Network Research Workshop
Proceedings of the Asia-Pacific Advanced Network 2014 v. 38, p. 65-77.
http://dx.doi.org/10.7125/APAN.38.10 ISSN 2227-3026

Mining and Predicting Smart Device User Behavior

Yuntao Fan , Hewu Li [1], Qian Wu , Wenqi Sun

Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing, China

E-Mails: fanyt1988@gmail.com; lihewu@cernet.edu.cn; wuqian@cernet.edu.cn; swq_persistence@126.com

* Tel.: +86-010-6260-3307; Fax: +86-010- 6260-3308

Abstract: Three types of user behavior are mined in this paper: application usage, smart device usage and periodicity of user behavior. When mining application usage, the application installation, most frequently used applications and application correlation are analyzed. The application usage is long-tailed. When mining the device usage, the mean, variance and autocorrelation are calculated both for duration and interval. Both the duration and interval are long-tailed but only duration satisfies power-law distribution. Meanwhile, the autocorrelation of both duration and interval is weak, which makes predicting user behavior based on adjacent behavior not so reasonable in related works. Then DFT (Discrete Fourier Transform) is utilized to analyze the periodicity of user behavior and results show that the most obvious periodicity is 24 hours, which is in agreement with related works. Based on the results above, an improved user behavior predicting model is proposed based on Chebyshev inequality. Experiment results show that the performance is good in accurate rate and recall rate.

Keywords: smart device; user behavior; autocorrelation; predicting model.

1. Introduction

Mobile services and applications have experienced explosive development in recent years. All the personalized services are based on the understanding of user behavior. Smart device is the

most intimate equipment for users, and thus the mining of smart device usage behavior is the most important area of mining user behavior, and can contribute much to personalized services.

There have been many researches on the mining of mobile user behavior. [1] predicted user's mood by mining the usage of smart devices. [2] focused on the payment behavior on smart devices. [3]studied how applications are used to save energy. [4-5] recommended applications by analyzing applications usage behavior on smart phones. [6] classified applications by natural language processing method using the data from app store. [7] studied the relationship of application usage and geographical position. [8-9] collected much information such as position, time and sensor data and predicted user behavior. [10] classified users by their behavior.

Although there are many researches on mining mobile user behavior, the focuses of these researches are various. There still lack researches on the mining of application usage, smart device usage and time feature of user behavior. This paper collects sufficient data and mines user behavior on the above three aspect.

2. Approach

2.1. Data Collection

The following three types of data are collected in this paper: application usage, smart device usage, and application installation.

For the application usage, the data format is ($user_i$, $time_j$, app_k), which means $user_i$ uses app_k at $time_j$. In Android, this can be obtained by method getRunningTasks() of class ActivityManager. The traditional telecommunication application, such as call and SMS are filtered out in this paper.

For the smart device usage, the data format is ($user_i$, $time_{j1}$, $time_{j2}$), means $user_i$ begins to use smart device at $time_{j1}$, and stops using it at $time_{j2}$. The start and end of using smart device is reflected in the on/off state of device's screen. In Android, this can be obtained by registering a BroadcastReceiver which can receive the event ACTION_SCREEN_ON and event ACTION_SCREEN_OFF.

The application installation can be obtained by method getInstalledPackages() of class PackageManager in Android. The build-in applications are not collected in this paper, such as phone, SMS, settings and so on.

The data collection code is integrated in specific version of the application At Tsinghua[11-12]. Users are notified of the data collection by an announcement and users can choose to decline the data collection.

From 4th, December, 2013 to 4th April 2014, there are 2690 users accepting the data collection. Users are identified by the MAC address of smart device.

2.2. Application Usage Statistics

There are 14,293 different applications installed by the 2690 users. Of all these users, the maximum application installation is 226, and the minimum is 1. The average application installation is 39.85, and the standard deviation is 26.88. The most popular applications is shown in Table 1.

