps were launched for very few times. It is similar to the finding in Rivron et al. (2016), that around 12% of the installed apps

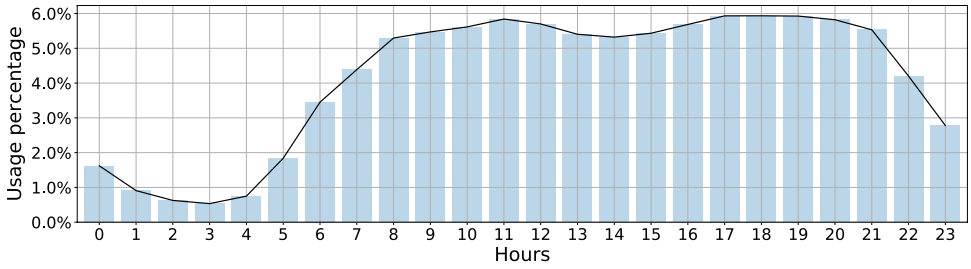


Fig. 3 App usage frequency across 24 h

Table 3 Top three most frequently used apps in each category

Category	Top 3 frequently used apps
SON and IM	WeChat(微信), QQ(QQ), Microblog(微博)
Browser and searching	UC Browser(UC 浏览器), QQ Browser(QQ 浏览器), Baidu(百度)
Phone and SMS	Phone(电话), Contacts(通讯录), Messages(短信)
News and reading	Tencent News(腾讯新闻), Headline Today(今日头条), Shuqi Novels(书旗小说)
Shopping	Taobao(淘宝), Meituan Groupbuy(美团), JD(京东)
Photography and beauty	Gallery(图库), Camera(相机), Kwai Video(快手)
Transportation	DiDi(滴滴打车), Mobile Ticket(铁路 12306), Uber(优步)
Travel	Qunar Travel(去哪儿旅行), Ctrip(携程旅行), Tongcheng Trip(同程旅行)
Car	Cubic Autohome(汽车之家), Autohome Mycar(违章查询助手), Violation Search(车轮违章查询)
Navigation	Baidu Map(百度地图), AMap(高德地图), Tencent Map(腾讯地图)
Parent and child	Babytree Pregnancy(快乐孕期), Baby Time(亲宝宝), Pregnant Housekeeper(怀孕管家)
Education	Calculator(计算器), Hundred Words(百词斩), Youdao Dictionary(有道词典)
Theme	MiLocker(小米百变锁屏), 91 Desktop(91 桌面), Xperia Home(Xperia 主页)
Launcher	Launcher(启动器), Huawei Desktop(华为桌面), OPPO Desktop(OPPO 桌面)
Weather	Moji Weather(墨迹天气), Meizu Weather(魅族天气), Weather(天气)
Clock	Clock(时钟), Alarm Clock(闹钟), Worldclock(世界时钟)
Calendar	Calendar(日历), Chinese Perpetual Calendar(中华万年历), Perpetual Calendar(万年历)
Lifestyle	QQdownloader(应用宝), Baidu Mobile Assistant(百度手机助手), 360 Mobile Assistant(360 手机助手)
Health and fitness	Meet You Period Tracker(美柚), Ledongli Sport(乐动力), Yuedong Sport(悦动圈)
Music and audio	Kugou Music(酷狗音乐), QQmusic(QQ 音乐), TTPOD(天天动听)
Media and video	IQIYI Video(爱奇艺视频), Tencent Video(腾讯视频), Youku Video(优酷视频)
Stock	Straight Flush(同花顺), East Money(东方财富), Dazhihui(大智慧)
Finance	Alipay(支付宝), China Construction Bank(中国建设银行), Mymoney Account(随手记理财记账)
Business	Mail(邮件), WPS Office(WPS), QQmail(QQ 邮箱)
Game casual and puzzle	Happyelements(开心消消乐), Wematch(天天爱消除), Candy Crush Saga(糖果传奇)
Game card and chess	Joy Doudizhu(欢乐斗地主), Doudizhu(JJ 斗地主), Tencent Mahjong(腾讯欢乐麻将全集)
Game other	Timi Run Everyday(天天酷跑), LOL Mobile(掌上英雄联盟), Airplane War(全民飞机大战)
System tool	System UI(系统 UI), Settings(设置), Packageinstaller(包安装器)

are used 80% of the time, suggesting that lots of apps are not used regularly.

We also investigate the app usage across 24 h, shown in Fig. 3. The x-axis is 24 h, and the y-axis is the percentage of the apps usage in the corresponding hour over the usage in 24 h. As we can see, the curve goes up and down across 24 h. There are two peaks at 12 p.m. and 5 p.m., respectively, when users use their smartphone the most. From 1 to 3 a.m., the usage percentage is very low, when most of our users are presumably sleep. The steepest curve going up happens from 4 to 8 a.m. in early morning, indicating smartphone usage increases dramatically in this period. Generally, users wake up and start to use their smartphones. The smartphone usage slightly decreases from 12 to 2 p.m., when users probably take a short break around noon. Users likely go to sleep from 9 to 0 a.m., causing a sharp decline in app usage from 9 p.m. to 0 a.m.

The top five popular categories are SON_and_IM (social online network and instant messaging), Phone_and_SMS, Launcher, System tool and Theme (e.g., screen locker, screen protector, desktop, wallpaper apps). To have a deep look into each category, we list the top three frequently used apps for each category, shown in Table 3.

3.3 Demographic attributes

Demographic data about each user was collected, including gender, income level, and age range. There were three income categories: low income [monthly income $\leq$ 3000 CNY (460 USD)], high income [monthly income $\geq$ 10,000 CNY (1535 USD)], and medium income. There were four age categories: 0–17, 18–24, 25–34, and 35+.

The demographics for the remaining 84,810 users are shown in Table 4. There are more female than male users

Table 4 The percentage of the remaining users in the dataset in each demographic

Percentage of users in each demographic attribute (%)								
Gender		Age range				Income level		
Female	Male	0–17	18–24	25–34	35+	Low	Medium	High
59.5	40.5	4.4	37.9	36.3	21.4	29.8	38.6	31.6

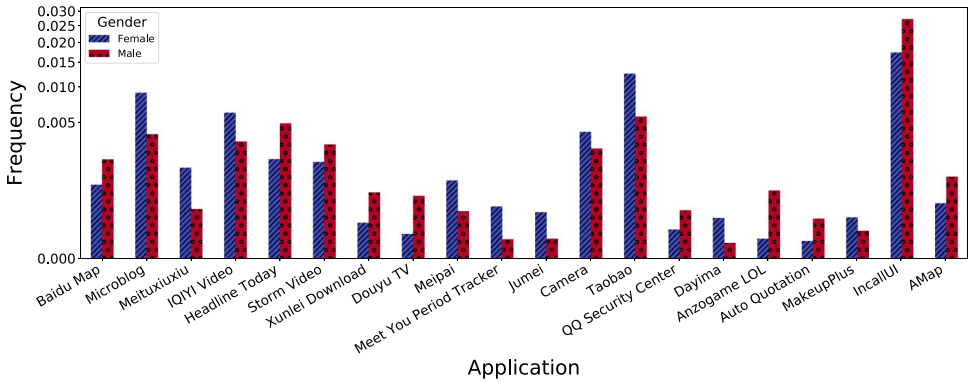


Fig. 4 Top 20 apps with the greatest differences in usage frequency between females and males

in our dataset (59.5% vs. 40.5%). Only 4.4% of users are in the age range of 0–17, with most users being adults. Most users are in the age range of 18–34, accounting for 74.2%. The income levels of the users are almost evenly distributed.

4 User differences in app usage frequency

We explore the user differences in terms of app usage frequency with respect to gender, age and income level.

4.1 Gender differences in app usage frequency

We first investigate the gender differences in app usage frequency in the individual app level. Here, we employ the method of Gradient Boosting Decision Tree (GBDT) (John Lu 2010) to select the apps for which the difference in usage frequency between females and males is significant. GBDT measures the significance of each app by retrieving its significance score after boosted trees being constructed. A score indicates how useful an app is in the construction of the boosted decision trees. The higher the significance score is, the greater the differences in the usage frequency between females and males are. We rank the apps according their significance score, and the first app is considered to be the one which has the most significant differences in the usage frequency between females and males. The top 20 significant apps for gender are shown in Fig. 4.

As we can see from Fig. 4, the app with the most differences in usage frequency between females and males is Baidu Map, a navigation app, for which the average usage frequency of male users is much bigger than that of females. The second one is Microblog app, for which females use more frequently in average. The third one is Meituxiuxiu that provides services for photography about taking pictures with smart beautifying functions and sharing photos. Besides, it is found that males more frequently use the apps related to news, downloading tools and live streaming videos for online games (e.g., Douyu TV), cars (e.g., Auto Quotation) and games [e.g., League of Legends (LOL)], while females more frequently use apps related to photography (e.g., Meipai, MakeupPlus), shopping (e.g., Taobao, Jumei, an app for fashion e-commerce with clothing and cosmetics targeting females), videos for drama (e.g., iQIYI Video), and period tracker (e.g., Meet you, Dayima). The differences of usage frequency of such apps between females and males indicate that gender has an influence on app usage behaviors to a certain degree.

