Cohort Modeling Based App Category Usage Prediction

Yuan Tian
University of Nottingham
Nottingham, UK
yuan.tian@nottingham.ac.uk

Ke Zhou
University of Nottingham
Nottingham, UK
ke.zhou@nottingham.ac.uk

Mounia Lalmas
Spotify
London, UK
mounia@acm.org

Yiqun Liu
Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

Dan Pelleg
Yahoo Research
Haifa, Israel
pellegd@acm.org

ABSTRACT

Smartphones utilize context signals, such as time and location, to predict users' app usage tailored to individual users. To be effective, such personalization relies on access to sufficient information about each user's behavioral habits. For new users, the behavior information may be sparse or non-existent. To handle these cases, app category usage prediction approaches can employ signals from users who are similar along one or more dimensions, i.e., those in the same *cohort*. In this paper, we describe a characterization and evaluation of the use of such cohort modeling to enhance *app category usage prediction*. We experiment with pre-defined cohorts from three taxonomies - demographics, psychographics, and behavioral patterns - independently and in combination. We also evaluate various approaches to assign users into the corresponding cohorts. We show, through extensive experiments with large-scale mobile app usage logs from a mobile advertising company, that leveraging cohort behavior can yield significant prediction performance gains than when using the personalized signals at the individual prediction level. In addition, compared to the personalized model, the cohort-based approach can significantly alleviate the *cold-start* problem, achieving strong predictive performance even with limited amount of user interactions.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**; **Smartphones**.

KEYWORDS

Mobile app usage; Cold start; User cohort; App usage prediction; Mobile user characterization; Demographics

ACM Reference Format:
Yuan Tian, Ke Zhou, Mounia Lalmas, Yiqun Liu, and Dan Pelleg. 2020. Cohort Modeling Based App Category Usage Prediction. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20), July 14–17, 2020, Genoa, Italy.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3340631.3394849

1 INTRODUCTION

Smartphones are increasingly seen as large-scale, non-intrusive sensors of human activity, relating the physical and social space of people's lives, and how people interact with their devices. Smartphones have become increasingly important in our daily life; we use them through multiple apps to communicate with friends, check emails, take pictures, and play games, etc. Therefore, various stakeholders in the mobile industry [3, 24] are keen to understand how users engage with different apps, including phone operators, manufacturers, advertising companies, and service providers. It has been shown in the past that many contextual features, such as time, location, last used app and other device signals, can be used to predict app usage [2, 13, 20, 30, 32].

Personalization of app usage prediction has been investigated in many prior studies [2, 13, 20, 30, 32]. The ability to tailor prediction results to a particular individual enables a wealth of opportunity to better satisfy their particular needs. Personalized models are typically learned from observed usage behavior and context information (such as temporal/periodic pattern and sensor signals), which are either used directly [2, 30] or converted into a different representation (e.g., graph) to build more general models and improve personalization tasks [21, 42].

Despite the value of personalized models, one drawback is that they require lots of user historical interaction information to become effective. Every time a new user comes, a new prediction model needs to be trained for a period until it can predict users app usage correctly. The personalized prediction model can be very sensitive to the data available and might not perform well for new users. This is generally referred to as the *cold-start* problem. One way to alleviate this problem is by finding *cohorts* of users who share common attributes or experiences with the current user. Given a user, we can leverage the app usage behavior of other members of their cohort(s) to enhance prediction by providing signals if insufficient information is available for this user.

Modeling aggregate user behavior in existing app prediction approaches is commonly performed with collaborative filtering (CF) techniques [9], where groups of similar users (based on factors such as liking the same item [28] or previous used apps [26]) has been shown to work well. However, CF only exploits usage history

UMAP '20, July 14–17, 2020, Genoa, Italy

ACM ISBN 978-1-4503-6861-2/20/07…$15.00
https://doi.org/10.1145/3340631.3394849

information and explicit user rating feedback, ignoring the context information when people use various apps.

In this work, different from CF, we propose using cohorts to enhance *app category usage prediction* by exploiting various dimensions, including contexts, user characteristics and user interactions. Our method creates predefined cohorts covering three aspects: demographics (e.g., age, gender), psychographics (e.g., interests, way of living) and behavioral patterns (e.g., engagement frequency, revisitation patterns). Rather than limiting ourselves to these predefined sets, we also propose cohorts modeling methods that assign users to a combination of cohorts. We demonstrate through extensive experiments with a large-scale app usage log data that our cohort modeling methods can yield significant improvements over a personalized prediction model.

Finally we demonstrate that compared to existing approaches, our proposed cohort modeling method can significantly alleviate the *cold-start* problem, as it can achieve strong predictive performance for new users, even with limited amount of user interactions available. Moreover, users' interpretable cohort information can provide more transparency and expose the reasoning behind the prediction, which has been shown to be useful in improving the effectiveness of such recommendations [37]. Note that we do not directly compare our proposed approach with CF in this work given it is difficult to model all the dimensions we consider into the CF framework. Rather, our main focus is to demonstrate the effectiveness of our cohort-based approach, compared to personalized models, especially for the cold-start scenario.

Our papers make the following contributions:

- We establish a comprehensive taxonomy to generate cohorts using logs readily available for mobile app usage: demographics, psychographics, and behavioral patterns.
- We demonstrate that modeling user interests within these cohorts can enhance state-of-the-art app category usage prediction personalization methods, leading to significant gains in the prediction performance.
- Our proposed cohort modeling method can effectively alleviate the user cold-start issues compared with the personalized prediction models, especially when limited amount of user interaction data is available.

To our knowledge, our paper is the first to extensively utilize users' cohort information in predicting large-scale mobile app category usage.

2 RELATED WORK

There are three relevant areas: (1) personalization of app usage prediction based on temporal and contextual features; (2) enhancing the recommendation system based on user grouping method; and (3) approaches for mobile user modeling.