Table 1. The top 20 application installation

installation	application	installation	application
2690	AtTsinghua	1011	WPS Office
2006	QQ	920	Fetion
1960	Wechat	828	Alipay Fast payment
1333	Baidu map	751	Adobe Flash Player
1328	Youdao Dictionary	689	Taobao
1328	RenRen	673	360 assistant
1166	UC Browser	642	wandoujia
1120	Alipay Wallet	619	public comments
1103	Sina microblog	564	Baidu Cloud
1054	Sogou input method	545	Adobe Reader

Of all the installed applications, some are frequently used and some are rarely used. The usage frequency is apparently long-tailed, as shown in Table 2. Notably, there are nearly 70 percent applications never used during the four months.

Table 2. The usage frequency of all applications

number of applications	usage frequency
1	≥1,000,000
12	≥100,000
99	≥10,000
173	≥5,000
449	≥1,000
643	≥500
996	≥200
1295	≥100
2626	≥10
4252	≥1

10041	0

The most frequently used application is shown in Table 3.

Table 3. Top 20 frequently used applications

application	usage	application	usage
Wechat	1813680	Chrome	114182
QQ	582805	QQ Browser	105393
UC Browser	453541	GOLockscreen	92735
MiLocker	369451	Baidu	91986
Renren	332480	91 Assistant	79637
Browser	240559	iReader	69526
Word lockscreen	178188	Youku video	64426
At Tsinghua	158121	MiHome	60737
Push Service	139911	Mini Thunder	55581
Baidu Postbar	121729	GO Safe home	55572

Applications are not independent from each other. In a specific period of device usage, users usually switch from one application to another. A period means users are using devices all the time during the period, when the screen is never off. The switch behavior reflects the correlation of applications. To describe this, a n×n matrix C is introduces, where n means the number of all applications. In a period of device usage, if users switch from app_i to app_j, then c_{ij}++. The correlation of applications is shown in Table 4.

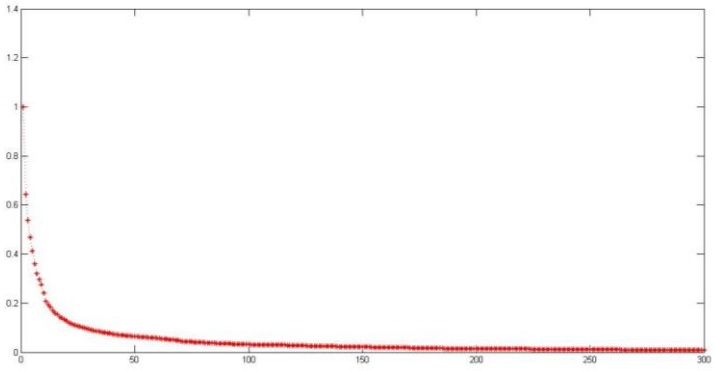
Table 4. Top 20 correlations of application pairs

application pairs	correlation	application pairs	correlation
Wechat: QQ	20546	Wechat: AtTsinghua	6734
Wechat: Renren	19713	Renren: QQ	6560
QQ: Wechat	19265	Wechat: Browser	5300
Renren: Wechat	15628	QQ: UCBrowser	4938
Wechat:UCBrowser	12993	Browser: Wechat	4870
AtTsinghua: Wechat	11792	UCBrowser: QQ	4579
UC Browser: Wechat	11360	AtTsinghua: QQ	4349
Wechat: MiLock	10703	Renren: UCBrowser	3796
MiLock: Wechat	10547	AtTsinghua:UCBrowse	3660
QQ: Renren	7716	MiLock: QQ	3474

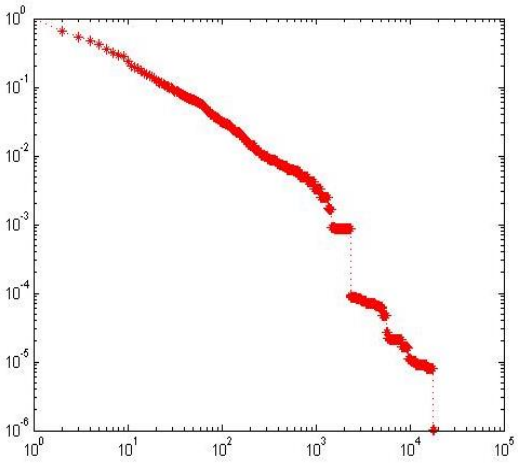
2.3. Device Usage Statistics

First the total time of smart device usage in one day is calculated. In the span of four months, of all the collected users, the longest time of device usage is 324.2 minutes in one day, and the shortest usage is 2 minutes, with non-use excluded. The average usage is 53.0 minutes, and standard deviation is 42.4 minutes. So we can see that there are no giant gap between the most active users and the least active users.