We then explore the gender differences in app usage frequency in the category level. For each category, we calculate the average usage frequency of females and males, respectively, and then calculate the difference in the average usage frequency between females and males. The 28 categories (the category of 'Others' is not included) are ranked in a descending order according to the differences, shown in Fig. 5. The category with the greatest difference in apps usage frequency between female and male users is Car, which males use more frequently, and the second category is Parent_and_child that is more frequently used by females. Similar to the findings in the individual app level, males more frequently use the apps in the categories of transportation, navigation, stock, business, finance, news_and_reading, and browser_and_searching, while females use the apps in the categories of shopping, photography_and_beauty, and

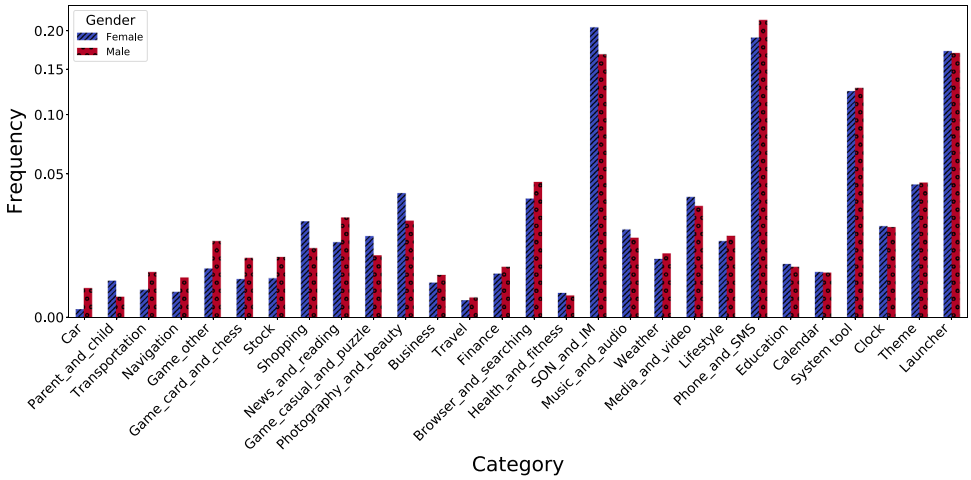


Fig. 5 The differences in categories between females and males

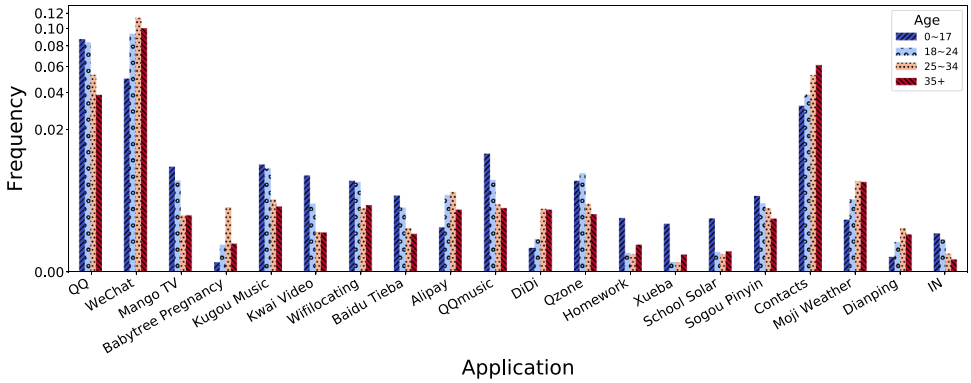


Fig. 6 Top 20 apps with the greatest differences in usage frequency among users in different age ranges

SON_and_IM more frequently. For the types of games, males prefer to play games in Game_card_and_chess and Game_other (e.g., role playing games, action games, simulation games, and adventure games), while females play casual and puzzle games more frequently.

4.2 Age differences in app usage frequency

The age differences in app usage frequency are also investigated. Similarly, the apps with significant differences in usage frequency among different age groups are found by the method of GBDT, shown in Fig. 6. It is interesting to find that the younger users (0–17 and 18–24) use QQ series apps more frequently, such as QQ (an IM App), QQmusic and Qzone (a online social network). They may use QQ and Qzone for instant message and online social activities in daily life. Compared with younger users, the users in the age range of 25–34 and 35+ use WeChat and Contacts for instant message or contact with others, since they use these two apps more frequently. Moreover, we find that, some emerging apps that are with increasing popularity in recent years, such as Mango TV, Kugou Music, Kwai Video and Baidu Tieba, attract more younger users in our dataset. For these apps, the usage frequency of the users (0–17 and 18–24) is much higher than that of those older than 24, suggesting that younger users may be more easily to adopt new things. Besides, there are also great differences in the usage frequency of the apps that provide specific services. For instance, the users with the age range of 25–34 use the BabyTree Pregnancy (for pregnancy and raising babies) much more frequently than other age groups while the users with the age range of 0–17 do not use the app at all. The two apps about study assistance, Xueba and School Solar that provide services for students, are used more frequently by the younger users with the age range of 0–17 and 18–24 in our dataset.

Similarly, we analyze the age differences in app usage frequency in the category level. There are some categories for which the usage frequency are significantly different among users with different age ranges, for example, the categories related to expenditure and financial activities (e.g., Travel, Transportation, Finance, Stock and Shopping). For instance, the elder users (25–34 and 35+) use apps in the categories of Finance and Stock much more frequently than the younger users in our dataset. But for the categories of Game_other, Music_and_audio, Media_and_video, Theme, Education, Photography_and_beauty, the younger users use more frequently.

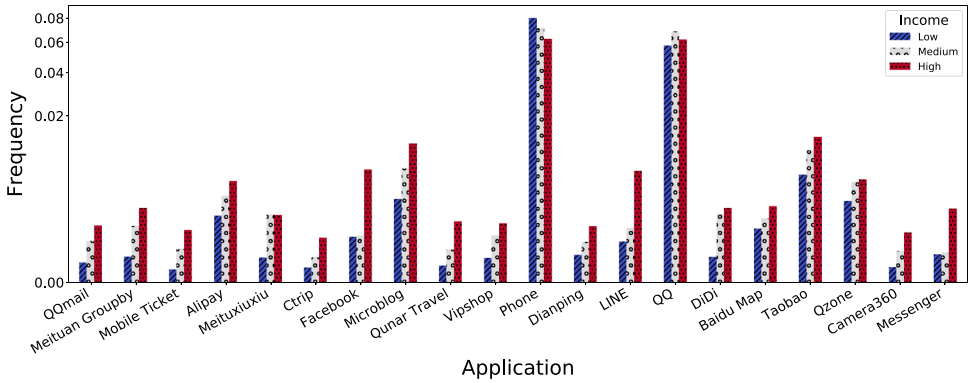


Fig. 7 Top 20 apps with the greatest differences in usage frequency among users with different income levels

4.3 Income level differences in app usage frequency

For the users with different income levels, they also have differences in usage frequency of some apps, shown in Fig. 7. By applying the method of GBDT, we find that users have significant differences in the usage frequency, especially the apps related to expenditure and financial activities (e.g., Meituan Groupbuy, Mobile Ticket, Alipay, Ctrip, Qunar Travel, Vipshop, Taobao). Moreover, for these apps, the usage frequency of the users with higher income level is higher. In addition, the users with high income level use the apps related to social network more frequently (e.g., Facebook, Microblog, LINE, and QQ), while the users with low income level use phone apps more.

The findings in the category level are similar to the ones mentioned above. The users with higher income level use more frequently the apps in the categories of Shopping, Finance, Travel and Business. Compared with the users with higher income level, those with low income level use Phone_and_SMS more. They also use the categories of Game_card_and_chess, Media_and_video more frequently.

5 User differences in app usage with time context

In this section, we investigate the user differences in app usage with respect to time context, e.g., do app usage trends with time differ across users with different attributes? does the usage peak appear in the same hour slot? In order to answer the questions, we calculate the usage percentage of each app on 1 h slot for each user, by dividing the usage frequency of the app on the corresponding hour slot by the sum of the usage frequency of all the apps used on the hour slot. Then, the average usage frequency of each app on 1 h slot for user groups with one specific attribute is obtained. For example, for female users, their average usage frequency of each app on 1 h slot is calculated, by dividing the sum of the usage percentage of each app on 1 h slot for each female user by the number of female users in the dataset.

We first compare the usage along with 24 h over all the apps among user groups, like females and males, users with different age ranges and income levels. As we can see in Fig. 8a, there are slight differences between females and males in the trend of the curves. The usage percentage in each hour slot is similar. In the hours from 0 to 10 a.m. the average usage percentage of females is a littler smaller than that of males, while from 11 a.m. to 11 p.m. the average usage percentage of females is slightly bigger than that of males. For the app usage percentage across 24 h among users with different age ranges, shown in Fig. 8b, that of the 0–17 users are obviously different from the users with the other three age ranges. The app usage percentage of these younger users fluctuates more significantly across 24 h, especially increasing from 10 a.m. to 12 p.m., and declining from 12 to 2 p.m. Such fluctuation is probably because of the school class schedule, since the young users are students. As we can see from Fig. 8c, the app usage percentage across 24 h of the users with medium and high income level is similar, and that of the ones with low income level is slightly different. The users with low income level use apps more frequently from 5 to 9 p.m., compared with the ones with high and medium income level.

