Mining User Attributes Using Large-Scale APP Lists of Smartphones

Sha Zhao, Gang Pan, Yifan Zhao, Jianrong Tao, Jinlai Chen, Shijian Li, and Zhaohui Wu

***Abstract*—Prevalence of smartphones is changing people's lifestyle. Mobile applications (abbr. APPs) on a smartphone serve as entries for users to access a wide range of services. What APPs installed on one's smartphone, i.e., APP list, convey lots of information regarding his/her personal attributes, such as gender, occupation, income, and preferences. This paper addresses the discovery of user attributes from an APP list. We develop an attribute-specific representation to describe user characteristics and then model the relationship between an attribute and an APP list. A large-scale real-world data set with APP lists of more than 100 000 smartphones is used for evaluation. Our approach achieves the average equal error rate of 16.4% for 12 predefined user attributes. To our best knowledge, this is the first work to explore mining of user attributes from installed APP lists.**

***Index Terms*—APP lists, mobile sensing, smartphones, user attributes, user mining.**

I. Introduction

With the rapid growth of smartphones, more than 1.5 billion people worldwide have been covered. Frequent use of smartphones generates massive personal historical information. Since a smartphone is usually tightly associated with a same person, it reveals rich clues regarding his/her behaviors, lifestyle, preferences, and so on. Typical personal information includes: 1) location signal, derived from GPS and cell towers; 2) individual activity, derived from an accelerometer, a camera, gyroscope, etc.; 3) social signal, derived from call detail record (CDR), GPS, WiFi/Bluetooth connection, and contacts; and 4) personalized information, such as interests and preferences. Smartphones are providing us an opportunity to understand users well.

The behavior information recorded by smartphones can help users understand themselves objectively and extensively. In the real world, users usually express subjective ideas about themselves, such as, "I often keep regular hours for work and rest," "I am good at communication." However, what people say is quite often an unreliable proxy for what they do [1]. Behavioral observations are surprisingly weakly related to cognitive reports [2]. Behaviors recorded by smartphones can help users discover objective and unobservable information about themselves. The detailed records in mobile phones can be considered as a partial life log. They are important for understanding users extensively.

Understanding users well can help us to improve devices, services, and applications. In particular, it is very important for personalization of applications, such as personalized Web search, personalized recommendation, targeted advertising, and smart environments. Services can be recommended to the users according to their requirements, interests, or habits. Advertisements can be actively pushed to targeted users. Devices in a smart environment can adjust adaptively according to users' interests, preferences, or requirements.

User attribute is a common and simple way to describe a user's characteristics. In other words, it expresses a user with a series of attributes. A user attribute could be, for example, gender, occupation, income, interests, and preferences. Recently, there have been a couple of works about mining user attributes based on mobile phone data, for example, using one's location information from GPS and CDR to find users' daily mobility patterns [3], mining social events of user interests from CDR and GPS [4], using call log, media, calendar, application usage in phones to infer users' demographic information [5].

Although some of user attributes can be mined from personal information of GPS, CDR, WiFi connections, and contacts, there are still some limitations in mining user attributes. The mined attributes are limited in location, individual activity, and social signals. There are still lots of user attributes to be mined, which reveal personalized information. For example, is he/she a student? Does he/she have a baby? Woman or man? Does he/she like watching movies?

This paper addresses mining user attributes from one's APP list. APP list, i.e., what APPs installed on one's smartphone, conveys lots of information regarding his/her personal attributes. Compared with other personal data, APP lists have several advantages for mining user attributes. First, APP lists on smartphones can intuitively reflect users' basic attributes, interests, preferences, and living habits in a wide range. APPs on smartphones can be considered as one's entry to access services. Almost all the smartphone functions are implemented by APPs. What APPs installed reveals ones' life requirements. For example, women may like to install APPs about dressing, yoga, beauty, and shopping, whereas men may prefer to install APPs on sports, cars, and news. Second, APP list is more easily to

Manuscript received July 31, 2014; revised November 12, 2014; accepted November 30, 2014. This work was supported in part by the Program for New Century Excellent Talents in University under Grant NCET-13-0521, by the National Key Basic Research Program of China under Grant 2013CB329504, by the National Key Technology R&D Program under Grant 2012BAH94F03. *(Corresponding author: Gang Pan.)*

S. Zhao, G. Pan, Y. Zhao, J. Tao, S. Li and Z. Wu are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (e-mail: szhao@zju.edu.cn; gpan@zju.edu.cn; yifanzhao@zju.edu.cn; taojianrong0@zju.edu.cn; shijianli@zju.edu.cn; wzh@zju.edu.cn).

J. Chen is with Zhejiang Merit Internet Technology Company, Hangzhou 310013, China (e-mail: chenjl@getui.com).

Digital Object Identifier 10.1109/JSYST.2015.2431323

be accessed under the condition of privacy protection. It does not need historical information. We may quickly figure out the users' attributes at the moment when accessing the users' current APP lists.

In this paper, we propose a framework to mine user attributes from APP lists. We develop an attribute-specific representation to describe user characteristics, and then model the relationship between an attribute and an APP list. Thus, the attribute mining problem is transformed into a classification problem. The approach is evaluated by a large-scale data set of APP lists of more than 100 000 smartphones.

The contributions of our paper are fourfold:

1) We tackle user attribute mining with APP lists installed on smartphones. APP lists intuitively reflect users' interests, preferences, and living styles. To our best knowledge, this is the first work to explore mining of user attributes from ones' installed APP lists.
2) We propose an information gain (IG)-based approach to measure the relationship between a user attribute and an APP. The purpose is to evaluate which APPs are more important to distinguish a given user attribute, so that we can select relevant APPs to build an effective APP-based user representation.
3) We transform the user attribute mining problem to a classification problem by building an APP-based user representation for each user attribute. A large-scale real-world data set of 100 281 smartphones is used for evaluation. Our approach achieves the average equal error rate (EER) of 16.4% for 12 predefined user attributes.
4) We find that the frequency of APPs follows Zipf's law. It indicates that only a few APPs have very large installation and numerous APPs are with a small installation.