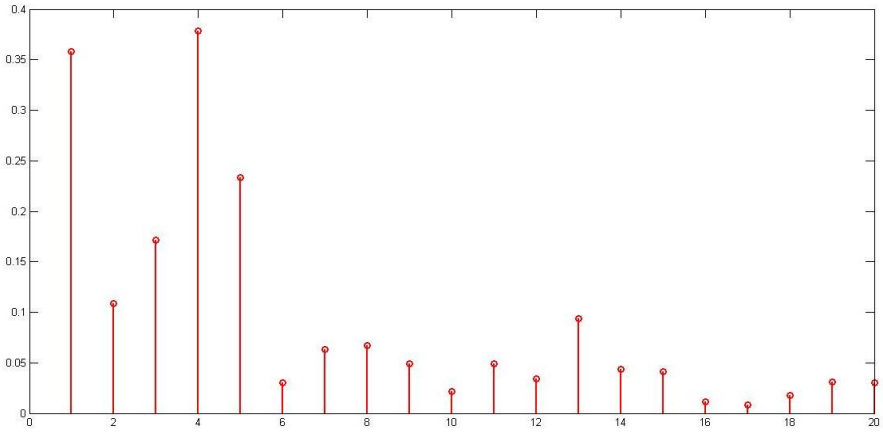
And then the duration of device usage at a time is analyzed. Here duration has the same meaning as period in 2.2, during which the screen is never off. Of all the durations, the average duration is 60.9 seconds, the standard deviation is 241.5 seconds, the maximum is 299.3 minutes, the minimum is 0.7 seconds and the coefficient of variation is 396.6%. As for one user, the duration is also different. The CDF (cumulative distribution function) of all these durations is shown in Figure 1(a). The function in log-log coordinates is nearly linear, as shown in Figure 1(b). Through the R-square test, the correlation coefficient is 0.9373, so it is concluded that the duration of smart device usage obeys the power-law distribution. And then the autocorrelation is analyzed in this paper. Autocorrelation analysis is usually used to reflect the degree of correlation between the values of the same sequence in different time. The first twenty points of autocorrelation are calculated and shown in Figure 1(c).



(a)



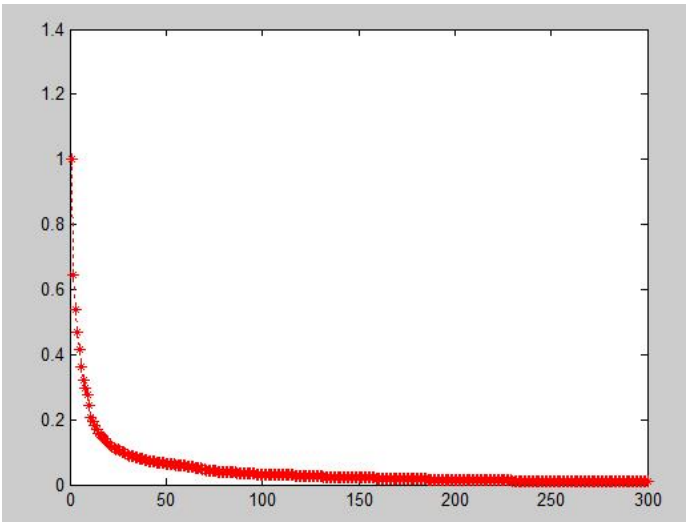
(b)