As shown in Fig. 8, the differences in app usage percentage across 24 h is not so significant among the user groups with different attributes. But, according to our observation, the users with different attributes, such as females and males, high and low income levels, have significant differences in the usage percentage along with time of some specific apps. Here, we apply the method of GBDT to select the most significant apps for which the users have significant differences in the usage with time context. For each user attribute, we present two example apps to show the differences in app usage with respect to time context.

Figure 9 shows the usage percentage in 24 h of the top two significant apps for females and gender, Meituxiuxiu (a photography app providing services for photography and beautifying) and AutoHome (an app is related to cars). The male users and females perform significant differences in the usage of Meituxiuxiu (see Fig. 9a). As

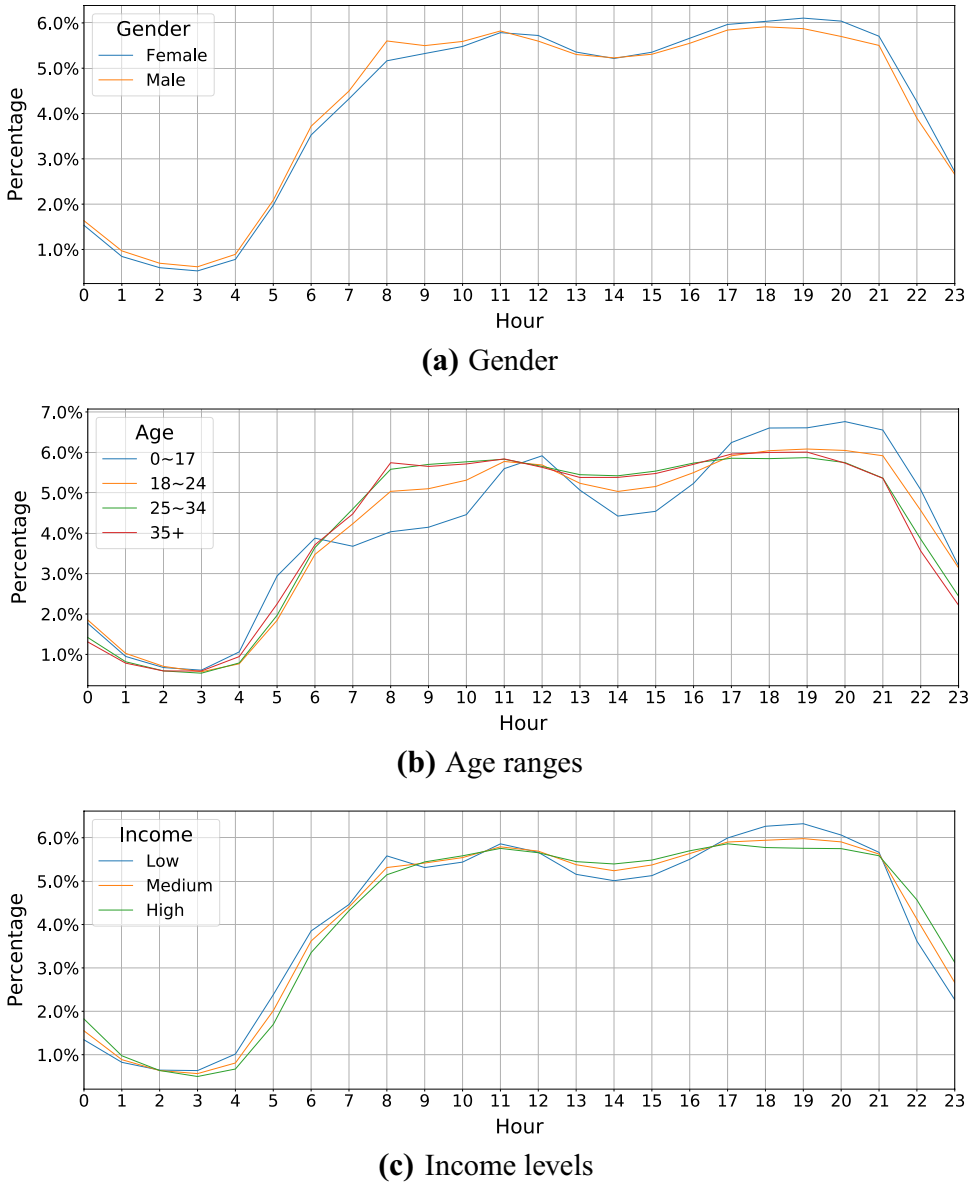
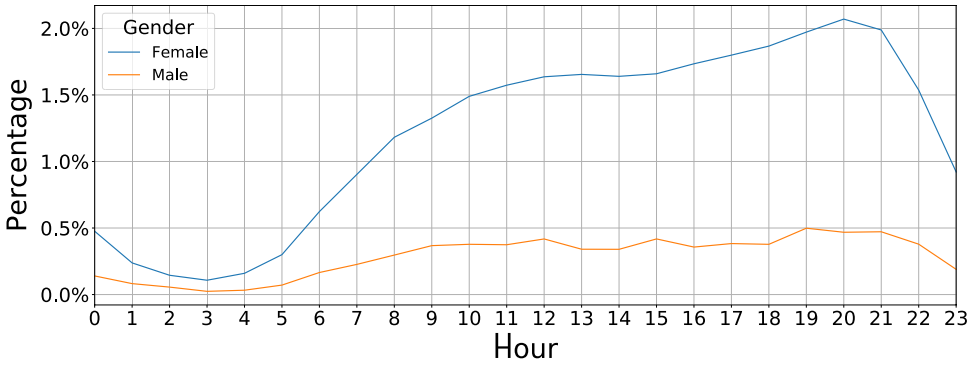


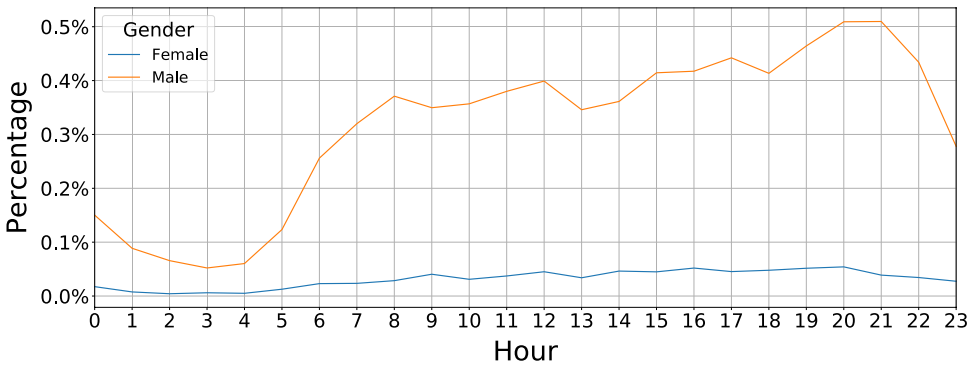
Fig. 8 App usage frequency across 24 h for the users with different attributes: **a** gender, **b** age ranges, **c** income levels

we can see from Fig. 9a, there is a big gap between the two curves, indicating the significant differences in the usage of Meituxiuxiu. Females use the app much more frequently during the whole day except the sleep hours (0–5 a.m.). The females' usage percentage on Meituxiuxiu grows quickly at 7 a.m., then it reaches to an obvious peak at 8 p.m., and begins to decrease at 9 p.m. Compared to females, the males use the Meituxiuxiu app more rarely, and the percentage curve varies smoothly during the whole day. But for the app of AutoHome, the curve trends are just opposite. Males use the AutoHome much more frequently while the females very rarely. The males' usage percentage grows quickly from 5 to 8 a.m., slightly fluctuates from 8 a.m. to 7 p.m., and reaches the peak from 8 to 9 p.m.