Most of the previous research work on app usage prediction (exact app or app category) is based on both the temporal patterns and sensor signals collected, and only the personalized prediction model is explored. Tan et al. [20, 22, 32] tried to predict the app usage patterns based on the periodic pattern and specific using times for each app. Huang et al. [2, 13, 30, 36] applied contextual information in app usage prediction, like location (based on Wifi access points) and user profile configuration (silent mode, etc.). In those works, the number of participants is small or from one social community (e.g. college students). Only the recent research from [2] uses a large-scale dataset to perform their prediction model, which is also the only work analyzed the user cold-start issues by leveraging the app installation list as the prediction basis for new users. Most of the other works do not work well for cold-start. In this paper, we propose to predict users' app category usage based on cohorts and show that this method help alleviating the user cold-start problem.

Collaborative filtering algorithm (CF) [1, 31] can also be used to find people with similar interests and leverage their activities and preferences to provide relevant recommendations. However, the app usage prediction problem differs from traditional collaborative filtering settings, such as the Netflix rating prediction problem, in many aspects. First, user interaction with items such as apps is *brief* and *repetitive* in nature, whereas items like movies and books are usually watched/read once. Second, the user feedback of app usage is inherently implicit in the form of item clicks, as opposed to explicit feedback like ratings or comments. Additionally, app usage has a temporal ordering of clicks within sessions. Lastly and most importantly, app usage prediction must be made available *dynamically* as the user interacts with the system. The cold-start problem also exists in CF recommendation systems [6, 23, 29, 40] and Natarajan et al [26] tried to solve it by clustering users based on their (sparse) one-step item transition probabilities. In our work, we propose a user cohort modeling method that goes beyond item transition patterns.

Some researchers looked at modeling smartphone users based on their app usage behaviour. Jones et al. [15] identified three distinct clusters of users based on their app revisitation patterns: *checkers* who exhibit brief but quick revisit patterns, *waiters* who are split between short-medium revisitations and long revisitations, and *responsives* who exhibit sometimes brief and sometimes long revisit patterns. Zhao et al. [38] identified 382 distinct kinds of users from more than 10,000 individuals. In their work, users are represented by the average usage weight of each app category in different time periods. Zhao et al. [16, 38] also show that the way users engage with their apps is related to their demographics. Finally, Li et al. [18] reported how the choice of device models impact app selection, revealing the significance of device models on app usage.

To summarize, although many studies have been conducted on mobile app usage prediction and mobile user modeling, no existing works have been conducted on modeling the user cohort for the purpose of app category usage prediction. In addition, less research has been undertaken on the user cold-start problem. In this work, we aim first to answer the question of whether users' app category usage can be predicted based on users' cohorts information. If yes, we then aim to answer our second question of whether the proposed framework can help the user cold-start issue. For simplicity, we will often refer to app usage to mean app category usage in the rest of this paper, unless otherwise stated.

3 DATA OVERVIEW

The dataset used in this paper is collected from a mobile analytics and advertising platform at Yahoo. We collected a sample of mobile usage logs from a week in March 2017 from US-based users. Each

Table 1: Statistics of our dataset.

Age	%users	Gender	%users
13-17	7.1%	female	51.8%
18-24	15.9%	male	48.2%
25-34	22.8%	**OS**	**%users**
35-54	43.9%	android	57.3%
55+	6.9%	ios	42.5%

log consists of the user's general app usage information, such as demographics, operating system, timestamp, app category, and app usage duration. All the data was anonymized by removing all personally identifiable data prior to processing. To reduce bias from users with low level of engagement, we restricted our sample to those users who interacted with apps from at least five different categories. These steps resulted into a dataset of approximately 400,000 sessions, 4,000 unique apps and 5000 users.

Table 1 shows some statistics. 51.8% of app users are female, and most of the logs are generated by users between 35 and 54 years old. 99.8% of the devices are operated by Android or iOS. Our dataset contains 45 app categories (consistent with the Google Play App taxonomy [10]), ranging from social, communication to business. The most popular app categories include social, lifestyle, productivity, tools and utilities.

4 COHORT MODELING

One goal of this work is to extract informative and interpretable user cohorts based on different aspects of users' characteristics and app usage behaviors. The cohorts are used to draw "portraits" of users, which we believe will help predicting app category usage. A user cohort is a group of people who share common characteristics or experiences within a defined time-span. From previous work, three major dimensions have been used to classify users [11]: demographics, psychographics and behavioral. Demographics is the most popular dimension; it includes age, gender, occupation, education, religion, race, and location. Psychographics brings a better understanding of the users as a person by measuring psychological aspects, such as the way of living (lifestyle), interests and opinions [41]. Finally, the behavioral dimension focuses on the actual behavior of users, and includes spending/consumption habits, session frequency, usage rate, and loyalty status. In this work, we use these three dimensions to develop user cohorts based in a one-week time window. The taxonomy of user cohorts is detailed in Table 2.

4.1 Demographics

The first user cohort is based on user demographics. In our case, these are age, gender, and operating systems. Some studies have shown that age and gender have an important impact on how users use apps on their smartphones [38]. For example, male users may be more engaged with sports apps and female users use more shopping apps. We group users into two gender cohorts: male and female, and five age cohorts: 13-17, 18-24, 25-34, 35-54 and 55+. Li et al. [18] reported how the choice of device models can impact the adoption of app stores, app selection and abandonment, online time, and data plan usage. Their work revealed the significance of device models against app usage, and suggest taking into account the device models as an essential factor in app recommendation tasks. We use operating system and group users into three cohorts accordingly: Android, iOS, and others.

4.2 Psychographics

Besides demographics, the psychological characteristics of a user, such as specific needs, preferences, and interests may also be a strong driver of app usage. For example, a young man interested in cooking may tend to use more recipe apps even if this category of apps is not broadly popular within his demographic group. Such differences between individuals and the communities to which they belong might be reflected in app usage. Thus, considering the psychographics of users may provide important insights in predicting app usage. We define user cohorts from three aspects of users' psychographics (see Table 2): interests, the way of living, and communities.