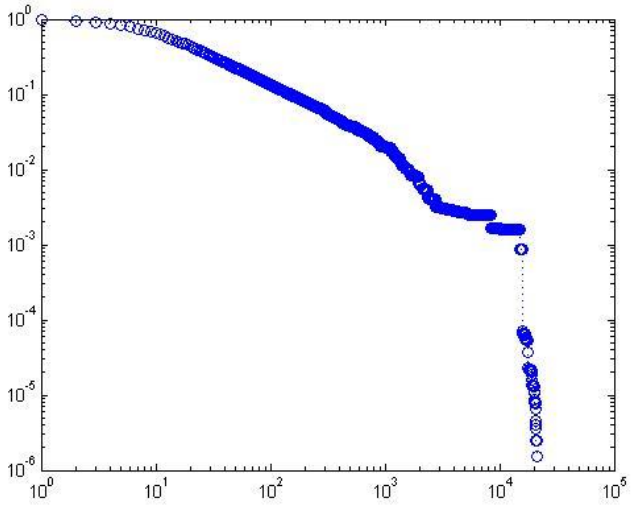
(c)

Figure 1. (a) The cumulative distribution function of durations. **(b)** The CDF in log-log coordinates. **(c)** Autocorrelation of durations.

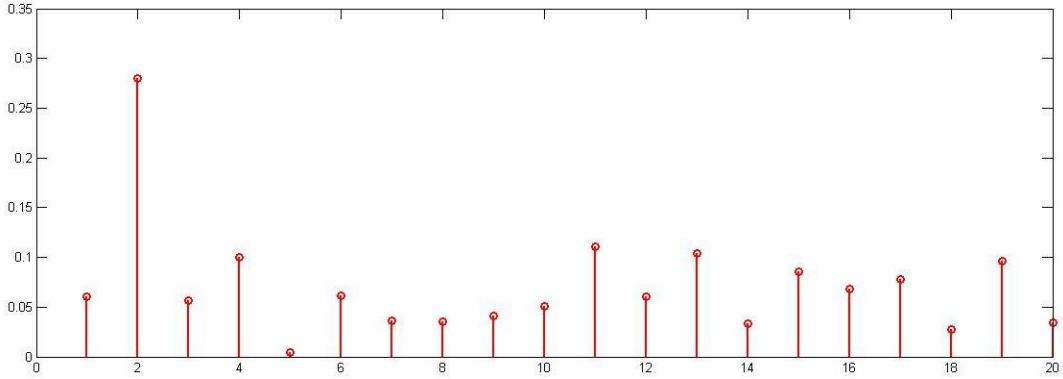
Last the interval of device usage is analyzed. Interval here means the span of two adjacent start time of smart device usage. The overnight intervals are filtered out in this paper. Of all the intervals, the average is 31.3 minutes, the standard deviation is 24.8 minutes, and the coefficient of variation is only 65.8%, which means the biggest and smallest values are both rare. The CDF, CDF in log-log coordinates and the autocorrelation are shown in Figure 2(a), Figure 2(b) and Figure 2(c) separately.


1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
50
100
150
200
250
300

(a)


10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}

(b)


0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
2
4
6
8
10
12
14
16
18
20

(c)

Figure 2. (a) The CDF of intervals. **(b)** The CDF in log-log coordinates. **(c)** Autocorrelation of intervals.

From Figure 2(a), the interval is also long-tailed, but Figure 2(b) shows that interval doesn't obey the power law distribution, the correlation coefficient is only 0.6934. While [13] has reviewed many researches on human behaviors and pointed out that many human behaviors, such as calling, sending short messages and sending emails all obey power law distribution. Here we find an exception.

From Figure 2(c), the autocorrelation is weak between adjacent device usages. So it is hard to predict how long the user will pick up his device again just according to the last few intervals. That is to say, the weak autocorrelation of both duration and interval make predicting user behavior based on adjacent behavior not so reasonable in related works.

2.4. Periodicity of User Behavior

DFT (Discrete Fourier Transform) is often utilized to analyze the periodic behavior. [14] has utilized DFT to understand user behavior. Here DFT is again performed in this paper to pave the way for the prediction of user behavior in Section 2.5.