Figure 10 shows the top two significant apps, Alarm and Mango TV, to describe the usage differences along with 24 h across users in different age ranges. As shown in Fig. 10a, the usage peak of 0–17 users is 5 a.m., 1 h earlier than the other three user groups. Maybe they need to get up early to go to school. There are usage peaks of alarm at 6 a.m. for both 18–24 and 25–34 users, and the peak is obvious. The peak for the 35+ users is not so obvious, and the usage of alarm at 6 a.m. is just a littler higher than that of at 6 a.m. There are also alarm usage peaks at night. At night, the alarm usage peak of 35+ users appears at 9 p.m., 1 h earlier than the younger users. In addition, the younger users use alarm much more frequently than the 35+ users. For the app of Mango TV, the usage percentage on each

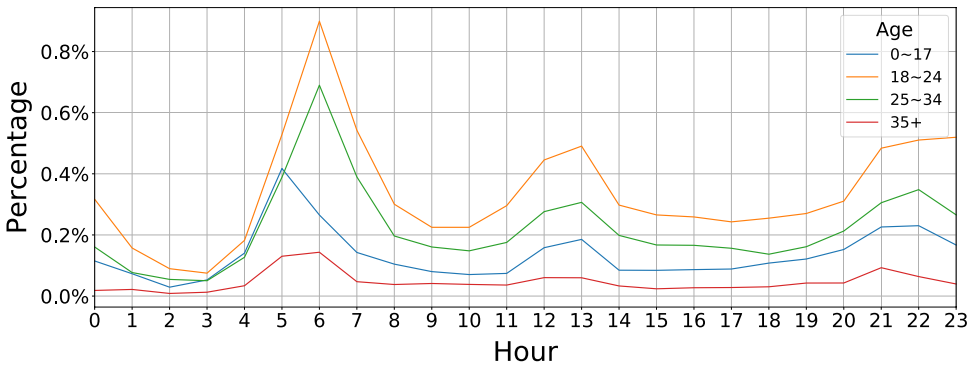


(a) Meituxiuxiu (Photography)

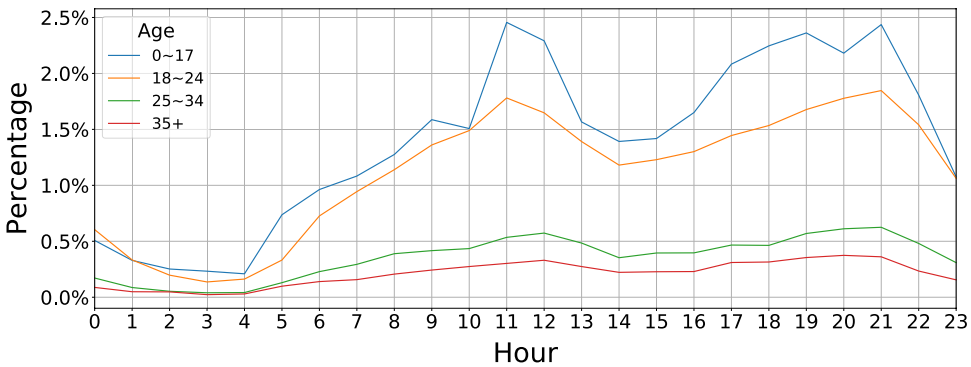


(b) AotoHome (Car)

Fig. 9 Gender differences in the usage across 24 h of apps: **a** Meituxiuxiu (photography); **b** AutoHome (car)

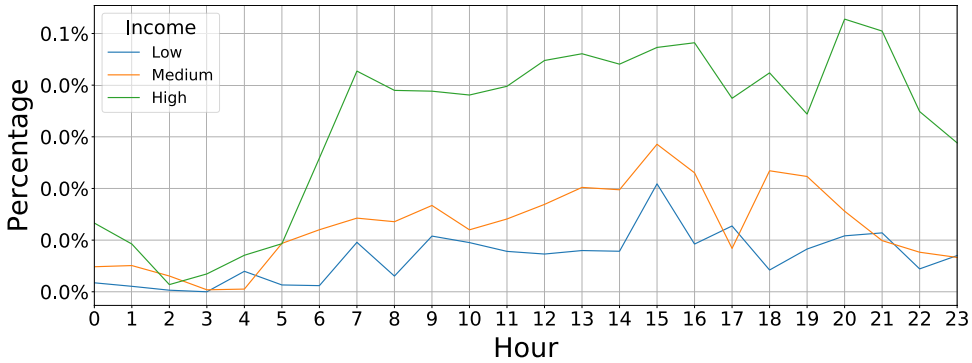


(a) Alarm Clock

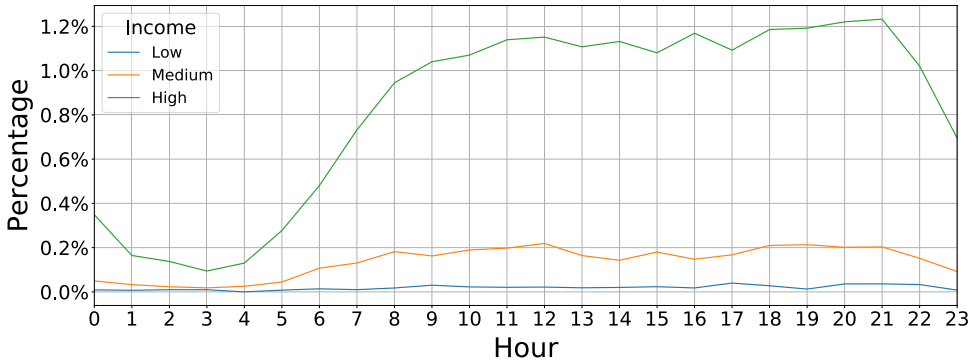


(b) MangoTV

Fig. 10 Age differences in the usage across 24 h of apps: **a** alarm clock; **b** Mango TV



(a) Flight Manager



(b) Qnar Travel

Fig. 11 Income differences in the usage across 24 h of apps: **a** Flight Manager; **b** Qnar Travel

hour slot of both 0–17 and 18–24 users is much higher than that of 25+ users. Mango TV is an app for playing online videos, such as variety shows, drama and cartoon, which attracts more younger users. There are two peaks, at both 11 a.m. and 9 p.m., for both 0–17 and 18–24 users.

We present the usage percentage in 24 h slots of Flight Manager and Qnar Travel, the top two apps with the most significant differences among users with different income levels, shown in Fig. 11. As we can see from Fig. 11a, the users with high income use Flight Manager much more frequently than the users with lower income level on each hour slot, and their usage percentage reaches the biggest value at 8 p.m. at evening. Compared to the high income level, the users with medium and low income level do not use the app so frequently after 8 p.m. at night, and the biggest usage percentage of them appears at 3 p.m. in afternoon. It suggests that the users with high income level are still on travel or business trip at evening or even night. For the app of Qnar Travel, the usage differences among users with different income levels are more significant, shown in Fig. 11b. The usage percentage in each hour slot of the users with high income level is much higher than that of the users with lower income levels. As we can see from Fig. 11b, users' income level is higher, the more frequently they use the app of Qnar Travel.

6 User differences in functions of the used apps

The functions of the used apps are potentially good indicators of gender, since users with different attributes may seek different functions and values in smartphone apps because of different needs and interests.

In order to discover the user differences in app functions at the individual app level, we extract the app functions through discovering the semantic topics from the description text of apps. Each topic indicates one kind of function, such as playing music and taking pictures. In order to learn latent semantic topics, each user is regarded as a document consisting of the words appearing in the description of all his/her used apps, and all the users constitute a corpus. Each user (document) is considered to have a set of topics that can be learned from his/her words. Here, we apply Latent Dirichlet Allocation (LDA) (Blei et al. 2003) to learn the semantic topics from the app description to extract the app function. LDA is a generative probabilistic model of a corpus, of which the basic idea is that documents are presented as random mixtures over latent topics, and each topic is characterized by a distribution over words.

Fig. 12 The word clouds of the top five different topics for gender

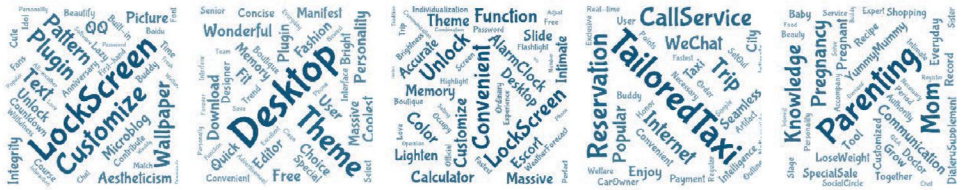


Fig. 13 The word clouds of the top five different topics for age

More specifically, we extract words using Jieba, a tool for Chinese text segmentation, to select nouns, verbs and adjectives, and used words to represent each user as a vector. Formally, a user u is represented by $u = (w_1, w_2, w_3, \ldots, w_j, \ldots, w_m)$, where w_j is the jth word, and its value is the term frequency, dividing the number of times that w_j occurs in the user representation by the count of all of the words. In other words, it refers to the occurring probability of w_j to the user u. Each user is taken as a word probability distribution $P(w|u)$ and input to LDA. LDA assign a user multiple topics with a probability distribution of the topics, indicating the probability that the topic belongs to the user. Each topic consists of a probability distribution of the words, indicating the probability that the word belongs to the topic. When generating a document (user), LDA posits that each word for each user is generated by randomly choosing a topic which belongs to the user in a certain probability, and then choosing the word that belongs to the topic in a certain probability. For a generated document (user), Eq. (1) shows how probably the word w_i is generated:

$$P(w_i|u) = \sum_{j=1}^{k} P(t_j|u)P(w_i|t_j) \tag{1}$$

where k refers to the number of topics. $P(w_i|u)$ is the probability of the ith word to the user u, $P(t_j|u)$ is the probability of the jth topic to the user u, and $P(w_i|t_j)$ is the probability of the ith word to the jth topic.