4.2.1 ***Interests.***

Users' specific interests may be indicative of future intent on app usage. For example, users who love sports might potentially access sports apps and consume more sports-related content than others. Therefore, it is important to consider users' interests when predicting app usage. To group users based on their app preferences, keeping the most popular app categories for each user based on their historical app access frequency is a straightforward way of doing this. However, previous researchers [19, 27] have found that the app popularity distribution follows Zipf's law, which indicates that only a few apps have high installation/usage whereas many apps have low installation/usage. Users grouped by their *absolutely* top app categories may lead to a skewed distribution meaning that most users may be classified into a few cohorts, mostly highly popular apps, such as social-networking, productivity, and communication.

Other than selecting the *absolutely* top app categories, we employ another strategy to select the *relatively* most popular app categories for each user by normalizing across all users. This strategy can properly represent users' app category preferences as well as keeping users' specific preference characteristics. Specifically, the top k apps for user u are selected based on the popularity score $P(a, u)$ for each app a, which is calculated based on usage frequency of that app category for user u and the usage frequency of the corresponding app category for all users:

$$P(a, u) = \frac{f(a, u)}{\sum\limits_{u_i \in U} f(a, u_i)} \tag{1}$$

where app $a \in A$, and A is the set of all the app categories engaged by u. $f(a, u)$ represents the usage frequency of the app category a for user u in a given period while U is the set of all users. Given this popularity score $P(a, u)$, we can select the top k app categories to represent the interests of user u. Through this normalization, the "niche" popular app categories are used for representing that user.

Selecting appropriate k for the top k app categories can be also crucial for the user representation. Consistent with prior results [2, 33, 34], we find that if we deem k as the number of all the app categories used by that user, 77.3% of the users can be uniquely identified (i.e., each of those users belongs to a user cohort that consists of exactly that one user). To empirically evaluate this, we present the results of the cohorts generated by varying k from 1 to

Table 2: User Cohort Taxonomy

Taxonomy	Features	Cohort Modeling Dimension	Cohort Dimension Illustration	Cohort Label Amount
Demographic	"Physical" Attributes	Gender	Female, Male	2
		Age	13-17, 18-24, 25-34, 35-54, 55+	5
	Technographics	Operation System	iOS, Android and Others	3
Psychographic	Interests	Absolute App Category Interests	Absolute Top K Preferred App Categories	45-961
		Relative App Category Interests	Relative Top K Preferred App Categories	45-1.5k
	Way of Living	Get up Time	Late-riser, Normal, Early-bird	3
		Bed Time	Night-bird, Normal, Early-to-bed	3
		Nocturnal phone use	Heavy-use, Normal, Light-use	3
	Community	Temporal App Interests	night communicators, evening TV watchers, weekend morning gamers, etc.	114±7
Behavioural	Engagement	Time Spent	Tourists, interested, average, active and VIP	5
		Access Frequency	Tourists, interested, average, active and VIP	5
	Revisitation	Revisitation Patterns	Checkers, Waiters, Responsives	3

Table 3: Description of the cohorts generated based on different selected app category sets (% User Identified: percentage of users are uniquely identified by the selected app categories; # Cohorts: number of cohorts generated; Avg. # of Users: average number of users in each cohort; Std: standard deviation of the number of users in each cohort.)

Interests Representation	% User Identified	# Cohort	Avg. # of Users	Std.
Absolutely Top 1	0%	45	65	142.67
Absolutely Top 2	4.0%	364	8	18.62
Absolutely Top 3	18.5%	961	3	5.01
Relatively Top 1	0%	45	62	37.82
Relatively Top 2	4.7%	582	5	5.21
Relatively Top 3	33.5%	1522	2	1.71

3, as shown in Table 3. We can observe that by selecting k equal to 3, most of the user cohorts consists of only 2-3 users whereas many users can be uniquely identified (i.e., many cohorts only consist of exactly one user). Therefore, we enumerate different interests representation, setting $k \leq 3$ in the rest of the work.

4.2.2 *Way of Living.*

We focus on when a user gets up or goes to bed, and how actively the user uses the phone during the night (midnight to 6 AM).

Get-up & Bed time. Murnane et al. [25] found that users' smartphone app usage patterns vary for individuals with different body clock types. In this work, we focus on when a user gets up or goes to bed. Zhao et al. [39] identified the get-up time and bed time by the charge cycle of smartphone batteries; however, this information is not available in our dataset. Following a similar methodology to [39], we assume that the users start to "stop" using the phone before getting to sleep and pick up the phone when they get up. If there is an idle time of phone usage for longer than 4 hours at night (i.e., the idle time starts between 8 PM and 5 AM; ends between 4 AM and 1 PM), we identify it as the sleeping time. We then use the timestamp of the start and end of this sleeping time respectively as the “go-to-bed” time T_b and “get-up” time T_g.

Nocturnal Phone Usage. This measures how actively a user uses the phone during the night. Following the methodology in [39], the total duration of all the active periods of app usage during night time (between midnight and 6 AM) D_n is computed as the feature for representing nocturnal phone usage.

After obtaining those variables (go-to-bed time T_b, get-up time T_g and nocturnal usage duration D_n), we need to further group them into user cohorts. We are particularly interested in those traits that make the users different from others. Following from [39], we first normalize those discrete features using z-score. By assuming those features follow the Gaussian distributions, we then calculate the mean and standard deviation for each of those features. As shown in prior work [39], those features far away from the mean of the feature with more than one standard deviation (std) can be utilized for representing the special user traits. Therefore, as shown in Table 2, for each “way of living” feature, a pair of semantic labels is generated for two ends of the feature distribution, i.e., lying outside of the interval of (mean ± std). For example, based on the distribution of get-up time T_g for all users, if a user gets up within the time period of mean ± std, we will label his/her cohort as “normal”. Otherwise, we will label him/her as “later-riser” if the user gets up later than the timestamp of “mean+std” and as “early-bird” if he/she gets up earlier than the timestamp of “mean-std”.