First the concept of active degree is introduced to quantify how active a user is to use smart device. For every minute, if a user is using smart device, the active degree of this minute is 1, otherwise 0. For every ten minutes, the active degree is the sum of every minute. Ten minutes is the minimum time unit in the following steps. And then DFT is performed and the PSD (power spectral density) is shown in Figure 3.

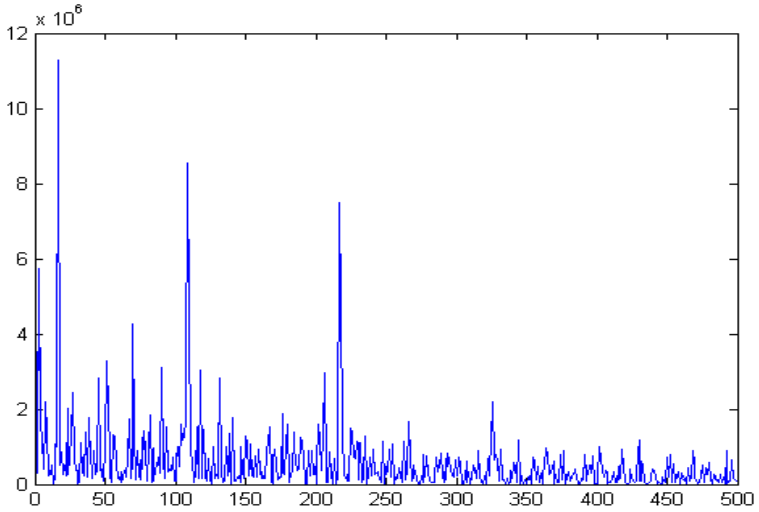


Figure 3. The power spectral density of user behavior.

In Figure 3, the one unit of abscissa is $2\pi/(NT)=6.69\times10^{-7}$Hz, where N=15645 and T=10min=600s. The PSD is long-tailed and here only first 500 values are presented. In Figure 3,

the power spectral density reaches the peak when the abscissa equals 17, whit the corresponding periodicity equaling $NT/(2\pi\times3600\times17)=24.4h$. So it is concluded that the most obvious periodicity of user behavior is 24 hours, which is in agreement with [14].

2.5. Predicting User Behavior

There are not too many researches on predicting user behavior. Of all the existed relevant works, [14] is the most classic one. [14] first analyzed the periodicity of user behavior utilizing DFT, and then utilized Chebyshev inequality to predict the top k applications user is probably to use at a specific time and put these applications on the home screen to make users launch their target applications quickly.

Despite the solid theoretical basis, there is still one thing left to be discussed in [14]. That is, user behavior is periodic and the most obvious periodicity is 24 hours, as shown in both [14] and this paper. But when performing predicting, [14] limit their focus in one day and predict the behavior at specific time x using history behavior at other time. For example, when [14] predicted the behavior at time 15, it used the history behavior at time 9:23 and time 22:08.

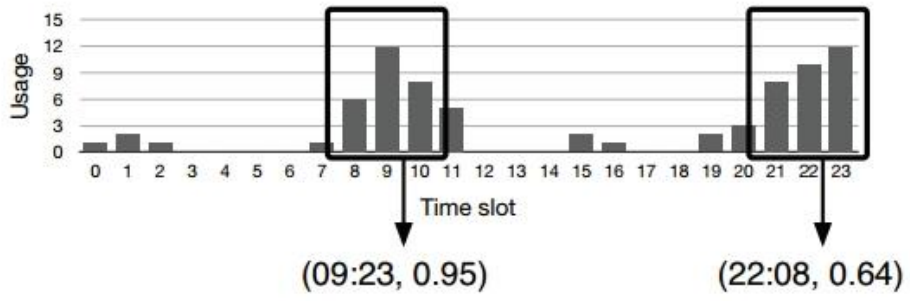


Figure 4. The predicting approach in [14].

There are two unreasonable points in this way. First, the most obvious periodicity is 24 hours, but [14] predicted the behavior using the other time's history behavior. Second, as shown in Figure 1(c) and Figure 2(c), the autocorrelation of duration and interval is neither significant, so to predict user behavior using adjacent behavior is not a good choice.