We obtain 300 topics and each has a probability distribution of words from LDA (the choice of 300 topics is explained in Sect. 7.3.2). In order to understand the differences of users with different attributes in app functions, we compute the significant topics for distinguishing the users in terms of gender, age and income level, respectively. Taking gender for an example, the significance score $s_{i,gender}$ of the ith topic for gender is computed by Eq. (2), then the topics are ranked in a descending order according to the significance score, and the first topic in the ranking is the one that females (or males) are the most different in. The top five different topics are selected for gender. Similarly, the top five different topics are selected for age and income level, respectively

$$\begin{aligned} s_{i,gender} &= \frac{Max(p_{i,female}, p_{i,male}) - Min(p_{i,female}, p_{i,male})}{Max(p_{i,female}, p_{i,male})} \\ p_{i,female} &= \frac{\sum_{j=1}^{N_F} F_{i,j}}{N_F}, \quad p_{i,male} = \frac{\sum_{j=1}^{N_M} M_{i,j}}{N_M} \end{aligned} \tag{2}$$

where $p_{i,female}$ and $p_{i,male}$ are the average probability of the ith topic for each female and each male, respectively, $F_{i,j}$ and $M_{i,j}$ refers to the probability that the ith topic belongs to the jth female and jth male, respectively, and N_F and N_M are the total number of females and males in our dataset.

To give a sense of what semantics the top different topics express, we select the top 30 words with the highest probability for each topic and generate a word cloud. We list the word clouds of the top five different topics for the users in terms of gender, age and income level, respectively, shown in Figs. 12, 13, and 14. For each of the present topics, we also listed the average probability for each user attribute. In each word cloud, the size of one word indicates the probability that it belongs to the topic, with higher probability corresponding to bigger size.

$p_{Low} = 0.0006$ $p_{Medium} = 0.0014$ $p_{High} = 0.0110$ | $p_{Low} = 0.0003$ $p_{Medium} = 0.0006$ $p_{High} = 0.0043$ | $p_{Low} = 0.0012$ $p_{Medium} = 0.0013$ $p_{High} = 0.0157$ | $p_{Low} = 0.0006$ $p_{Medium} = 0.0018$ $p_{High} = 0.0062$ | $p_{Low} = 0.0004$ $p_{Medium} = 0.0036$ $p_{High} = 0.0043$

Fig. 14 The word clouds of the top five different topics for income level

Figure 12 shows the word clouds of the top five most different topics for female and male users, as well as the average probability of each topic for per female and per male. The most different topic for gender is a topic related to games (e.g., LOL), for which the average probability of per male is around ten times bigger than that of per female. The second most different topic is a topic to media related to live streaming, sports, and match, which the male users are more interested in. In addition, the males use the apps related to cars, such as automobile and violation searching, seen in the fifth most different topic. As we can see from the third and fourth topic, female users are more interested in apps related to make-up, taking selfie and sharing photos in online social network, whose average probability for both the third and fourth topic is around six times bigger than males'. Such findings are similar to those gender differences discovered in terms of app usage frequency and app usage along with hour slots.

Figure 13 shows the word clouds of the top five most different topics for the users in different age ranges, and the average probability of each topic. As we can see from Fig. 13, for the top three most different topics, the average probability of 0–17 users and 18–24 users is much higher than that of 25–34 and 35+ users, suggesting that the younger users pay more attention to lock screen, desktop, wallpaper, and theme, especially the 0–17 users. Particularly, the average probability of the these three topics for 0–17 users is around 20 times bigger than that of the 35+ users. On the contrary, the 25–34 and 35+ users have much bigger average probability for the fourth and fifth topics, which convey the semantics of tailored taxi, trip, reservation, WeChat and parenting. These users use apps more frequently related to expenditure (e.g., trip, taxi, reservation), social activities (e.g., WeChat, call services) and raising babies (e.g., parenting, mom). It is reasonable that the 25+ users may have stable earnings to afford the trip or reservation services, and they likely have babies or kids to raise.

The top five different topics for income level is shown in Fig. 14. As we can see, for all the five topics, the income level of is the higher, the average probability of the topic for the user group is the bigger. The users with high income level have the biggest probability for all the five topics, while the ones with low income level have the smallest probability. Compared to the users with lower income level, the users with high income level are more interested in online social network (e.g., Line, Instgram, Facebook, Chat) and use the apps related to ticket purchase, tailored taxi and reservation more frequently. Such findings are also reflected in the aspects of app usage frequency and app usage along with time context.

7 The predictive ability of app usage records

In this section, we first extract features from app usage records based on the user differences mentioned above, and then investigate the ability of individual features and combinations of different features to predict user attributes: gender, age and income level, respectively.

7.1 Features extracted from app usage records

7.1.1 App-based user representation

Given the significant differences in used apps in terms of usage frequency, we intuitively exploit the used apps to represent users for inferring user attribute. In detail, we take each used app as a dimension and represent each user as an app-based vector. If an app is used, the corresponding value of its dimension is set to 1, and if not, the value is 0. Formally, a user u is represented by $u = (a_1, a_2, a_3, \ldots, a_k, \ldots, a_m)$, where a_k is for the kth app, and it has two values, 1 and 0, for indicating whether the app is used or not. In this case, u will be very sparse, since most users only use very few apps compared to the complete set of apps in our dataset. Not all the apps are useful for describing a user. If an irrelevant or redundant app is removed, it will not affect attribute inference. In order to increase the computational efficiency, we use the discriminative apps selected by GBDT to compactly represent each user.

7.1.2 App-time based user representation

Users determine what apps to use is usually related to temporal context. App usage behaviors on smartphones have been shown to exhibit specific temporal pattern (Jesdabodi and Maalej 2015; Xu et al. 2011; Zhao et al. 2016). For example, SMS and Phone are shown to have an evenly distributed pattern, whereas apps like news or weather apps are used more frequently in morning hours. Thus, we introduce temporal feature for attribute prediction. More specifically, we focus on the apps that are frequently used in our dataset. We select top 10,000 frequently used apps, and take 24 h slots into consideration. We use the app usage times in each hour slot to represent one user as a vector of 10,000 (*frequently used apps*) × 24 (hour slots) for a total of 240,000 dimensions. Formally, a user u was represented by $u = (at_1, at_2, \ldots, at_i, \ldots, at_{240,000})$, where at_i means the usage frequency of one app in the corresponding hour slot. In this way, the user representation vector is dramatically long and sparse. We also apply GBDT to select the top app-time based features to represent each user to improve computational efficiency.

7.1.3 App usage sequence based user representation

Apps on smartphones are used in order. We treat a series of apps used in a certain period as a sequence. Apps are often used in conjunction with other relevant apps to serve one need. For example, if a user launches the 'eBay' app, the next app is likely to be the 'PayPal'. The app sequence indicates one's need to a certain degree. Considering users with different demographic attributes could have different needs, we try to capture the characteristics of app usage sequence from app usage behaviors to build user representation.

We apply Doc2Vec (Le and Mikolov 2014) to model the app usage sequence, to learn the user representation. Doc2Vec predicts the next word by exploring a paragraph and a word sequence in a given context in the paragraph. More specifically, every word is mapped to a unique vector, as well as each paragraph. Word vectors are averaged, concatenated, or summed as a feature vector that is concatenated with the paragraph vector for predicting the next word. Taking the analogy to word and document modeling, we can treat each user as a document and each app as a word, to model the app usage sequence. The user and words are embedded in vectors. During the training procedure, user and word vectors are updated until convergence. By doing so, one user representation vector is obtained, which explicitly encodes many app usage patterns. We can feed the user representation vector directly to classifiers for demographic attribute prediction. Figure 15 illustrates Doc2Vec model based on app usage sequence, where Tom's app usage sequence of 'eBay', 'WhatsApp' and 'eBay' is input to predict 'PayPal'.

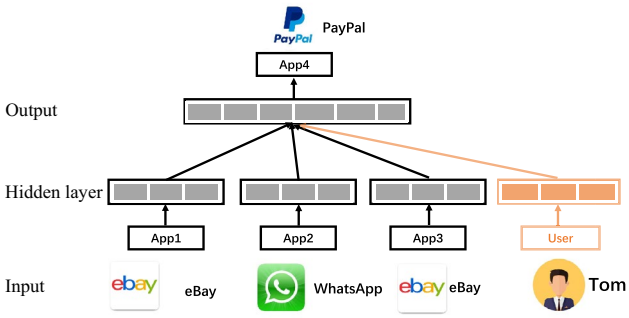


Fig. 15 Illustration of Doc2Vec model based on app usage sequence

7.1.4 Category-based user representation

Based on the usage differences in 29 app categories among users with different attributes, we convert the differences to features for user representation. To be specific, we represent each user's app usage using the categories and usage percentage in different hours. Thus, each user is represented by a vector of 29 (*categories*) × 24 (*hours*) for a total of 696 dimensions. A user was formally represented by $u = (c_1, c_2, \ldots, c_i, \ldots, c_{696})$, where c_i is the usage percentage of one category in the corresponding hour over all the app usage.

7.1.5 Topic-based user representation

App descriptions indicate apps' key functions reflecting users' needs and interests. Users with different demographic attributes may seek different functions in smartphone apps based on their different needs and interests. Considering the differences in app functions conveyed by topics that were learned from app descriptions, we applied the n topics to represent each user as a vector. One user can be represented by $u = (d_1, d_2, \ldots, d_k, \ldots, d_n)$, where d_k is the kth topic, and the value is the probability the topic belongs to the user.