The remainder of this paper is organized as follows. In the next section, the related work is reviewed. The data set of APP lists we used is described in Section III. The proposed approach is introduced in Section IV. In Section V, the experiment results are demonstrated. Finally, conclusion and discussion are given.

II. Related Work

Mining individual traits and attributes using various cues, such as website browsing logs, activities in online social networks, and mobile phone data, have been actively investigated in the past decade.

First, there have been some works mining users' attributes from website browsing logs. Users' attributes such as age, gender, occupation, and education level may be discovered by analyzing their online behavior from websites [6]–[8]. For instance, Weber *et al.* [8] analyzed a large query log of 2.3 million anonymous registered users from a Web-scale U.S. search engine, in order to jointly analyze their online behavior in terms of who they might be (demographics), what they search for (query topics), and how they search (session analysis). The results showed that there were important differences in search behavior among different demographics in terms of the topics they search for, and how they search. Wang *et al.* [7] analyzed the relationship between users' clicking behaviors and the category of the news story to model users' interests by mining Web log data of an adaptive news system.

Similarly, some works have been done to mine user attributes on the basis of the properties of social network sites, such as Twitter, Sina Weibo, and Facebook [9]–[12]. The key content from social network sites is analyzed to discover users' attributes. For instance, Rao *et al.* [12] classified latent user attributes, including gender, age, regional origin, and political orientation solely from Twitter user language or similar highly informal content. Liu *et al.* [11] identified users' interests by extracting keywords from the largest microblogging website in China, Sina Weibo. Kosinski *et al.* [10] used the data set of 58 000 volunteers who provided their Facebook Likes to discover a range of highly sensitive personal attributes, including sexual orientation, ethnicity, personality traits, age, gender, and other personal information.

In addition, many previous works have been done on mining user attributes from mobile phone data sets, such as GPS, CDR, sensor data, and phone usage information. Personal information may be discovered by analyzing the mobile phone data. For example, people's daily mobility may be mined to some extent [13] and important places in daily life such as work office and home may be identified by analyzing the location information derived from GPS call logs and cell tower [14], [15]. Users' activities or indentities can be recognized using the sensor data, such as accelerometer, gyroscope [16], [17]. Reddy *et al.* [18] combined accelerometer and GPS to recognize five transportation modes, including still, walk, run, bike, and motor. Users' social networks can also be revealed to some extent according to communication events, logs of discovered Bluetooth devices, WLAN MAC address, and GPS [19], [20]. Based on real logs of mobile users, including usage logs of applications, GPS data, system information, global system for mobile communications data, call logs, and sensor data, users' usage patterns of mobile application in real-time and mobile environment may be discovered [21], [22]. Shin and Dey [23] proposed a model to automatically detect users' problematic usage smartphones. Mo *et al.* [24] used GPS, call log, media, Bluetooth, calendar, acceleration, and application use frequency to predict phone users' demographic information, such as gender, job type, marital status, and age. Chittaranjan *et al.* [25] mined users' personality traits (Extraversion, Agreeableness, Conscientiousness, Emotional Stability and Openness to Experience) from smartphone usage data. In literature, the information of APP list, however, has not been employed to mine users' attributes.

III. Data Set Description and Demography

A. Data Set Overview

The data set used in this paper is about APP installation logs of smartphones, provided by a mobile Internet company in China. It contains 100 281 smartphones and 80 896 Android applications. The installation log consists of 2 811 910 records. Each installation record consists of three fields, i.e., user ID,

User ID	Installation package	APP name
004ac0891bb123d89e80ff182699f2d	com.tencent.minihd.qq	QQ
004ac0891bb123d89e80ff182699f2d	com.tencent.mm	微信
004ac0891bb123d89e80ff182699f2d	com.UCMobile	UC浏览器
004ac0891bb123d89e80ff182699f2d	com.besttone.FortuneStreet.plugin	股票财经
004ac0891bb123d89e80ff182699f2d	buke.besttone.caipiao.plugin	号百彩票
008f23e4ec85676e5d823abebb2043f1	com.soufun.app	搜房
008f23e4ec85676e5d823abebb2043f1	com.autonavi.xmgd.navigator	高德导航
008f23e4ec85676e5d823abebb2043f1	com.cubic.autohome	汽车之家
008f23e4ec85676e5d823abebb2043f1	com.tencent.mobileqq	QQ
008f23e4ec85676e5d823abebb2043f1	com.sina.weibo	微博
064d732ebc796fc64403a9ccf52a11ac	com.geili.gou	美丽购
064d732ebc796fc64403a9ccf52a11ac	com.sankuai.meituan	美团
064d732ebc796fc64403a9ccf52a11ac	com.sina.weibo	微博
064d732ebc796fc64403a9ccf52a11ac	com.eg.android.AlipayGphone	支付宝钱包
064d732ebc796fc64403a9ccf52a11ac	com.taobao.taobao	淘宝

Fig. 1. Sample of APP installation logs in the data set.

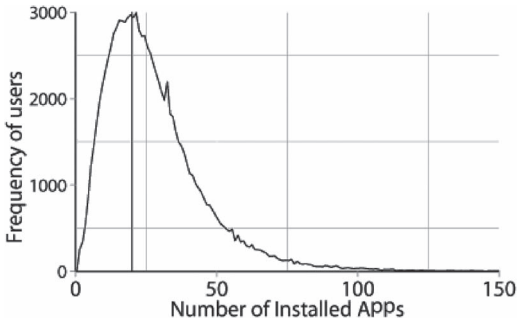

Fig. 2. Frequency of users in term of installed APPs.

installation package, and APP name. A sample of installation records is shown in Fig. 1.