Here we build a new model to predict user active degree. First let's make clear the problem. The aim is to predict the user behavior at the n^{th} day using behavior data from day 1 to day n-1.

Let $a_{i,x}$ represent the user active degree at time x in day i. The smallest time unit is ten minutes. Smooth $a_{i,x}$ with ten minutes and turn it into $a'_{i,x}$. Calculate the mean and variance of $a'_{i,x}$ ($1 \le i \le n-1$), and notate as E and V separately. Use the notation A to represent the real user active degree which is to predict. According to Chebyshev inequality, expression (1) decides the relationship of value of A and its probability, where P(x) means the probability of event x, ε stands for any positive value.

$$P[\ |A - E\ | \ge \varepsilon\ ] \le D / \varepsilon^2 \quad (1)$$

There are two ways to understand expression (1). In one way, given the acceptable threshold of error probability P_{th}, we can get the minimum ε, notated as ε_{min}, which satisfies the inequality $D / \varepsilon^2 \leq P_{th}$. The ε_{min} is also the biggest deviation between the real active degree A and the mean of history active degree E. In other way, we have the confidence of $(1-P_{th})$ to say that the deviation between A and E is smaller than ε_{min}. In the other way, given the acceptable threshold of predicting deviation ε_{th}, we can get the biggest error probability $P_{max}= D / \varepsilon_{th}^2$. In other way, the probability of deviation between A and E bigger than ε_{th} is smaller than P_{max}.

The calculation of mean and variance of the first n days behavior is shown in expression (2) and expression (3).

$$E_n = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{2}$$

$$D_n = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} - E_n^2 \tag{3}$$

To get the mean and variance of the n^{th} day, data of first (n-1) days is all used, so the data of first (n-1) days should be all kept in storage and thus the space complexity is O(n). Besides, to calculate the variance, there are (n+2) times of multiply operation should be performed, so the time complexity is O(n).

Then iterative formulas (4) and (5) are drawn up to bring down the complexity.

$$E_{n+1} = \frac{x_1 + x_2 + \ldots + x_n + x_{n+1}}{n+1} = \frac{n \cdot E_n + x_{n+1}}{n+1} \tag{4}$$

$$D_{n+1} = \frac{x_1^2 + x_2^2 + \ldots + x_n^2 + x_{n+1}^2}{n+1} - E_{n+1}^2 = \frac{n \cdot (D_n + E_n^2) + x_{n+1}^2}{n+1} - E_{n+1}^2 \tag{5}$$

From (4) and (5), to calculate the mean and variance of $n+1^{th}$ day, only mean and variance of the n^{th} day and user behavior of the $n+1^{th}$ day are needed. That means, the iterative formulas don't require to keep data of the first (n-1) days, and bring down the space complexity from O(n) to O(1). Meanwhile, the calculation of the variance of the $n+1^{th}$ day only perform 5 times of multiply operation, and bring down the time complexity from O(n) to O(1).

3. Evaluation of the Predicting Model

3.1. Evaluation index

Two indexes are defined here to evaluate the predicting model in Section 2.5.

The first one is accurate rate, defined as the probability of the real active rate falls in the expected interval. According to Chebyshev inequality, the accurate rate can't be smaller than $(1-P_{th})$. Although the upper bound can't be lower because there is a distribution to reach the upper bound as pointed out in [15], usually Chebyshev inequality's upper bound is loose. So the accurate rate is still do be examined.

The second one is recall rate, defined as the proportion of the length of predicted active time to the length of real active time. The reason to introduce the index of recall rate is that if there are many active time slot but the model can only predict a few of them, then the prediction is meaningless even if the accurate rate is 100%.