7.2 Implementation and performance metrics

We trained different classifiers to infer gender, including LR (Logistic Regression) (Hosmer et al. 2013), GBDT (Gradient Boosting Decision Tree) (He et al. 2014), and DNN (Deep Neural Network) (LeCun et al. 2015) using different features. In our DNN model, features were input into a wide layer, followed by three hidden layers of fully connected Rectified Linear Units (ReLU). There are 16, 32, and 64 neurons on the first, second, and third hidden layer, respectively.

Table 5 The gender classification results

Feature	Classifier	ACC (%)	Precision-macro	Recall-macro	F1-macro
App (1000)	GBDT	76.29	0.7541	0.7525	0.7532
	LR	78.13	0.7740	0.7687	0.7709
	DNN	78.42	0.7768	0.7726	0.7744
Topic (300)	GBDT	**81.43**	0.8080	0.8050	0.8064
	LR	77.82	0.7720	0.7626	0.7662
	DNN	79.29	0.7866	0.7837	0.7843
Sequence (500)	GBDT	**83.01**	0.8253	0.8200	0.8223
	LR	83.08	0.8263	0.8204	0.8229
	DNN	83.36	0.8291	0.8235	0.8259
Category (696)	GBDT	76.03	0.7534	0.7426	0.7463
	LR	74.12	0.7350	0.7182	0.7228
	DNN	75.37	0.7463	0.7382	0.7406
App-Time (1000)	GBDT	80.75	0.8006	0.7986	0.7996
	LR	79.25	0.7870	0.7780	0.7815
	DNN	79.53	0.7890	0.7864	0.7868
All (3496)	GBDT	83.26	0.8283	0.8215	0.8244
	LR	82.68	0.8219	0.8159	0.8184
	DNN	**83.29**	0.8271	0.8254	0.8260

In the training procedure, a cross-entropy loss was minimized with gradient decent on the output of the sampled softmax. We employed a five-fold cross-validation policy. The sampled dataset was randomly divided into five folds as evenly as possible. In each round, four folds were used for training classifiers and the rest for validation. Thus, any user for testing will never simultaneously appear in the training set and testing set. We repeated the procedure five times and report the averages of the tests below.

We used three criteria to measure the performance of the classification: *ACC*, *precision_macro*, *recall_macro* and *F1 score_macro*. ACC refers to the classification accuracy, which is computed by dividing the number of true positive smaples and true negative samples by the number of all the samples in the testing set. A macro-average computes the metric independently for each class and then takes the average (treating all classes equally).

7.3 Results

7.3.1 Prediction results

We investigated the predictive ability of individual features and combinations of different features using different classifiers, for gender, age, and income level, respectively. In particular, for the app-based user representation we used top 1000 significant apps selected by GBDT, for app-time based features we selected the top 1000 important features via GBDT, and for app sequence feature we learned user representation vectors of 500 dimensionality by Doc2Vec model, where four apps were input to predict the next app. The combination of different features was obtained by concatenation operation. Tables 5, 6, and 7 summarize the performance of the classifiers over the samples for gender, age and income level, respectively. The ACC of the top 2 best individual features and the combination of all the features are highlighted in bold, respectively.

We investigate the predictive ability of individual features and combinations of different features using different classifiers. As we can see from Tables 5, 6, and 7.

1. As shown in Table 5, the best performance for inferring gender is using the DNN model with the combination of all the features, with an ACC of 83.29%, precision of 0.8271, recall of 0.8254, and F1 score of 0.8260, while the category-based feature performs the worst (compared to all the feature sets), e.g., with an ACC of 74.12%, precision of 0.7350, recall of 0.7182, and F1 score of 0.7228 for an LR model. The top two most powerful individual features for distinguishing gender are sequence-based and topic-based features.
2. As we can see from Table 6, the best performance for inferring age is using the GBDT model with the individual feature, app-time based feature, with an ACC of 66.94%, precision of 0.6398, recall of 0.5613, and F1 score of 0.5820, while the category-based feature performs the worst (compared to all the feature sets) for an LR model, about 15% lower in ACC than the best performance. The second most powerful individual features for distinguishing age is the sequence-based feature.
3. It can be seen from Table 7, the best performance for inferring income level is using the GBDT model with

Table 6 The age classification results

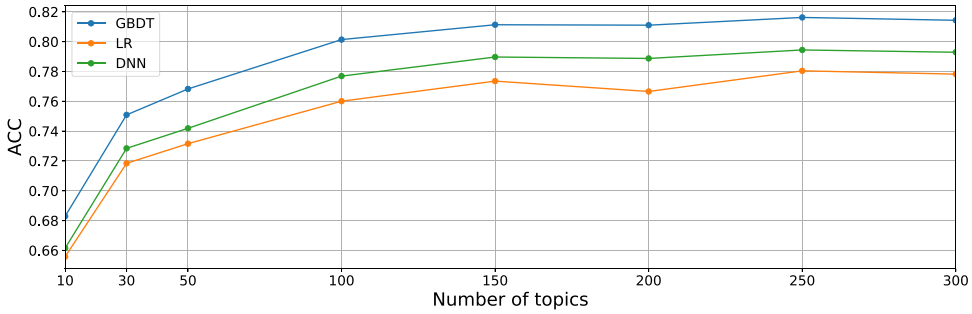
Feature	Classifier	ACC (%)	Precision-macro	Recall-macro	F1-macro
App (1000)	GBDT	61.43	0.5836	0.5031	0.5217
	LR	64.54	0.6302	0.5452	0.5685
	DNN	64.87	0.6226	0.5597	0.5806
Topic (300)	GBDT	65.42	0.6258	0.5650	0.5855
	LR	59.83	0.5700	0.4616	0.4728
	DNN	61.82	0.5858	0.5185	0.5376
Sequence (500)	GBDT	65.43	0.6276	0.5528	0.5749
	LR	65.09	0.6268	0.5398	0.5619
	DNN	**65.92**	0.6289	0.5746	0.5937
Category (696)	GBDT	57.84	0.5580	0.4713	0.4904
	LR	54.95	0.5281	0.4153	0.4215
	DNN	56.10	0.5403	0.4433	0.4573
App-Time (1000)	GBDT	**66.94**	0.6368	0.5613	0.5820
	LR	62.79	0.6143	0.5099	0.5314
	DNN	64.12	0.6242	0.5340	0.5564
All (3496)	GBDT	66.56	0.6431	0.5647	0.5870
	LR	65.64	0.6314	0.5472	0.5678
	DNN	**67.03**	0.6431	0.5895	0.6079

Table 7 The income level classification results

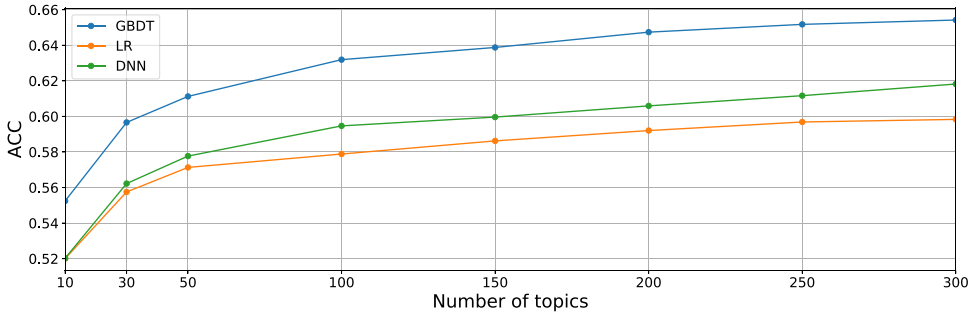
Feature	Classifier	ACC (%)	Precision-macro	Recall-macro	F1-macro
App (1000)	GBDT	64.63	0.6523	0.6545	0.6521
	LR	67.82	0.6798	0.6891	0.6826
	DNN	**69.53**	0.7008	0.7041	0.6998
Topic (300)	GBDT	63.76	0.6484	0.6411	0.6443
	LR	57.23	0.5846	0.5756	0.5787
	DNN	60.76	0.6186	0.6120	0.6141
Sequence (500)	GBDT	64.51	0.6547	0.6491	0.6516
	LR	64.53	0.6465	0.6558	0.6498
	DNN	66.71	0.6736	0.6745	0.6729
Category (696)	GBDT	57.87	0.5912	0.5813	0.5852
	LR	52.82	0.5376	0.5302	0.5332
	DNN	54.92	0.5614	0.5523	0.5554
App-Time (1000)	GBDT	**71.43**	0.7192	0.7220	0.7193
	LR	63.08	0.6369	0.6381	0.6367
	DNN	66.52	0.6712	0.6729	0.6711
All (3496)	GBDT	**69.15**	0.6993	0.6963	0.6975
	LR	65.10	0.6531	0.6599	0.6557
	DNN	68.10	0.6908	0.6874	0.6866

the individual feature, app-time based feature, with an ACC of 71.43%, precision of 0.7192, recall of 0.7220, and F1 score of 0.7193, while the category-based feature performs the worst (compared to all the feature sets) for an LR model, about 20% lower in ACC than the best performance. The second most powerful individual features for distinguishing age is the app-based feature.