4.2.3 *Communities.*

Several studies have clustered users into different communities based on their temporal app usage patterns. For example, Zhao et al. [38] identified 382 distinct kinds of users using their clustering method based on the usage frequency of different app categories during specific time periods. Within their proposed clustering method, they identified different types of users and ultimately label them with a community label, such as night communicators, evening learners and car lovers. Different from *interests* described in Sec. 4.2.1, the *communities* capture the more fine-grained temporal app usage patterns.

In our work, we utilize the same methodology to assign each user to different communities. Based on our dataset, each user is represented by a vector C_v of 45 (categories) x 4 (time periods) x 2 (weekends and workdays) for a total of 360 dimensions. By applying the best performing k-means-MeanShift hybrid clustering algorithm described in [38], we obtain a total of 114±7 clusters.[1] This k-means-MeanShift clustering algorithm combines the benefits of multiple standard clustering algorithms, is computationally feasible, and finally is able to automatically determine the ultimate number of clusters. We find that our clustering results are similar to the findings in [38] that many meaningful communities exist in our clusters, such as night communicators, evening TV watchers and weekend morning gamers.

[1] Since we use 5-fold cross validation in our performance evaluation, different number of clusters are generated in different folds.

4.3 Behavioural

The third dimension is based on behavioral patterns. We consider two aspects of users' behavioral characteristics, which we show in Table 2: engagement and revisitation.

4.3.1 ***Engagement.*** Within the context of web browsing, Lehmann et al. [17] created five types of user groups based on their frequency of visiting the site over a month: tourists, interested, average, active and VIP users. They identified that the proportion of specific types of users based on their engagement will be different across websites. In this work, we also hypothesize that users with different engagement levels may behave differently in terms of app usage. Additionally, we measure not only users' engagement in terms of how frequently they access apps, but also the more fine-grained total time spent (i.e., total dwell time). The latter represents the total duration of users accessing mobile apps in the given period.

Following the five-level engagement level definition in [17], we propose the following strategy for defining users' engagement cohorts based on their mobile app behavior patterns. We group users into five different engagement cohorts of duration and frequency, respectively, by using the quantiles at 20%, 40%, 60%, 80% and 100% as the breakpoints, resulting in those five cohorts: tourists, interested, average, active and VIP.

4.3.2 ***Revisitation.*** Jones et al. [15] present a revisitation analysis of smartphone use. They propose that users could be clustered into three different types based on their revisitation patterns, where a revisitation curve for a particular user is constructed by considering the in-between duration in launching *any* app on their phone. They grouped the users based on the revisitation curves into three user cohorts: *checkers* (users exhibiting brief revisit patterns lightly skewed towards fast revisitation of less than 4 hours), *waiters* (users exhibiting longer revisit patterns longer than 16 hours), and *responsives* (users exhibiting a hybrid of brief and long revisit patterns).

Following [15], we use an exponential scale for revisit interval bins, which are 1, 2, 4, 8, 16 and 32 minutes; 1, 2, 4, 8, 16, 32 hours (i.e 1.3 days), 64 hours (i.e. 2.6 days), 128 hours (i.e. 5.3 days), and above (i.e >5.3 days). A revisitation curve characterizes a user by its 15-dimensional vector R_v, where each dimension corresponds to the frequency of revisits within the corresponding bin. These curves are like a "signature" of users' behavior in launching mobile apps. We iteratively apply k-means for a varying number of clusters and use within-groups sum of squares to plot the variations as a function of the different number of clusters. We then pick the "elbow" of the curve as the optimal number k of clusters. Based on this simple k-means method, we identify a substantial trichotomy of user cohorts within their revisitation patterns: checkers (which accounts for 39.6% of users), waiters (13.5%) and responsives (46.8%).

5 APP CATEGORY USAGE PREDICTION BASED ON COHORT MODELING

In this section, we introduce our approach of using cohort modeling for the purpose of app category usage prediction. We start by formalizing the prediction problem, and then discuss how to assign the cohort information when a new user comes. To assign users into cohorts of different granularity, we also describe how we build combined user cohorts. Finally, we evaluate the performance of the proposed prediction method, including for the user cold-start problem.

5.1 Problem Formulation

In our work, we aim to predict the app category a new user will use based on their cohort information. Our aim is to overcome the data sparsity issue and to guarantee efficient real-time prediction. Each user can be characterized by his/her cohort information, i.e., demographics, psychographics, and behavioral patterns. For each user, given the current time (time of day and day of week) and his/her cohorts information, we want to predict the app category he/she will use. This prediction task can be formalized as a multi-class prediction problem. There are $K = |A|$ classes for this prediction task, where A is the set of all app categories in the dataset.

The **app category usage prediction problem** is formally defined as follows: Given a list of app categories $\{a_1, a_2, ..., a_i\}$, the users' cohorts information C and temporal context T, the problem of app usage prediction is to find an app category $\hat{a}$ that has the highest probability of being used under C and T. Specifically, we aim to solve:

$$\hat{a} = \underset{a_i \in A}{\operatorname{argmax}} \, P(a_i | C, T)$$

5.2 User Cohorts Assignment for New Users

Since the user cohort based approach aims at addressing the user cold-start problem, we split the training and test set based on users instead of log entries. In the training set, the users are assigned to the specific cohorts based on their usage logs. For example, the community cohorts are generated based on the clustering results of users in the training set. During the test stage, for all test users whom have not been seen by the system before, we assign those users to existing cohorts in the training set and then proceed to the prediction task. During the assignment process, we need to compare the new users with all "old" users in the vector space, so we first introduce the representative vectors for representing users' cohorts information as vectors. For scalar based cohort: e.g., demographics, the representative vector refers to the one-hot encoded vector of the categorical scalar. For cluster-based cohorts: e.g., community and revisitation, the representative vector is the vector used in clustering. We then propose three assignment approaches to assign the test/new users to the most accurate cohort within all the different cohorts taxonomies as described in Section 4.

5.2.1 ***Nearest centroid classifier (NCC).*** If we want to determine which existing cohort a new user belongs to, the straightforward way is to find the nearest cohort. Therefore, we propose to use the nearest centroid classifier [35] as the first assignment methodology, which is a classification model in machine learning that assigns observations to the class of samples whose centroid is closest to the observation. In our scenario, given existing users' representative vectors $\{(\overrightarrow{x}_1, y_1), ..., (\overrightarrow{x}_n, y_n)\}$ with cohort labels $y_i \in Y$, we compute the per-cohort vectors:

$$\overrightarrow{\mu_l} = \frac{1}{|C_l|} \sum_{i \in C_l} \overrightarrow{x}_i$$

where C_l is the set of indices of samples belonging to cohort label $l \in Y$; the assignment function for the cohort label assigned to a new user $\overrightarrow{x'}$ is:

$$\hat{y} = \underset{l \in Y}{\operatorname{argmin}} \|\overrightarrow{\mu_l} - \overrightarrow{x'}\|$$

5.2.2 ***K-nearest neighbor classifier (KNNC).*** The second approach is to apply the k-nearest neighbor (KNN) [8] rule, which is one of the most straightforward non-parametric techniques in pattern classification. The basic idea of k-nearest neighbor classifier is: an object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. Here we set $k = \sqrt{n}$, where n is the amount of unique cohort labels. Similar to NCC, representative vectors are used when calculating the distance between two users. More specifically, given a new user x' and a similarity metric d based on Euclidean distance, KNN classifier performs the following two steps: (1) It runs through the whole training set computing d between the new user x' and each user in the training set. We state the k users in the training set that are nearest to x' as the set C. (2) It then estimates the conditional probability for each cohort label, that is, the fraction of users in C with that given cohort label:

$$\hat{y} = \underset{j \in Y}{\operatorname{argmax}} P(y = j | X = x') = \frac{1}{k} \sum_{i \in C} I(y^{(i)} = j)$$

where $I(x)$ is the indicator function which evaluates to 1 when the argument x is true and 0 otherwise. Finally, the new user x' gets assigned to the cohort label with the highest probability.

5.2.3 ***Classifiers trained based on the existing cluster labels (RF).*** Given the NCC assignment, some of the labels will be assigned based on the centroids of clustered results. However, assigning new points based on distance in a clustering algorithm is complex because the results of a clustering algorithm may be imperfect; they only present a snapshot of a (hopefully good) segmentation within the current data. With more data coming in, the cluster may change. Therefore to make the assignment more robust given a particular clustering segmentation, we can train an additional classifier where the resulting clusters are treated as different classes. In that way, we can account more intuitively for the non-robustness of the clustering labels. Besides, as we expect the clustering to reflect "some structure", it is a cheap and straightforward way to encapsulate that structure. Following this, the classifier learns $P(c|x)$ based on users' representative vectors and corresponding cluster labels. When a new user x' appears, we can directly predict which class the new user belongs to instead of assigning him/her based on distance or neighbors. So here, we propose to use an additional classifier based on the Random Forest algorithm, which is efficient in assigning the cohort label to new users:

$$\hat{y} = \underset{c \in Y}{\operatorname{argmax}} P(c|x')$$

5.3 Combination of Multiple User Cohorts

Besides predicting users' next app category solely based on one type of cohorts, we also want to consider multiple types of cohorts together. For example, as shown in Figure 1, a combined cohort with different types of demographic cohort features could be generated to describe a user. This combined cohort would have 30 different labels since there are potential 30 different compositions based on age, gender, and operating system cohort labels. For instance, we can observe one of the combined cohorts in Figure 1: "Android male user cohorts aged 25-34". It is possible to generate a new cohort based on a combination of any cohort dimensions listed in Table 2.

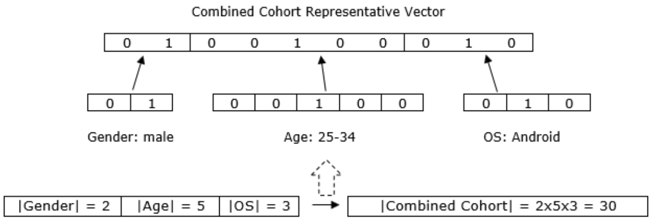


Figure 1: Structured representative vector of combined user cohorts: combining age, gender and operation system groups would potentially result in 30 different cohorts.

The cohort assignment for new users within these combined cohorts follow the same methodologies described in Section 5.2. The representative vectors x are updated by concatenating the original vectors x_i of each selected cohort:

$$x = [x_1, x_2, ..., x_i], i \in S_c$$

where S_c is the set of selected cohorts to be combined. Figure 1 illustrates the generation of such new representative vector.

5.4 Experimental Results

In this section, we firstly empirically demonstrate how our proposed cohorts can be used to improve the prediction of users' app category usage. Secondly, we investigate whether our proposed approach can help addressing the user cold-start issue when compared with other prediction mechanisms.

5.4.1 ***Experimental Setup.*** We apply 5-fold cross-validation to evaluate each model. At each time, we split all the users into training, validation and test set: the logs of three-fold of the users are used as the training set, one-fold is validation set and the remaining one-fold users are used as the test set.

5.4.2 ***App Usage Prediction.*** Our goal is to predict which app category the user will use next. We use a set of state-of-the-art algorithms to build models for our prediction problem: (1) XGBoost (XGB) [4], as an example of ensemble learning method; (2) K Nearest Neighbours (KNN) [5], as an example of the non-parametric method for classification; (3) L2-regularized Logistic Regression(LR) [12], as an example of linear classifier. The parameters in each model, e.g., K in KNN, the number of used trees, the maximum depth of the trees and the learning rate are tuned on the validation sets.

The features include the user cohorts and the temporal context (different hours of a day and days of a week). Additionally, we use the prediction model only based on context features as our benchmark, which takes all the users as they are "the same" (from one cohort). We report four metrics with 5-fold cross-validation: accuracy (acc), precision (pre), recall (rec), and F1-measure (F1). We test the prediction performance of the proposed user cohorts based

Table 4: Measurements of next app category prediction based on different cohorts information. All the results are statistical significant ($p < 0.01$) using the two tailed t-test compared to the temporal context only baseline.

User Cohorts	**# Cohorts**	**Assignment**	**Measurements**			
			acc	pre	rec	F1
Context Baseline						
Hour + Weekday	-	-	0.416	0.173	0.416	0.244
I. Single Cohort						
(a). Demographics						
Age	5	-	0.441	0.197	0.441	0.272
Gender	2	-	0.443	0.197	0.443	0.272
Operating system	3	-	0.447	0.234	0.447	0.291
(b). Psychographics						
Interests: Top1-Absolutely	45	-	**0.686**	**0.666**	**0.686**	**0.665**
Interests: Top2-Absolutely	364	NCC	0.559	0.515	0.559	0.496
Interests: Top3-Absolutely	961	KNNC	0.467	0.387	0.467	0.362
Interests: Top1-Relatively	45	-	0.555	0.528	0.555	0.510
Interests: Top2-Relatively	582	NCC	0.547	0.494	0.574	0.485
Interests: Top3-Relatively	1522	KNNC	0.418	0.393	0.418	0.402
Way of Living: Sleep Time	3	-	0.440	0.203	0.440	0.270
Way of Living: Get-up Time	3	-	0.439	0.206	0.440	0.271
Way of Living: Nocturnal	3	-	0.443	0.197	0.443	0.272
Communities	115	NCC	0.572	0.550	0.572	**0.517**
(c). Behavioural						
Time Spent	5	-	0.444	0.222	0.444	0.286
Frequency	5	-	0.444	0.222	0.444	0.286
Revisitation Pattern	3	NCC	0.444	0.219	0.444	0.283
II. Combinatory Cohort						
(d). Demographics						
Age + OperatingSys	15	NCC	0.443	**0.255**	0.443	**0.298**
Gender + OperatingSys	6	NCC	0.447	0.240	0.447	0.291
Age + Gender	10	NCC	0.436	0.206	0.436	0.271
Age + Gender + OperatingSys	20	NCC	0.436	**0.255**	0.436	**0.298**
(e). Psychographics						
Getup + Nocturnal	6	NCC	0.437	0.236	0.437	0.280
Sleep + Get-up + Nocturnal	17	NCC	0.431	**0.248**	0.431	**0.291**
Top1-Absolute + Community	363	NCC	0.682	0.662	0.682	0.660
Top1-Absolute + Nocturnal	80	NCC	0.681	0.661	0.681	0.659
Top1-Relative + Community	489	NCC	0.577	0.565	0.577	0.543
Top2-Absolute + Community	1034	KNNC	0.572	0.529	0.572	0.514
Top1-Relative +Nocturnal	88	NC	0.548	0.519	0.548	0.505
(f). Behavioural						
Time Spent + Revisitation	12	NCC	0.439	0.244	0.439	0.295
Frequency + Revisitation	12	NCC	0.439	0.244	0.439	0.294
(g). Across Taxonomies						
Age + OperatingSys + Revisitation	69	NCC	0.430	0.290	0.430	0.329
Age + OperatingSys + Time Spent	96	NCC	0.426	**0.292**	0.426	**0.332**
Age + OperatingSys + Top1-Absolute + Time Spent	649	NCC	0.600	0.552	0.600	0.565

on each cohort taxonomy individually and then when combined. We consider several combinations.

Table 4 presents the results.[2] For the combined user cohorts (Table 4.II), since there are a large number of compositions across different cohort taxonomies, we only report the results of the top performing combinations, on which we test the combinations among any two, three or four different cohorts. Note that we also only report those combinations for which we observe a performance boost compared to using any individual cohort feature. For the different cohort assignment methods we employ for new users (see Section 5.2), we only report the one with the best performance.[3]

Firstly, we find that compared to the context-only baseline approach, all the cohort based models achieve better performance on all metrics; all those improvements are found to be statistically significant ($p < 0.01$) through a two-tailed t-test. This demonstrates that not surprisingly, incorporating user characteristics on top of the temporal context help to improve app category usage prediction.

Secondly, we investigate more specifically the gains obtained by the single cohort models (Table 4a-c). We can observe that in general the psychographic cohort models (Table 4b) perform better than the demographic (Table 4a) and behavioural (Table 4c) cohort models by a large margin. When looking at the psychographic cohort models, we can find that the user interests cohort based on the "absolutely top 1 app category", and "communities" are the best predictive models. This indicates that users that have common interests or belong to the same communities may behave more similarly in their app usage behavior, which is not only constrained to their past, but also their future app usage. However, compared to the baseline, we observe only marginal improvements on the "way of living" (Table 4b), behavioral (Table 4c) and demographic (Table 4a) cohort models. This is not surprising as most of those models contain only a small number of cohorts (3-5) and are not sufficiently discriminative.

Finally, when examining the combined cohorts (Table 4d-g), we find that they generally perform better than when using any one of them respectively. For example, all of the "way of living" cohorts outperform any of them when individually used. When combining demographics and revisitation behaviour patterns (Table 4g), we observe an increase of 10% performance improvement, compared to using demographics only (Table 4d). However, it is worth noting that combining "interests" with any other cohorts (Table 4e) would result in only marginal improvements and sometimes even inferior performance. This implies that enriching the cohorts with additional information might not always necessarily help. Another interesting observation is that most of the time, the simple Nearest Centroid Classifier (NCC) cohort assignment approach outperforms KNCC and RF approaches.

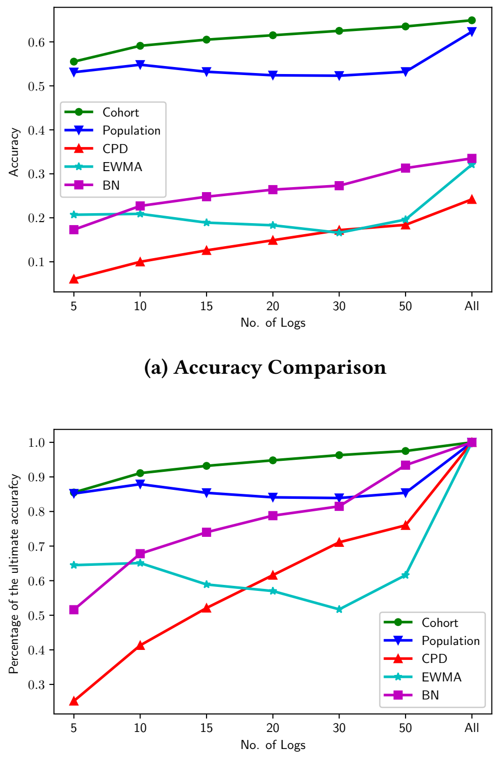
5.4.3 *User Cold-Start Problem.*

In this section, we focus on analyzing whether the cohort-based prediction model can help solving the user cold-start problem. We adopt the best performing cohort model (see Table 4) for the rest of the experiments.

The baseline approaches we compare against are both personalized and population-based prediction models. Although there are many personalized models (see Section 2), some of them are not applicable because they use additional information. We select CPD [32], EWMA [32], and BN [42] as our comparative baselines for personalized models as they can be used with our dataset. CPD (Cumulative Probability Distribution) computes the probabilities of each used app in all the specific time slots based on historical app usage time series for that user, and selects the app with the highest probability at the prediction time based on its time slot.

[2] Due to the limited space, only the results of XGB are reported, which performs the best among all the three classifiers.

[3] We use '-' in Table 4 to denote such scenario when there is no single winner for the three assignment approaches (i.e., all those methods result in the identical cohort assignment).



(a) Accuracy Comparison

(b) Convergence Speed Comparison

Figure 2: Performance comparison among different prediction models with limited amount of historical user logs. Three personalized baseline models: CPD [32], EWMA [32] and BN [42], one population-based baseline model [7] and our proposed cohort model.

EWMA (Exponentially Weighted Moving Average) replaces the cumulative probability in CPD with exponentially weighted moving average [14] so that the newer data points have higher influence in the prediction. BN [42] is a Bayesian Network model that relies on both app usage and time context and calculates a linear combination of the user's last used app and the second last used app for the final prediction. Regarding the population-based prediction models, following [7], we generate the baseline model based on our available predictive features and utilize random forest to combine all those features for the next app category prediction. We combine a set of extensive features that could be extracted from our dataset, which include hour, weekday, last used apps, historical popularity of users' app usage in different time windows, one hour, one day and all history, and periodicity (intervals between app usage) [22].

To explore whether the user cohorts based methodology perform better especially for new users (for which there are limited interaction data), we randomly select 20% of the users in our dataset, and use them to simulate the new users by increasing the number of interaction data (logs) available we consider for each user. Specifically, we extract different amount of logs from 5 to 50 for each user to simulate various severity of the cold-start problem. We hypothesize that models that handle well the cold-start problem tend to perform competitively even with very limited amount of user logs.

Figure 2 presents the prediction performances of all baseline models and our proposed cohort-based prediction model, given different amount of historical logs available for the test users (x-axis). The results are averaged across all the test users. Firstly, we can observe in Figure 2a that when the amount of user interactions is limited, all personalized models perform worse (accuracy is below 35%) than the population baseline models and the cohort models. CPD is the worst for the personalized model when there are logs, followed by EWMA and BN. However, both population and cohort models can achieve over 50% accuracy even with the limited amount of historical user data. Secondly, we observe from Figure2b that only the cohort-based model achieves over 90% of the best accuracy when only 10 entry logs are considered. The performance steadily increases as more and more logs are available. This demonstrates that our proposed user cohort based model outperforms both the personalized models and the population-based model for the user cold-start problem.

6 CONCLUSIONS

In this study, our goal was to identify meaningful user cohorts information to help with the app category usage prediction problem. We show that besides personalized prediction approaches, users' app category usage behavior can be predicted based on cohorts information. Based on our proposed taxonomies of user cohorts modelling, we found that psychographics (interests and community) perform best. Additionally, we identify that our proposed user cohorts based prediction outperforms both the personalized and population-based models on the user cold-start problem.

Through our study, we demonstrated the value of cohorts, especially for new users. This is promising as cohorts information could be used not only on their own but also in combination with other signals as they become more present. For a new user without much interaction data, general cohorts information such as interests or community could be collected, e.g., a user could label themselves as car lovers or young parents. Users' app category usage could be predicted with relatively high accuracy using this basic cohort information. The cohort labels can also be utilized to explain the prediction model, enabling the recommendation to be more transparent and interpretable [37].

There are several limitations of our work, which we would like to address in future work. Firstly, our dataset only consists of relatively short-term app usage and it would be interesting to study signals that could relate to long-term characterizations of user cohorts. Secondly, we mainly focus on next mobile app category prediction in our work. Our method is general while it is worthwhile extending this to further investigate our cohort-based methods on next app prediction [2]. Finally, the cohort taxonomy we define in our work is only a first step proof-of-concept, and can be refined with more fine-grained cohorts when relevant interaction or user profile information become available.

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering* 6 (2005), 734–749.

[2] Ricardo Baeza-Yates, Di Jiang, Fabrizio Silvestri, and Beverly Harrison. 2015. Predicting the next app that you are going to use. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 285–294.

[3] Stuart J Barnes. 2002. The mobile commerce value chain: analysis and future developments. *International journal of information management* 22, 2 (2002), 91–108.

[4] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2* (2015), 1–4.

[5] Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory* 13, 1 (1967), 21–27.

[6] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 191–198.

[7] Trinh Minh Tri Do and Daniel Gatica-Perez. 2014. Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing* 12 (2014), 79–91.

[8] Evelyn Fix and Joseph L Hodges Jr. 1951. *Discriminatory analysis-nonparametric discrimination: consistency properties.* Technical Report. California Univ Berkeley.

[9] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)* 23, 2 (2005), 147–168.

[10] Google. 2018. Select a category for your app or game. https://support.google.com/googleplay/android-developer/answer/113475?hl=en-GB

[11] Fadly Hamka, Harry Bouwman, Mark De Reuver, and Maarten Kroesen. 2014. Mobile customer segmentation based on smartphone measurement. *Telematics and Informatics* 31, 2 (2014), 220–227.

[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

[13] Ke Huang, Chunhui Zhang, Xiaoxiao Ma, and Guanling Chen. 2012. Predicting mobile application usage using contextual information. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 1059–1065.

[14] J Stuart Hunter. 1986. The exponentially weighted moving average. *Journal of quality technology* 18, 4 (1986), 203–210.

[15] Simon L Jones, Denzil Ferreira, Simo Hosio, Jorge Goncalves, and Vassilis Kostakos. 2015. Revisitation analysis of smartphone app use. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1197–1208.

[16] Farshad Kooti, Mihajlo Grbovic, Luca Maria Aiello, Eric Bax, and Kristina Lerman. 2017. iPhone's Digital Marketplace: Characterizing the Big Spenders. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 13–21.

[17] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. 2012. Models of user engagement. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 164–175.

[18] Huoran Li and Xuan Lu. 2017. Mining Device-Specific Apps Usage Patterns from Large-Scale Android Users. *arXiv preprint arXiv:1707.09252* (2017).

[19] Huoran Li, Xuan Lu, Xuanzhe Liu, Tao Xie, Kaigui Bian, Felix Xiaozhu Lin, Qiaozhu Mei, and Feng Feng. 2015. Characterizing smartphone usage patterns from millions of Android users. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*. ACM, 459–472.

[20] Zhung-Xun Liao, Po-Ruey Lei, Tsu-Jou Shen, Shou-Chung Li, and Wen-Chih Peng. 2012. Mining temporal profiles of mobile applications for usage prediction. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*. IEEE, 890–893.

[21] Zhung-Xun Liao, Shou-Chung Li, Wen-Chih Peng, S Yu Philip, and Te-Chuan Liu. 2013. On the feature discovery for app usage prediction in smartphones. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 1127–1132.

[22] Zhung-Xun Liao, Yi-Chin Pan, Wen-Chih Peng, and Po-Ruey Lei. 2013. On mining mobile apps usage behavior for predicting apps usage in smartphones. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 609–618.

[23] Jovian Lin, Kazunari Sugiyama, Min-Yen Kan, and Tat-Seng Chua. 2013. Addressing cold-start in app recommendation: latent user models constructed from twitter followers. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 283–292.

[24] Nielsen Mobile. 2008. Critical mass: The worldwide state of the mobile web. *Nielsen Company* (2008).

[25] Elizabeth L Murnane, Saeed Abdullah, Mark Matthews, Matthew Kay, Julie A Kientz, Tanzeem Choudhury, Geri Gay, and Dan Cosley. 2016. Mobile manifestations of alertness: Connecting biological rhythms with patterns of smartphone app use. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 465–477.

[26] Nagarajan Natarajan, Donghyuk Shin, and Inderjit S Dhillon. 2013. Which app will you use next?: collaborative filtering with interactional context. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 201–208.

[27] Thanasis Petsas, Antonis Papadogiannakis, Michalis Polychronakis, Evangelos P Markatos, and Thomas Karagiannis. 2017. Measurement, modeling, and analysis of the mobile app ecosystem. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)* 2, 2 (2017), 7.

[28] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM, 175–186.

[29] Alan Said, Ernesto W De Luca, and Sahin Albayrak. 2010. How social relationships affect user similarities. In *Proceedings of the International Conference on Intelligent User Interfaces Workshop on Social Recommender Systems, Hong Kong*.

[30] Choonsung Shin, Jin-Hyuk Hong, and Anind K Dey. 2012. Understanding and prediction of mobile application usage for smart phones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 173–182.

[31] Xiaoyuan Su and Taghi M Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* 2009 (2009).

[32] Chang Tan, Qi Liu, Enhong Chen, and Hui Xiong. 2012. Prediction for mobile application usage patterns. In *Nokia MDC Workshop*, Vol. 12.

[33] Zhen Tu, Runtong Li, Yong Li, Gang Wang, Di Wu, Pan Hui, Li Su, and Depeng Jin. 2018. Your Apps Give You Away: Distinguishing Mobile Users by Their App Usage Fingerprints. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 138.

[34] Pascal Welke, Ionut Andone, Konrad Blaszkiewicz, and Alexander Markowetz. 2016. Differentiating smartphone users by app usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 519–523.

[35] Wikipedia. 2018. Nearest centroid classifier — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Nearest_centroid_classifier [Online; accessed 4-February-2018].

[36] Tingxin Yan, David Chu, Deepak Ganesan, Aman Kansal, and Jie Liu. 2012. Fast app launching for mobile devices using predictive user context. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*. ACM, 113–126.

[37] Yongfeng Zhang, Xu Chen, et al. 2020. Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101.

[38] Sha Zhao, Julian Ramos, Jianrong Tao, Ziwen Jiang, Shijian Li, Zhaohui Wu, Gang Pan, and Anind K Dey. 2016. Discovering different kinds of smartphone users through their application usage behaviors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 498–509.

[39] Sha Zhao, Yifan Zhao, Zhe Zhao, Zhiling Luo, Runhe Huang, Shijian Li, and Gang Pan. 2017. Characterizing a user from large-scale smartphone-sensed data. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM, 482–487.

[40] Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. 2011. Functional matrix factorizations for cold-start recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 315–324.

[41] Ruth Ziff. 1971. Psychographics for market segmentation. *Journal of Advertising Research* 11, 2 (1971), 3–9.

[42] Xun Zou, Wangsheng Zhang, Shijian Li, and Gang Pan. 2013. Prophet: What app you wish to use next. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM, 167–170.