1) User ID: the unique identity of the sampled smartphone. Each ID is anonymized for the privacy purpose.
2) Installation package: it can be used to identify an APP.
3) APP name: most of the APP names are in Chinese and English, and a few in Korean, Japanese, and other languages.

Fig. 2 shows the frequency of users in terms of the number of installed APPs. The horizontal axis is the number of APPs in an APP list, and the vertical axis is the frequency of users, that is, how many users have installed the exact number of APPs. It can be seen that many users have nearly 20 APPs, and just a few users install more than 50 APPs.

B. Zipf's Law of APPs

We ranked the APPs according as how many people installed them (frequency of APPs). We found that frequency of APPs follows Zipf's law, as shown in Fig. 3, where the x-axis is the logarithm of APP rank and the y-axis is the logarithm of the frequency of APPs. A straight line is fitted, as shown in red in Fig. 3, expressed in (1). It means there are a few APPs with very large installation and very numerous APPs with a small installation

$$\log(\text{Frequency of APPs}) = 6.29 - 0.76\log(\text{Rank of APPs}). \quad (1)$$

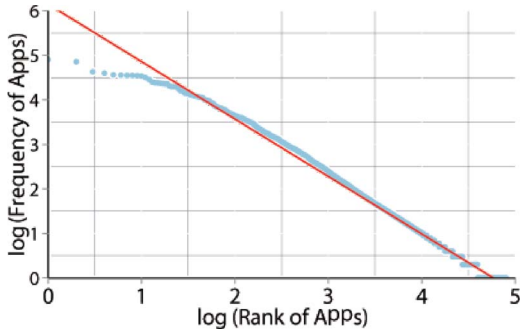

Fig. 3. Frequency of APPs follows Zipf's law.

IV. Proposed Framework

In order to mine user attributes from APP lists, it is necessary to first represent a user. Here, we use an APP-based vector to represent the user. For a given attribute, only some of the APPs are useful for identifying the attribute. We develop an approach to measure how important an APP is for the given attribute. This will lead to an attribute-specific user representation. After each user attribute is simply defined as a two-value label, the mining problem becomes the two-class problem.

A. APP List-Based User Representation

Intuitively, we can exploit one's APP list to represent a user. In detail, we take each APP as a dimension and represent each user as an APP-based vector. If an APP is installed, the corresponding value of its dimension is set to 1, and on the contrary, the value is 0. Formally, user u is represented by

$$u = (a_1, a_2, a_3, \ldots, a_k, \ldots, a_m) \quad (2)$$

where a_k is for the kth APP, and it has two values, i.e., 1 and 0, for indicating whether the APP is installed. In this case, u will be very sparse, since most users only have very few APPs out of all the APPs.

B. Attribute-APP Relationship Measure

For a given attribute, if all the APPs are used to build the user representation, the user vector will be dramatically long. Not all the APPs are useful for describing a user for a given attribute. If an irrelevant or redundant APP is removed, it will not affect the attribute mining. In order to increase the computational efficiency, we have to know which APPs are important to build the user representation for the specific attribute.

The APP's relevance to an attribute can be measured by the amount of information it brings about for the attribute. In information theory, entropy is the amount of information contained in a piece of data. For example, information about a user attribute has its own entropy. IG measures the expected reduction in entropy by learning the state of a random variable [26]. Here, we employ IG of an APP to measure the expected reduction in entropy of a user attribute. More information gain, more relevance of the APP to the user attribute.

IG is a popular approach employed as a term importance criterion [27], particularly in decision tree. It also works well with text categorization and has often been used particularly for

the high-dimensionality of the feature space [28], [29]. We use the IG to compute the relevance of an APP to a user attribute. Given a user attribute, we rank the APPs according to their information gain. The first APP is considered to be the most relevant to the attribute. According to the ranking result, we can choose those most relevant APPs to construct a compact user vector space.

Formally, the IG of an APP A to a specific attribute Ψ is defined as

$$\mathrm{IG}(\Psi, A) = H(\Psi) - H(\Psi|A) \tag{3}$$

where $H(\Psi)$ represents the entropy of the user attribute, and $H(\Psi|A)$ represents the entropy of Ψ conditioned on APP A. They are defined by (4) and (5), respectively

$$H(\Psi) = -\sum_{\psi \in \Psi} P(C_\psi) \log P(C_\psi) \tag{4}$$

where C_ψ represents the group of users whose attribute value of Ψ is ψ

$$\begin{aligned} H(\Psi|A) = &-P(A=1) \sum_{\psi \in \Psi} P(C_\psi|A=1) \log(C_\psi|A=1) \\ &- P(A=0) \sum_{\psi \in \Psi} P(C_\psi|A=0) \log(C_\psi|A=0) \end{aligned} \tag{5}$$

where $A = 1$ means the APP A is installed for a user, and $A = 0$ means it is not installed.

When given a set of samples of APP lists and the corresponding ground truth of the user attribute Ψ, all the probability in the right hand of (4) and (5) can be easily estimated by statistics. Thus, for each APP, its IG to the attribute can be computed by (3).

C. Mining User Attributes With Classification

With the IG-based measure, we can evaluate how important an APP is for the attribute. This will lead to an attribute-specific user representation, that is, each user attribute has its own optimal APP-based user representation.

In this paper, we regard an attribute as a label. Thus, the problem actually is to know whether a user has the label or not. In other words, there are only two categories for a user attribute. From the viewpoint of classification, it is a two-class problem, where one class consists of the users with the label, whereas the other class consists of those without the label.