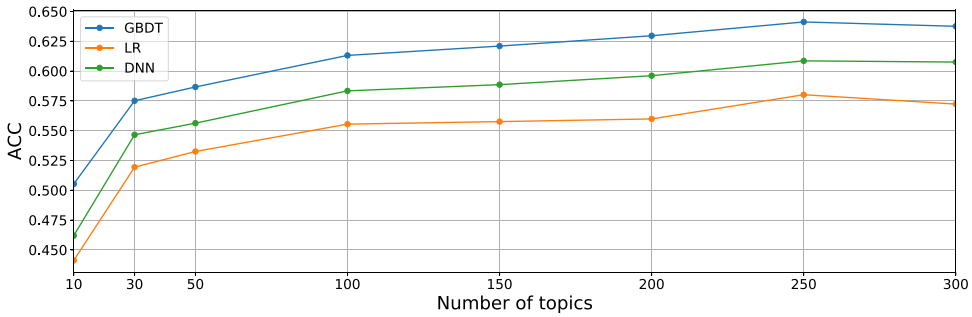
4. For different user attributes, the most powerful features for distinguishing the attributes are different. In general, the app-time based, sequence-based and topic-based features are relatively powerful for distinguishing all the three user attributes. The category-based feature is the most weak indicator for inferring all the attributes.
5. Combining different features do not provide a significant performance improvement. As we can see from the listed tables, the combination of all the features can slightly improve the prediction performance (83.29% vs. 83.01% when singly use sequence-based feature when



(a) Gender



(b) Age



(c) Income

Fig. 16 Performance with varying number of topics for user representation: **a** gender, **b** age, and **c** income

inferring gender; 67.03% vs. 66.94% when solely use app-time based feature for age), or even perform a little worse than the most powerful individual feature (69.15% vs. 71.43% when only use app-time based feature for inferring income level).

7.3.2 Performance with varying number of topics for user representation

As mentioned above, topic-based feature is one of the most powerful individual features for distinguishing user attributes. According to our observation, the number of topics that are extracted from app description and used to represent users has a significant influence on the prediction performance. Thus, we investigate the effect of varying the number of topics for inferring gender, age and income level, respectively. We experimented with 10, 30, 50, 100, 150, 200, 250, and 300 topics and compare the corresponding performance of *ACC*, shown in Fig. 16. As we can see, the performance grows quickly when the number of topics varies from 10 to 150, and changes slightly after 250 for all the three user attributes. That is to say, using less than 150 topics learned by LDA is not enough for optimizing attribute inferring. Using 300 topics for user representation is advisable for inferring gender, age and income level.

8 Discussion

Through our analysis, we have discovered the differences of users with different attributes in app usage behaviors in terms of usage frequency, app functions and app usage time context, and inferred user attributes from their app usage behaviors. We found that demographics like gender, age and income level play a strong role on app usage behaviors.

8.1 Implications

We foresee many opportunities of this study for improving the user experience of smartphone usage. Mobile phone designers, mobile carriers, and application developers can improve services and devices based on the user differences in app usage behaviors. In particular, mobile phone designers can design smartphones that are targeted towards improving the user experience of users with different attributes by providing features that the users may value more than others. For example, males may value an improved GPS sensor since they use navigation apps more frequently. Mobile carriers could allow for the customization of what apps are made available for users based on the most discriminative apps and the top interests. For example, mobile carriers can pre-install apps related to study assistants for students. Application developers can recommend apps to users according to their interests, and provide personalized applications. After inferring attributes from the app usage behaviors, app developers can recommend specific apps that the users with a specific attribute are the most interested in.

8.2 Limitations

Although we can investigate the differences of users from the recent task lists, we must acknowledge the limitations of the dataset used in this study. First, the dataset consisted of recent task lists that were collected once every hour. This low sampling rate can cause us to miss information about app usage. Second, from the recent task lists, we do not know how long each app is used, how often it is used, and in which order the task list changes. This kind of information could be very helpful to more precisely characterize usage behaviors. This was a known tradeoff of using an existing dataset vs. collecting our own, and is one we will address in a future data collection of our own. Third, 1 month is not enough to understand how user attribute affect app usage behaviors and users' behaviors could change over time (Yu et al. 2019). Besides, app usage behaviors could be affected by other context information, such as location (Yu et al. 2018) and network connection (4G or WiFi), while we only analyze the temporal context. Finally, the user attributes studied were limited to demographics of gender, age and income level. In the future work, we will take more user attributes into study.

9 Conclusion

In this work, we have presented an empirical study of investigating smartphone user differences from their app usage behaviors. We conducted our study based on a large-scale dataset from the smartphones of 106,672 Android users from multiple provinces in China. For each smartphone, the dataset contains hourly updates on the ten most recently used apps, for the month of September, 2015. We demonstrated the differences of the users with different attributes (gender, age, and income level) in terms of usage frequency, usage with temporal context and functions. Then, we extracted features from their app usage behaviors to infer the gender, age and income level of each user. Sequentially, we investigated the predictive ability of individual features and combinations of different features. We achieved the accuracy of 83.29% for gender with the combination of all the features, 69.94% for age (four age ranges) using the individual feature of app usage with time context, and 71.43% for income levels (three income levels) using the individual feature of app usage with time context. Finally, we discussed the implications of our findings and the limitations of this work. In the future work, we will explore the differences among other user attributes, such as personality, education level, and occupation, from their app usage behaviors, since our method can potentially be readily extended to a series of other user attributes.

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

References

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

Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

Böhmer, M., Hecht, B., Schöning, J., Krüger, A., Bauer, G.: 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, pp. 47–56. ACM (2011)

Brdar, S., Culibrk, D., Crnojevic, V.: Demographic attributes prediction on the real-world mobile data. In: 2012 Nokia Mobile Data Challenge Workshop (2012)

Cao, H., Lin, M.: Mining smartphone data for app usage prediction and recommendations: a survey. Pervasive Mob. Comput. **37**, 1–22 (2017)

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

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

De Bock, K., Van den Poel, D.: Predicting website audience demographics for web advertising targeting using multi-website click stream data. Fund. Inf. **98**(1), 49–70 (2010)