Let the tolerant deviation ε_{th} be 2, and the threshold of the confident probability P_{th} be 0.7. We define the property L as $D / \varepsilon_{th}^{2} \leq 1 - P_{th}$. Of all the time slots which satisfy property L, we predict that the active degree is in the interval $[E-\varepsilon_{th}, E+\varepsilon_{th}]$. If real active degree really falls in interval $[E-\varepsilon_{th}, E+\varepsilon_{th}]$, then we has made an accurate prediction. The proportion that accurately predicted time slots to all the time slots during which the real active rate falls in $[E-\varepsilon_{th}, E+\varepsilon_{th}]$ is calculated as recall rate.

3.2. Experiment Results

Using ten days data as training set, accurate and recall rate are calculated every day then after and then the means of these results are calculated. The data is divided into three types: weekdays, weekends and winter vacation. The results are listed in Table 5.

Table 5. Experiment results in different stages.

stage	accurate rate	recall rate
weekdays	86.3%	63.2%
weekends	72.4%	44.5%
winter vacation	80.5%	51.2%

The average predicting performance is weekdays>winter vacation>weekends. It is concluded that users are most irregular during weekends.

4. Conclusions

Three types of smart device user behavior are analyzed in this paper: application usage, device usage and periodicity of device usage. Through these analyses, the deficiency of related works is pointed out. An improved user behavior predicting model is proposed and experiment results show that the model has good performance in accurate rate and recall rate.

Acknowledgements

This work is supported by China 863 Project: The 5 Generation Wireless Network Architecture and Technologies Development grants 2014AA01A701; Tsinghua University Initiative Scientific Research Program grants 20131089339; National Sci-Tech Major Special Item grants 2012ZX03002015-003; National Natural Science Foundations of China: ”Next

Generation Internet" grants 61161140454 and the EU FP7 under grant number PIRSES-GA-2013-610524.

References

1. Ma Y, Xu B, Bai Y, et al. Daily mood assessment based on mobile phone sensing[C]//Wearable and Implantable Body Sensor Networks (BSN), *2012 Ninth International Conference on. IEEE*, 2012: 142-147.
2. Yang K C C. Exploring factors affecting the adoption of mobile commerce in Singapore[J]. *Telematics and informatics*, 2005, 22(3): 257-277.
3. Pathak, Abhinav, Y. Charlie Hu, and Ming Zhang. "Where is the energy spent inside my app?: fine grained energy accounting on smartphones with eprof." *Proceedings of the 7th ACM european conference on Computer Systems*. ACM, 2012.
4. Yin P, Luo P, Lee W C, et al. App recommendation: a contest between satisfaction and temptation[C], *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013: 395-404.
5. Yan, Bo, and Guanling Chen. "AppJoy: personalized mobile application discovery." *Proceedings of the 9th international conference on Mobile systems, applications, and services*. ACM, 2011.
6. Functionality-Based Clustering using Short Textual Description: Helping Users to Find Apps Installed on their Mobile Device
7. Lu, EH-C., Vincent S. Tseng, and Philip S. Yu. "Mining cluster-based temporal mobile sequential patterns in location-based service environments." *Knowledge and Data Engineering*, IEEE Transactions on 23.6 (2011): 914-927.
8. Shin, Choonsung, Jin-Hyuk Hong, and Anind K. Dey. "Understanding and prediction of mobile application usage for smart phones." *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012.
9. Huang, Ke, et al. "Predicting mobile application usage using contextual information." *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012.
10. Hu, Duan, et al. "We Know What You Are--A User Classification Based on Mobile Data." Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), *IEEE International Conference on and IEEE Cyber*, Physical and Social Computing. IEEE, 2013.
11. http://mobileapp.tsinghua.edu.cn/
12. https://play.google.com/store/apps/details?id=com.attsinghua.main

13. Wang, Bing-Hong, Tao Zhou, and Chang-Song Zhou. "Statistical physics research for human behaviors, complex networks, and information mining." *Journal of University of Shanghai for Science and Technology*, Shanghai Ligong Daxue Xuebao 34.2 (2012).
14. Liao, Zhung-Xun, et al. "Mining temporal profiles of mobile applications for usage prediction." Data Mining Workshops (ICDMW), *2012 IEEE 12th International Conference on. IEEE*, 2012.
15. http://en.wikipedia.org/wiki/Chebyshev%27s_inequality