Therefore, mining of user attributes could be solved by classification. For each user attribute, its APP-based user representation could be constructed via the IG-based measure between each APP and the user attribute. Then, a binary classifier will take the attribute-specific user representation as the input to determine whether a user has the label or not.

There are many binary classifiers proposed in the literature. Support vector machine (SVM) has shown good generalization performance for solving classification problems [30]. It maps the input points into a high-dimensional feature space and finds a separating hyperplane that maximizes the margin between two classes in this space. Equation (6) shows the objective function to find the optimal hyperplane. To find the optimal hyperplane, SVM uses the dot product functions in feature space that are called kernels [31]

$$\min \frac{1}{2}\|w\|^2 \qquad \text{s.t.}, y_i = \left(w^T x_i + b\right) \geq 1, i = 1, \ldots, n \tag{6}$$

where w is the normal vector to the hyperplane, the parameter $b/\|w\|$ determines the offset of the hyperplane from the origin along the normal vector w, x_i represents the ith point, y_i is either 1 or -1.

In this paper, we employ SVM to solve the two-class problem. For each user attribute, we will train an SVM classifier with the attribute-specific user vector as its input.

V. Experiments

Here, we will evaluate performance of our approach using the large-scale data set of APP lists introduced in Section III. We simply regard each smartphone in the data set as a user. In order to verify the proposed approach, 12 user attributes are predefined.

A. Filtering of Preinstalled APPs

For some smartphones, there are a few APPs preinstalled, which cannot reflect the user's interests and preferences. We manually filtered the preinstalled APPs from the original data set. We observe that many preinstalled APPs are bound to smartphone brands, mobile operators, and smartphone operating systems, such as Samsung In-App Purchase, China Unicom, and Android System Services. Finally, 1876 preinstalled APPs were filtered, and there were 79 020 APPs left in total for the follow-up evaluation.

B. Twelve Predefined User Attributes

In order to evaluate our approach, we should have ground truth of user attributes. A large-scale ground truth, however, is difficult to get. To cope with this problem, this paper employs two strategies to get the ground truth for the experiments.

1) Predefinition of user attributes via smartphone models

Each smartphone in our real-world data set is accompanied with its phone model. Then, for each smartphone, we crawled its model-related data from websites. Finally, we design two user attributes: smartphone price and smartphone size. According to the 2013 annual report on Chinese smartphone market by iiMedia,[1] price is the primary factor to be considered in the purchase of smartphones, which accounts for 69.6% of all the factors. Smartphone price can reflect user income or consumption to a certain extent. In addition, smartphone size is another important factor, which accounts for 20.2%. Size may reveal user preference of phone usage to some extent.

2) Predefinition of user attributes via niche APPs

Some niche APPs usually are connected with concentrative target user groups. In a sense, a niche APP may mean an underlying user attribute. Based on this observation, we select ten

[1] http://www.iimedia.cn/36504.html

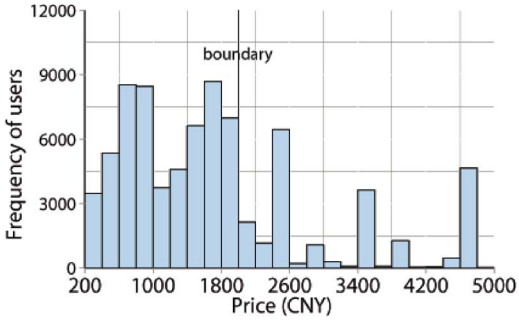

Fig. 4. Histogram of price.

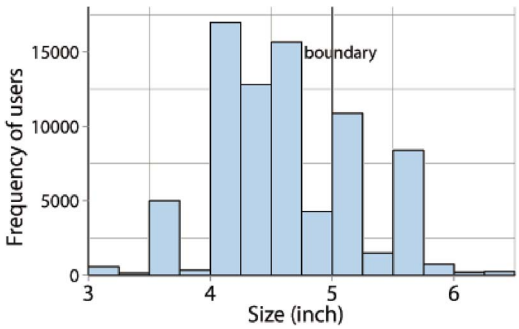

Fig. 5. Histogram of size.

niche APPs as ten underlying user attributes. We remove all the ten niche APPs from the data set for fair evaluation, that is, for both training data and test data, the ten niche APPs have never been used for user representation.

We will describe how the 12 predefined user attributes are extracted in the following.

1) Price and Size of Smartphones: Our data set has the information on model of each smartphone. We ranked the phone models in a descending order, and selected the top 700 phone models. The 700 models cover 78 022 users. Then, we crawled the price and size from the two websites.[2] We calculated the frequency of users in term of price and size, as shown in Figs. 4 and 5, respectively.

From Fig. 4, the price of 2 000.00 (CNY) is an obvious dividing point. Many of users use mobile phones with a price lower than 2 000.00, and only about 25% of users prefer the price higher than 2 000.00. We chose the value of 2 000.00 as a dividing point and divided users into two groups: low-price group (negative samples) and high-price group (positive samples).

As shown in Fig. 5, many users use mobile phones with a size smaller than 5.0 in, and only about 20% of users use mobile phones bigger than 5.0 in. We chose 5.0 as a dividing point and divide users into two groups, i.e., small-size group (negative samples) and large-size group (positive samples).

2) Ten User Attributes by Niche APPs: We chose ten niche APPs and took each APP as an underlying user attribute. The ten predefined user attributes are shown in Fig. 6. When we used APP lists to represent users, the ten APPs were removed. Each of the ten niche APPs can categorize users into two groups: positive group, in which users install the niche APP,

[2]http://mobile.zol.com.cn/ and http://www.cnmo.com/

APPs	Attributes	Installation package
QQ电影票	Movie_fan	com.tencent.movieticket
号百彩票	Lottery	buke.besttone.caipiao.plugin
股票财经	Stocks	com.besttone.FortuneStreet.plugin
艺龙旅行	Travel	com.dp.android.elong
搜房网	Housing	com.soufun.app
高德导航	Driving	com.autonavi.xmgd.navigator
超级课程表	Student syllabus	com.xtuone.android.syllabus
美团	Group_buying	com.sankuai.meituan
美丽购	Beauty shopping	com.geili.gou
粉粉日记	Pinknote	pinkdiary.xiaoxiaotu.com

Fig. 6. Ten niche APPs for user attributes.

and negative group consisting of users without the APP. The ten attributes are detailed in the following.

i. **Movie_fan**. It provides film services, such as online booking, real-time query, movie review, and movie introduction. This APP reveals the user's interest in movies.
ii. **Lottery**. It provides lottery services such as betting record query and share, online social interaction with other users. It reveals the user's interest in lottery.
iii. **Stocks**. It provides stock services and reveals the user's interest in stock investment.
iv. **Travel**. This APP provides travel services, such as hotel reservation, flight ticket booking, and location-based query.
v. **Housing**. It is an APP about domestic real estate and house rental, such as buying, selling or renting houses. The users who plan to buy, sell, or rent houses may install this APP.
vi. **Driving**. It is a driving navigation APP. Its user most likely owns a car.
vii. **Student syllabus**. It provides online class scheduling and recommendation for students. This APP indicates the user's occupation may be a student.
viii. **Group_buying**. It provides groupon and discount in many aspects, such as restaurants, shopping, and hotels.
ix. **Beauty shopping**. This APP is an online shopping platform particularly designed for females. It provides latest news about dresses, shoes, handbags, etc. It indicates its users most likely are women and prefer shopping online for fashion.
x. **Pinknote**. It is an assistant APP for women's life. Users can keep diaries, make schedules, keep accounts, etc. It reveals information on user gender.

C. Implementation and Performance Measures

We used the LibSVM [32] for SVM implementation. We tested several kernel functions and found the radial basis function performs the best. The positive samples are much less than negative samples, particularly for the ten user attributes by niche APPs. For the training sets with unbalanced class sizes, we adjusted the weights of samples by increasing the weight of positive samples in the cost function.

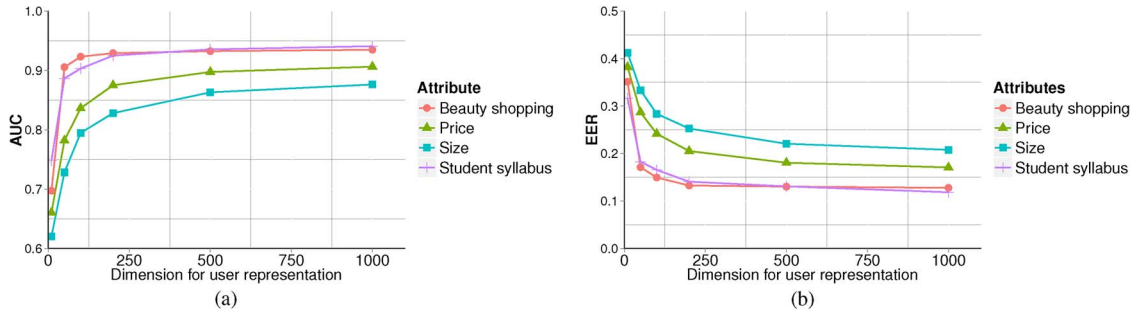

Fig. 7. Performance with varying user representation dimension for four attributes of price, size, beauty shopping, and student syllabus. (a) AUC. (b) EER.

To train the classifier for each predefined attribute, we employed the fourfold cross-validation policy. The sampled data set was randomly divided into fourfolds as evenly as possible. In each round, threefolds were 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 this procedure four times.

We used two criterions, i.e., EER and area under curve (AUC), to measure performance of our approach. The EER means where the false positive rate (FPR) and false negative rate (FNR) are equal on the receiver operating characteristics (ROC) curve. Thus, the EER is equal to FPR or FNR. Smaller EER means better performance, i.e., a lower error rate of the classifier. AUC is the area under the ROC curve. Bigger AUC means better performance, a higher probability that a randomly chosen negative example will have a smaller estimated probability of belonging to the positive class than a randomly chosen positive example. It is computed by [33]

$$A = \frac{S_p - N_p(N_p+1)/2}{N_n N_p} \tag{7}$$

where N_p is the number of the positive samples in the test set and N_n is the number of the negative samples, S_p is the sum of the ranks of the positive samples.

D. Experimental Results

We conducted experiments for the 12 predefined user attributes to evaluate our approach. After filtering the preinstalled APPs and filtering phone models for two user attributes (price and size), as aforementioned, finally, we got a data set with 78 022 users for the following experiments.

1) Optimization of User Representation Dimension: In the attribute-specific user representation, each APP can be taken as a dimension. In order to improve the computational efficiency, we have to optimize the dimension of user representation. We experimented with 10, 50, 100, 200, 500, and 1 000 dimensions to represent users. In our approach, the IG-based measure was applied to rank APPs and select the top APPs for a given user attribute.

We took four predefined attributes for the experiment, price, size, beauty shopping, and student syllabus. The performance of AUC and EER with different dimension is shown in Fig. 7. We can see that, when the dimension increases, the performance eventually becomes steady from the dimension of nearly 500 for all the four user attributes. Considering efficiency of computation, the dimension of user representation is set to 500 in the following experiments.

2) Attribute-APP Relationship: In this experiment, we demonstrate relationship between attributes and APPs. Fig. 8 shows four user attributes (i.e., price, size, beauty shopping, and student syllabus) and their top ten relevant APPs, with their IG as well. It can be found that, for both price and size, the IG is very close among the top ten APPs, whereas for beauty shopping and student syllabus, the IG varies a lot (the IG of the most relevant APP is much more than that of others).

According to Fig. 8(a), users in the high-price group have a preference to the APPs of Qunar travel, Taobao (an APP for online shopping), a few casual games (Elimination game, Rhythm master, Pencil pilot, and Link link). The most important APP for the price attribute is Qunar travel, which is an application developed for travelling. It indicates that the group preferring high phone price may travel more frequently than the group of low phone price. The casual games are also very important, which indicates that the users in high-price group may often play these games for entertainment. As shown from Fig. 8(b), the users in the large-size group have a preference to APPs of animation, video, some games (Pencil pilot, Car racing), etc. They may need a screen with larger size to enjoy animation, videos or playing games.

According to Fig. 8(c), female users who have the attribute of beauty shopping, have a preference to beauty camera, Meitu photoshop, photo booth, beauty shopping, Jumei cosmetics, and mushroom street. These APPs have very significant features of women. As shown from Fig. 8(d), the users who have the attribute of student syllabus, have a preference to Youdao dictionary (an English learning dictionary), learning assistant, RenRen (Chinese online social network similar to Facebook). It fits the occupation of students very well.

3) Performance of Mining User Attributes: To compare with other classifiers, we tested two other algorithms, Adaboost and multi nominal logistic regression. The comparison performance of AUC and EER is shown in Fig. 9. We can see that SVM performs best for most user attributes. The AUC of most attributes is above 0.9. The EER of most attributes is smaller than 0.2. This experimental result shows the advantages of SVM. The best EER (2.8%) is of the APP of Stocks, and the worst EER (25.5%) is of phone size. The overall EER achieves 16.4%.

4) Performance With Different Sample Size: In order to investigate how the sample size affects the performance, we experimented with 2 000, 5 000, 10 000, 20 000, 30 000, 40 000,

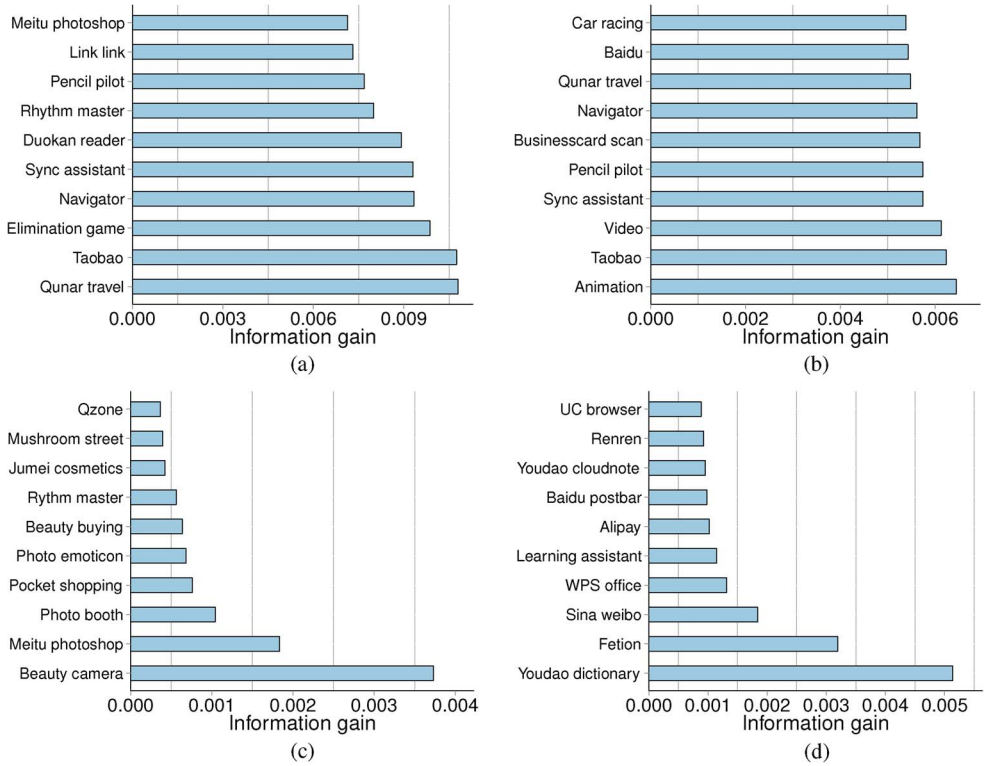

Fig. 8. IG for four attributes of price, size, beauty shopping, and student syllabus. (a) Price. (b) Size. (c) Beauty shopping. (d) Student syllabus.

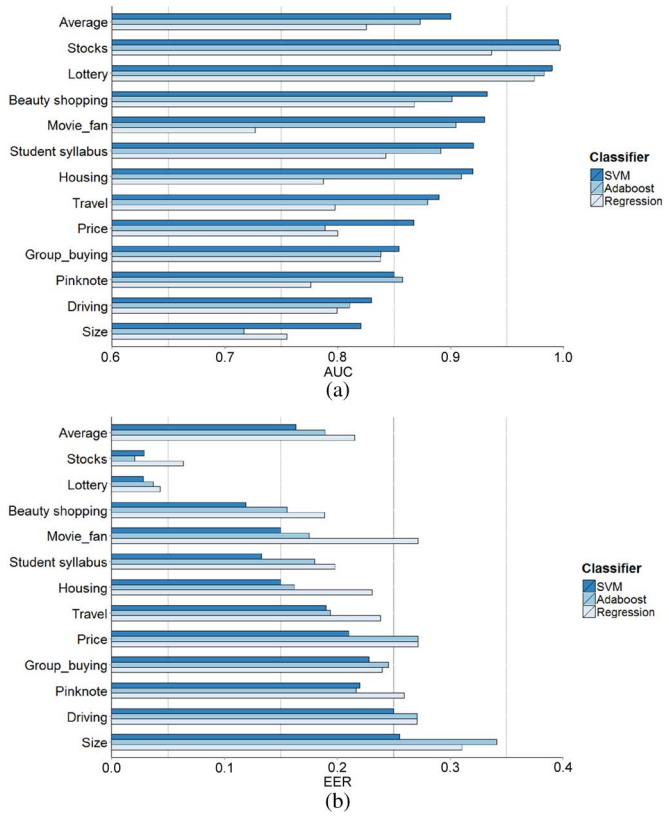

Fig. 9. Comparison performance with Adaboost and logistic regression for the 12 attributes. (a) AUC. (b) EER.

50 000, 60 000, 70 000, and all the samples to train the classifier. Then, we compared their performance of AUC and EER, as shown in Fig. 10.

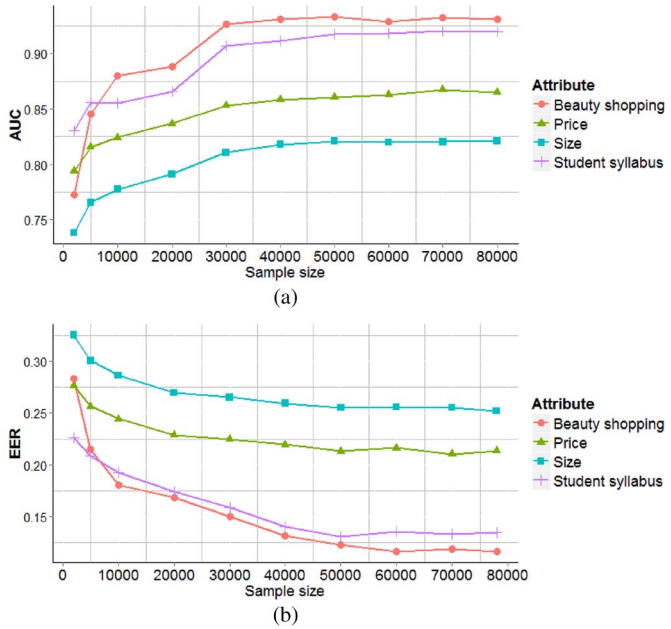

Fig. 10. Performance with varying sample size for four attributes of price, size, beauty shopping, and student syllabus. (a) AUC. (b) EER.

In this experiment, we took four predefined attributes, price, size, beauty shopping, and student syllabus. For each predefined user attribute, the proportion of positive and negative samples is kept roughly the same in each sampling. The SVM classifier was trained by fourfold cross-validation policy. To test the trained classifier, we used the left samples in the whole data set. We repeated sampling seven times for each sample size, and then averaged the performance.

Fig. 10 shows the performance with varying sample size. From Fig. 10, we can see that, with the sample size increasing, performance eventually increases and becomes steady at the size of 60 000 for all the four user attributes.

VI. Conclusion and Discussions

One's APP list of smartphones reveals a lot of underlying user properties. In this paper, we develop an effective framework to mine user attributes from APP lists. In the framework, an APP-based user representation is established to describe user characteristics. In order to evaluate which APPs are more important to distinguish a given user attribute, we define an IG-based measure to select relevant APPs for building an effective APP-based user representation. This paper only considers those user attributes, which could be discretized as two values. Experiments were conducted for 12 predefined user attributes and the overall EER achieved about 16.4%. This work is a promising step toward mining user attributes from users' APP lists on smartphones.

Although APP lists have lots of information about user attributes, they still have some defects. First, APP lists do not contain information on how frequently an APP is used. Some APPs installed may rarely be used, whereas others may be frequently used. It would be much more information if we know when and how much time an APP is used. We will try to cope with this issue in the future work.

For limitation in the data with ground truth of real-world user attributes, in this paper, we use niche APPs to design user attributes for verification. Niche APPs, however, cannot straightly indicate high-level user attributes. Our future work will try to develop a data set with ground truth to explore more user attributes.

References

[1] I. Deutscher, *What We Say/What We Do: Sentiments & Acts*. Glenview, IL, USA: Scott, Foresman, 1973.

[2] H. R. Bernard and P. D. Killworth, "Informant accuracy in social network data II," *Human Commun. Res.*, vol. 4, no. 1, pp. 3–18, Mar. 1977.

[3] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using mobile phone location data," *IEEE Pervasive Comput.*, vol. 10, no. 4, pp. 36–44, Jun. 2011.

[4] V. A. Traag, A. Browet, F. Calabrese, and F. Morlot, "Social event detection in massive mobile phone data using probabilistic location inference," in *Proc. PASSAT-SOCIALCOM*, 2011, pp. 625–628.

[5] S. Brdar, D. Culibrk, and V. Crnojevic, "Demographic attributes prediction on the real-world mobile data," in *Proc. Mobile Data Challenge Nokia Workshop, Conjunction Int. Conf. Pervasive Comput.*, 2012, pp. 1–5.

[6] K. De Bock and D. Van den Poel, "Predicting website audience demographics forweb advertising targeting using multi-website clickstream data," *Fundamenta Informaticae*, vol. 98, no. 1, pp. 49–70, Mar. 2010.

[7] W. Wang, D. Zhao, H. Luo, and X. Wang, "Mining user interests in web logs of an online news service based on memory model," in *Proc. NAS*, 2013, pp. 151–155.

[8] I. Weber and A. Jaimes, "Who uses web search for what and how," in *Proc. WSDM*, 2011, pp. 15–24.

[9] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in *Proc. ICWSM*, 2011, pp. 450–453.

[10] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Nat. Academy Sci.*, vol. 110, no. 15, pp. 5802–5805, Feb. 2013.

[11] Z. Liu, X. Chen, and M. Sun, "Mining the interests of Chinese microbloggers via keyword extraction," *Frontiers Comput. Sci.*, vol. 6, no. 1, pp. 76–87, Jan. 2012.

[12] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *Proc. 2nd Int. Workshop Search Mining User-Generated Contents*, 2010, pp. 37–44.

[13] S. Phithakkitnukoon, Z. Smoreda, and P. Olivier, "Socio-geography of human mobility: A study using longitudinal mobile phone data," *PloS ONE*, vol. 7, no. 6, Jun. 2012, Art. ID 0039253.

[14] S. Isaacman *et al.*, "Identifying important places in people's lives from cellular network data," in *Proc. PerCom*, 2011, pp. 133–151.

[15] G. Pan *et al.*, "Trace analysis and mining for smart cities: Issues, methods, and applications," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 120–126, Jun. 2013.

[16] J. Wu, G. Pan, D. Zhang, G. Qi, and S. Li, "Gesture recognition with a 3-D accelerometer," in *Proc. UIC*, 2009, pp. 25–38.

[17] Y. Zhang *et al.*, "Accelerometer-based gait recognition by sparse representation of signature points with clusters," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1864–1875, Sept. 2015.

[18] S. Reddy *et al.*, "Using mobile phones to determine transportation modes," *ACM Trans. Sens. Netw.*, vol. 6, no. 2, p. 13, Feb. 2010.

[19] S. A. Muhammad and K. Laerhoven, "Discovery of user groups within mobile data," in *Proc. Mobile Data Challenge Nokia Workshop, Conjunction Int. Conf. Pervasive Comput.*, 2012, pp. 1–6.

[20] J. Zheng and L. M. Ni, "An unsupervised learning approach to social circles detection in ego bluetooth proximity network," in *Proc. UbiComp*, 2013, pp. 721–724.

[21] C. Tan, Q. Liu, E. Chen, and H. Xiong, "Prediction for mobile application usage patterns," in *Proc. Mobile Data Challenge Nokia Workshop, Conjunction Int. Conf. Pervasive Comput.*, 2012, pp. 1–4.

[22] Y. Xu *et al.*, "Preference, context and communities: A multi-faceted approach to predicting smartphone app usage patterns," in *Proc. ISWC*, 2013, pp. 69–76.

[23] C. Shin and A. K. Dey, "Automatically detecting problematic use of smartphones," in *Proc. UbiComp*, 2013, pp. 335–344.

[24] K. Mo, B. Tan, E. Zhong, and Q. Yang, "Report of task 3: Your phone understands you," in *Proc. Mobile Data Challenge Nokia Workshop, Conjunction Int. Conf. Pervasive Comput.*, 2012, pp. 1–6.

[25] G. Chittaranjan, J. Blom, and D. Gatica-Perez, "Mining large-scale smartphone data for personality studies," *Pers. Ubiquitous Comput.*, vol. 17, no. 3, pp. 433–450, Mar. 2013.

[26] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[27] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. ICML*, 1997, pp. 412–420.

[28] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Inf. Process. Manage.*, vol. 42, no. 1, pp. 155–165, Jan. 2006.

[29] H. Uuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Syst.*, vol. 24, no. 7, pp. 1024–1032, Oct. 2011.

[30] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, Jun. 1998.

[31] C. F. Lin and S. D. Wang, "Fuzzy support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 464–471, Mar. 2002.

[32] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–39, Apr. 2011.

[33] D. J. Hand and R. J. Till "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Mach. Learn.*, vol. 45, no. 2, pp. 171–186, Nov. 2001.

Sha Zhao received her B. Sc. degree in computer science from Jinan University, Zhuhai, Guangdong, China, in 2011. She is currently a Ph.D. candidate of the College of Computer Science and Technology, Zhejiang University, Hangzhou, China.

Her research interests include pervasive computing, mobile sensing, data mining, and machine learning.

Gang Pan received B.Sc. and Ph.D. degrees in computer science from Zhejiang University, Hangzhou, China, in 1998 and 2004, respectively.

He is currently a Professor with the College of Computer Science and Technology, Zhejiang University. He was with the University of California, Los Angeles, Los Angeles, CA, USA, as a Visiting Scholar, from 2007 to 2008. He has published more than 100 refereed papers. His research interests include pervasive computing, computer vision, and pattern recognition.

Yifan Zhao received the B.Sc. degree in computer science from Zhejiang University, Hanzhou, China, in 2012. He is currently working toward the M.S. degree in the College of Computer Science and Technology, Zhejiang University.

His research interests include machine learning and data mining.

Jianrong Tao received the B.Sc. degree in computer science from Huazhong University of Science and Technology, Wuhan, China, in 2014. He is currently working toward the M.S. degree in the College of Computer Science and Technology, Zhejiang University, Hangzhou, China.

His research interests include machine learning and data mining.

Jinlai Chen received the B.Sc. degree from Zhejiang University, Hangzhou, China, in 2007.

He is currently with Zhejiang Merit Internet Technology Company, who takes charge of the cooperation with us.

His research interests include user characterizing, data mining, and recommendation system.

Shijian Li received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2006.

In 2010, he was a Visiting Scholar with the Institute Telecom SudParis, Evry, France. He is currently with the College of Computer Science and Technology, Zhejiang University. He 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 a Reviewer or PC Member of more than ten conferences.

Zhaohui Wu received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 1993. From 1991 to 1993, he was with the German Research Center for Artificial Intelligence as a joint Ph.D. student in the area of knowledge representation and expert system.

He is currently a Professor of computer science with Zhejiang University and the Director of the Institute of Computer System and Architecture, Zhejiang University. He has authored five books and over 200 refereed papers. His current research interests include intelligent systems, semantic grid, and ubiquitous embedded systems. He is with the editorial boards of several journals and has served as a Program Committee member for various international conferences.