Fast, L.A., Funder, D.C.: Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. J. Personal. Soc. Psychol. **94**(2), 334 (2008)
Frey, R., Xu, R., Ilic, A.: Reality-mining with smartphones: detecting and predicting life events based on app installation behavior (2015)
Frey, R.M., Xu, R., Ilic, A.: Mobile app adoption in different life stages: an empirical analysis. Pervasive Mob. Comput. **40**, 512–527 (2017)
He, X., Pan, J., Jin, O., Xu, T., Liu, B., Xu, T., Shi, Y., Atallah, A., Herbrich, R., Bowers, S., et al.: Practical lessons from predicting clicks on ads at facebook. In: ADKDD2014, pp. 1–9. ACM (2014)
Herring, S.C., Paolillo, J.C.: Gender and genre variation in weblogs. J. Socioling. **10**(4), 439–459 (2006)
Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X.: Applied logistic regression, vol. 398. Wiley, New York (2013)
Jesdabodi, C., Maalej, W.: Understanding usage states on mobile devices. In: UbiComp2015, pp. 1221–1225. ACM (2015)
John Lu, Z.: The elements of statistical learning: data mining, inference, and prediction. J. Royal Stat. Soc. Series A (Stat. Soc.) **173**(3), 693–694 (2010)
Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: ICML2014, pp. 1188–1196 (2014)
LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
Li, H., Lu, X.: Mining device-specific apps usage patterns from large-scale android users. arXiv preprint arXiv:1707.09252 (2017)
Li, H., Lu, X., Liu, X., Xie, T., Bian, K., Lin, F.X., Mei, Q., Feng, F.: Characterizing smartphone usage patterns from millions of android users. In: Proceedings of the 2015 Internet Measurement Conference, pp. 459–472. ACM (2015a)
Li, S., Wang, J., Zhou, G., Shi, H.: Interactive gender inference with integer linear programming. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, pp. 2341–2347 (2015b)
Liu, X., Li, H., Lu, X., Xie, T., Mei, Q., Feng, F., Mei, H.: Understanding diverse usage patterns from large-scale appstore-service profiles. IEEE Trans. Softw. Eng. **44**(4), 384–411 (2018)
Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. J. Artif. Intell. Res. **30**, 457–500 (2007)
Malmi, E., Weber, I.: you are what apps you use: demographic prediction based on user's apps. arXiv preprint arXiv:1603.00059 (2016)
Mo, K., Tan, B., Zhong, E., Yang, Q.: Report of task 3: your phone understands you. In: 2012 Nokia Mobile Data Challenge Workshop, pp. 18–19. Citeseer (2012)
Ouyang, Y., Guo, B., Guo, T., Cao, L., Yu, Z.: Modeling and forecasting the popularity evolution of mobile apps: a multivariate hawkes process approach. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **2**(4), 182 (2018)
Peltonen, E., Lagerspetz, E., Hamberg, J., Mehrotra, A., Musolesi, M., Nurmi, P., Tarkoma, S.: The hidden image of mobile apps: Geographic, demographic, and cultural factors in mobile usage (2018)
Qin, Z., Wang, Y., Cheng, H., Zhou, Y., Sheng, Z., Leung, V.C.M.: Demographic information prediction: a portrait of smartphone application users. IEEE Trans. Emerg. Top. Comput. **6**(3), 432–444 (2016)
Rivron, V., Khan, M.I., Charneau, S., Chrisment, I.: 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), pp. 266–273. IEEE (2016)
Seneviratne, S., Seneviratne, A., Mohapatra, P., Mahanti, A.: Your installed apps reveal your gender and more!. ACM SIGMOBILE Mob. Comput. Commun. Rev. **18**(3), 55–61 (2015)
Tu, Z., Fan, Y., Li, Y., Chen, X., Su, L., Jin, D.: From fingerprint to footprint: cold-start location recommendation by learning user interest from app data. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **3**(1), 26 (2019)
Wang, Y., Tang, Y., Ma, J., Qin, Z.: Gender prediction based on data streams of smartphone applications. In: Proceedings of International Conference on Big Data Computing and Communications, pp. 115–125. Springer (2015)
Wang, J., Wang, L., Wang, Y., Zhang, D., Kong, L.: Task allocation in mobile crowd sensing: state-of-the-art and future opportunities. IEEE Internet Things J **5**(5), 3747–3757 (2018)
Wang, J., Wang, Y., Lv, Q.: Crowd-assisted machine learning: current issues and future directions. Computer **52**(1), 46–53 (2019)
Wei, X., Huang, H., Nie, L., Zhang, H., Mao, X.L., Chua, T.S.: I know what you want to express: sentence element inference by incorporating external knowledge base. IEEE Trans. Knowl. Data Eng. **29**(2), 344–358 (2017)
Xu, Q., Erman, J., Gerber, A., Mao, Z., Pang, J., Venkataraman, S.: Identifying diverse usage behaviors of smartphone apps. In: IMC2011, pp. 329–344. ACM (2011)
Xu, R., Frey, R.M., Fleisch, E., Ilic, A.: Understanding the impact of personality traits on mobile app adoption-insights from a large-scale field study. Comput. Hum. Behav. **62**, 244–256 (2016a)
Xu, R., Frey, R.M., Ilic, A.: Individual differences and mobile service adoption: An empirical analysis. In: Proceedings of IEEE Second International Conference on Big Data Computing Service and Applications, pp. 234–243. IEEE (2016b)
Ying, J.J.C., Chang, Y.J., Huang, C.M., Tseng, V.S.: Demographic prediction based on users mobile behaviors. 2012 Nokia Mobile Data Challenge Workshop pp. 1–6 (2012)
Yu, D., Li, Y., Xu, F., Zhang, P., Kostakos, V.: Smartphone app usage prediction using points of interest. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(4), 174 (2018)
Yu, Z., Du, H., Yi, F., Wang, Z., Guo, B.: Ten scientific problems in human behavior understanding. CCF Trans. Pervasive Comput. Interact. **1**(1), 3–9 (2019)
Zhao, S., Ramos, J., Tao, J., Jiang, Z., Li, S., Wu, Z., Pan, G., Dey, A.K.: 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, pp. 498–509. ACM (2016)
Zhao, S., Pan, G., Zhao, Y., Tao, J., Chen, J., Li, S., Wu, Z.: Mining user attributes using large-scale app lists of smartphones. IEEE Syst. J. **11**(1), 315–323 (2017a)
Zhao, S., Ramos, J., Tao, J., Jiang, Z., Li, S., Wu, Z., Pan, G., Dey, A.K.: Who are the smartphone users? Identifying user groups with apps usage behaviors. GetMobile Mob. Comput. Commun. **21**(2), 31–34 (2017b)
Zhao, S., Zhao, Y., Zhao, Z., Luo, Z., Huang, R., Li, S., Pan, G.: Characterizing a user from large-scale smartphone-sensed data. In: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing (workshop), pp. 482–487. ACM (2017c)
Zhao, S., Xu, F., Luo, Z., Li, S., Pan, G.: Demographic attributes prediction through app usage behaviors on smartphones. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, pp. 870–877. ACM (2018)
Zhao, S., Jiang, Z., Ramos, J., Luo, Z., Li, S., Dey, K.A., Pan, G.: User profiling from their use of smartphone applications: a survey. Pervasive and Mobile Comput. https://doi.org/10.1016/j.pmcj.2019.101052 (2019a)

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

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

Zou, X., Zhang, W., Li, S., Pan, G.: Prophet: What app you wish to use next. In: Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication (poster), pp. 167–170. ACM (2013)

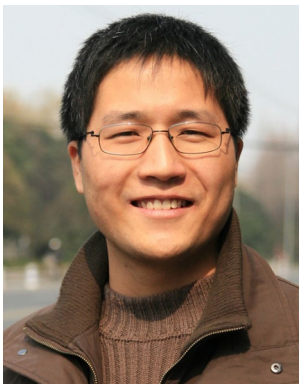
Sha Zhao is currently a Postdoctoral Research Fellow of the College of Computer Science and Technology, Zhejiang University. She received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2017. Zhao visited the Human-Computer Interaction Institute at Carnegie Mellon University as a visiting PhD student from September 2015 to September 2016. Her research interests include pervasive computing, mobile sensing, data mining, and machine learning. She won the Best Paper Award of ACM UbiComp'16.

Feng Xu is currently a Master candidate of the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. He received his B. Sc. degree in Computer Science and Technology from Zhejiang University, Hangzhou, China, in 2018. His research interests include machine learning and data mining.

Yizhi Xu is currently a Master candidate of the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. He received his B. Sc. degree in Electrical Engineering from Zhejiang University, Hangzhou, China, in 2017. His research interests include machine learning and data mining.

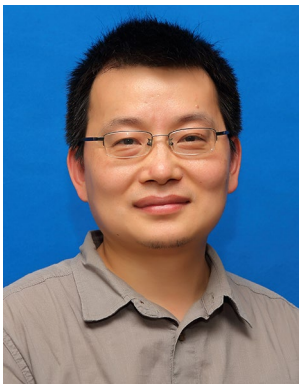
Xiaojuan Ma is an assistant professor of Human-Computer Interaction (HCI) at the Department of Computer Science and Engineering (CSE), Hong Kong University of Science and Technology (HKUST). She received the Ph.D. degree in Computer Science at Princeton University. She was a post-doctoral researcher at the Human-Computer Interaction Institute (HCII) of Carnegie Mellon University (CMU), and before that a research fellow in the National University of Singapore (NUS) in the Information Systems department. Before joining HKUST, she was a researcher of Human-Computer Interaction at Noah's Ark Lab, Huawei Tech. Investment Co., Ltd. in Hong Kong. Her background is in Human-Computer Interaction. She is particularly interested in data-driven human-engaged AI and Human-Robot Interaction in the domain of education, health, and design.

Zhiling Luo was an Assistant Professor in Computer Science at Zhejiang University, China. He received his B.S. and Ph.D. degree in Computer Science from Zhejiang University in 2012 and 2017, respectively. He was the visiting scholar of Georgia Institute of Technology, US, in 2016. His research interests include service computing, machine learning and data mining.



Shijian Li is currently with the College of Computer Science and Technology, Zhejiang University. He received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2006. In 2010, he was a Visiting Scholar with the Institute Telecom Sud-Paris, Evry, France. He was published over 40 papers. His research interests include sensor networks, ubiquitous computing, and social computing. Dr. Li serves as an Editor of the International Journal of Distributed Sensor Networks and as Reviewer or PC Member of more than ten conferences.

Anind Dey is a Professor and Dean of the Information School and Adjunct Professor in the Department of Human-Centered Design and Engineering. Anind is renowned for his early work in context-aware computing, an important theme in modern computing, where computational processes are aware of the context in which they operate and can adapt appropriately to that context. His research is at the intersection of human-computer interaction, machine learning, and ubiquitous computing. For the past few years, Anind has focused on passively collecting large amounts of data about how people interact with their phones and the objects around them, to use for producing detection and classification models for human behaviors of interest. He applies a human-centered and problem-based approach through a collaboration with an amazing collection of domain experts in areas of substance abuse (alcohol, marijuana, opioids), mental health, driving and transportation needs, smart spaces, sustainability, and education. Anind was inducted into the ACM SIGCHI Academy for his significant contributions to the field of human-computer interaction in 2015.



Gang Pan received the B.Eng. and Ph.D. degrees from Zhejiang University, China, in 1998 and 2004, respectively. He is currently a professor of the Department of Computer Science, and deputy director of State Key Lab of CAD&CG, Zhejiang University, China. From 2007 to 2008, he was a visiting scholar at the University of California, Los Angeles. His current interests include artificial intelligence, pervasive computing, brain-inspired computing, and brain-machine interfaces. He has authored over 100 refereed papers, and 35 patents granted. Dr. Pan received three best paper awards (e.g. ACM UbiComp'16) and three nominations from premier international conferences. He is the recipient of IEEE TCSC Award for Excellence (Middle Career Researcher), CCF-IEEE CS Young Computer Scientist Award, and the State Scientific and Technological Progress Award. He serves as an Associate Editor of IEEE Trans. Neural Networks and Learning Systems, IEEE Systems Journal, Pervasive and Mobile Computing, and ACM Proceedings of Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT).