Human-centric Computing and Information Sciences
a SpringerOpen Journal

RESEARCH Open Access

Modeling and discovering human behavior from smartphone sensing life-log data for identification purpose

Rischan Mafrur*, I. Gde Dharma Nugraha and Deokjai Choi

*Correspondence: rischanlab@gmail.com
School of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea

Abstract

Today, personal data is becoming a new economic asset. Personal data which generated from our smartphone can be used for many purposes such as identification, recommendation system, and etc. The purposes of our research are to discover human behavior based on their smartphone life log data and to build behavior model which can be used for human identification. In this research, we have collected user personal data from 37 students for 2 months which consist of 19 kinds of data sensors. There is still no ideal platform that can collects user personal data continuously and without data loss. The data which collected from user's smartphone have various situations such as the data came from multiple sensors and multiple source information which sometimes one or more data does not available. We have developed a new approach to building human behavior model which can deal with those situations. Furthermore, we evaluate our approach and present the details in this paper.

Keywords: Modeling, Human behavior, Smartphone sensing, User personal data, Behavior mining

Background

Nowadays, smartphones capability have increased significantly. A smartphone has equipped with a high processor, bigger memory, bigger storage and etc. With this equipment, smartphones have the capability to running complex applications. Many sensors also have embedded to the smartphone. With those sensors and log capability of smartphone, we can develop many useful systems or applications in different domains such as healthcare (elderly monitoring system [1, 2], human fall detection [3, 4]), transportation (monitoring road and traffic condition [5]), personal [6, 7] and social behavior [8, 9], environmental monitoring (pollution [10], weather) and etc. To develop such systems, we have to collect user personal data and then analyze it. In this research, we have collected user personal data to identify human behavior. Every person has unique behavior (behavior model). An example case, in the context of daily behavior: Bob is research student in one of a university in Korea. Every working day, he wakes up, takes a shower, breakfast, and goes to his campus at 8:40 AM. He is living in a dormitory, he walks from dormitory to his lab (campus) takes 10 min. Usually, he arrives in his lab at 9 AM and then sits on his chair and starts working. This example is one of the human daily routines in a working day. Based on this story, we can used Bob's smartphone sensors data to define and build Bob's behavior model.

Commonly, researchers who work in this field only focus on one feature to achieved their goals. For examples, authors in [11] focus to use accelerometer sensor for human gait identification, authors in [12] focus to use accelerometer sensor for basic activity recognition, and authors in [13] uses magnetic field sensor for location identification and etc. The result of those researches is very promising. However, one feature that come from one sensor have disadvantages. Sensors on a smartphone have limitation and sometimes become unreliable. So the data that comes from one sensor can be uncertain. This uncertainty condition has an impact the quality of the data since there are many data loss. The result for each approach also come from experiment environment. In the experiment environment, volunteers were asked to do some activity that have been decided by researchers. The case will be different when the approach is implemented to the real user in the real environment. To overcome those problems, we tried to use realistic dataset, it means the data is come from the realistic environment. The are the characteristics of realistic dataset which are defined by us as follows:

- In the realistic environment, the user has different types and brands of a smartphone. Each smartphone has different types of sensors, hardware specification and capabilities.
- We could not expect the human actions and their activities, they will do actions and activities as they want.
- There is no ideal data collection platform that can record user personal data for every day 24 h non-stop, it will drain the battery and spend smartphone resources.
- There is no ideal data collection that can record all the data without any data loss.

Based on those reasons, we propose an approach to modeling human behavior based on user smartphone data log by combining many sensors data rather than only focus on one sensor. When we decide to use many of sensors rather than focus only one sensor, we have to realize that the data from a smartphone are heterogeneous data. In this approach, we tried to develop our system which can deal with those situations (realistic data).

In terms of user personal data collection, there are two ways to collect users personal data based on user involvement. First, participatory sensing and then the second, opportunistic sensing. Participatory sensing means the application still need user's intervention to complete their task. The examples for such applications such as the application that needs the user to taking text input for each time period, taking a picture and etc. On the other hand, opportunistic sensing means application does not need user's intervention to complete their task. Users not involved in making decisions instead a smartphone itself make decisions according to the sensed and stored data. In this research, to collect user personal data, we follow opportunistic method because we do not want to bother user much. Based on those data, we identified human behavior and create their behavior model.

Our contribution in this work are: (1) We have developed an application data collector based on opportunistic method; (2) We have developed system that can identify human behavior based on their smartphone personal data; (3) Also we have developed system which can create human behavior model.

Related works

In this section, we explain about previous work which related with exploring user personality and user smartphone log. Smartphone log consist of many of data such as contact, call log, SMS log, GPS, Wi-Fi, Bluetooth, etc. We can choose which data or information features that we want to explore. For example, from contact data we can explore many things. In [14], they collect the contact list and tried to analyze using several features such as communication intensity, regularity, medium, and temporal tendency. By using machine learning techniques and their proposed method, they achieved up to 90 % accuracy to classify life facets/type of relation in contact (family, work, social). Another interesting research conducted by [15]. They proposed *SmartPhonebook*, an artificial assistant like method which recommends the candidate callees whom the users probably would like to contact in a certain situation. Their approach used social contacts based on the contact patterns that constructed based on user emotional states and behaviors from the mobile log. They use Bayesian networks for handling the uncertainties in the mobile environment. Another example rather than using contact is proposed in [16] that used smartphone log to studies about the business relationship among the users. The proposed method tried to predict the spending behavior for couples in terms of their tendency to explore diverse businesses, become loyal customers, and overspend. The methods tried to predicts customers type such as loyal customers or overspend. Another research is based on location features. In [17], the authors learn about the role of proximity, location, and user personality, such as friendship, to understand user behavior. Their result shows three things which is (1) friendship (SMS contacts and Facebook friendship) in proximity has a significant impact on traffic consumption, (2) personality tends to impact application preference and consumption, and (3) applications can have different contextual usages based on the location. Another research which is focus on location is [18]. In this paper they utilizing location information which obtained from phone sensors (GPS, WiFi, GSM, accelerometer sensors). They proposed a new framework to discover places of interest based on the location where the user usually goes and stays for a while.

Those previous works show that we can exploit call log, SMS log, contact, GPS, and smartphone sensor for many purposes. We still have many of android features that we can explore. In [19], the author tried to investigate how user traits can be inferred by a single snapshot of installed apps. They use SVM with minimal external information such as the religion, relationship status, spoken languages, and countries of interest, and the user is a parent of small children or not. They collected data from over 200 smartphone user, and the list of installed apps, by using their approach, they can achieve over 90 % of precision.

There are also the research which is had the study related with user personality but in different directions. In [20], the authors use virtual world (secondlife.com) to examine how satisfaction in the virtual world was affected by personality differences. They are involving 297 students engage in a virtual tutorial group in Second life and they found that small variations in personality between the virtual and real world groups, such as being helpful, sociable, seeking recognition, or submissive, could lead to greater satisfaction of the discussion.

Not only user personality that can be predicted based on smartphone log data, but also happiness [21], stress [22], mood [23], or maybe we can create application which can help human doing daily routines [24]. In [21], the authors provide the evidence that we can predict the happiness of human based on their phone log. In this paper, the authors proposed a method using Random Forest classifier to recognize daily happiness of person which obtained from the mobile phone usage data (call log, SMS, and Bluetooth proximity data), and background noise. They achieved 80.81 % of accuracy for classifying 3-class daily happiness (happy, neutral, and unhappy). In [22], the authors proposed new approach for daily stress recognition based on human behavior metrics derived from the mobile phone activity (call log, SMS log, and Bluetooth interaction). The approach is based on Random Forest and Gradient Boosted Machine algorithms. Their approach not only on the term of recognition but also for features extraction, selection, and the ensemble recognition model which combines a number of models for each different weather conditions and personality dispositions. They use two classes classification problem (stressed and unstressed) and with theirs approach, they achieved 72.39 % of accuracy. It is could be proof that individual daily stress can be predicted from smartphone data. In [23] have proof that phone log can be used for predicting the user mood. The author in this paper tried to develop smartphone service called *MoodSense*. On this research, 25 iPhone users was studied and only six information features from a mobile log (SMS, email, phone call, application usage, web browsing, and location) was used. By using simple clustering classifier, the proposed method achieved 61 % accuracy on average and improved to 91 % when inference is based on the same participant's data.

There are also previous researches which focus on personality classification but most of them use the Big Five personalities (Extraversion, Agreeableness, Conscientiousness, Emotional Stability and Openness to Experience). In [25], the authors develop a conceptual model that explains a relationship between user Big Five personality and their satisfaction with basic mobile phone services such as call, message, 3G services. The main propose of this paper is several implications for designing of mobile phone services. In [26], the authors said by using smartphone log and their approach, they can predict Big five personality types of users. The result in this paper shows that their approach achieved 42 % better than random and on this research they found that Extraversion and Neuroticism were the traits that were best predicted in their study.

The last one is research which is similar with our work and we only found one. In [27], the authors develop the *mFingerprint* framework, which is user modeling framework which can uniquely depict user. They also use heterogeneous data sensors such as GPS, WiFi, and Bluetooth and soft sensors including app usage logs. The application that they used for collecting data was developed based on *Funf* library. The purpose of this framework is also for user identification. The different between this framework and our proposed system is in the methods/approaches that used. The features that they used are conditional entropy and frequency based footprint features such as conditional features on time and on location. The approach that they used based on designing a discriminative set of statistical features to capture mobile footprints. Based on their method, they achieved 94.68 % accuracy for 4 users, 93.14 % for 10 users and remains 81.30 % for

22 users. In our research, we proposed novel approach to identify a user based on user smartphone log data and find the similarity pattern from their daily activities.

Data collection and processing

Data acquisition

We develop our Data Collector Application for Android Smartphone based on *Funf* library. The *Funf* Open Sensing Framework is an Android-based extensible framework, originally developed at the MIT Media Lab, for doing phone-based mobile sensing. *Funf* provides a reusable set of functionalities enabling the collection and configuration for a broad range of data types. *Funf* is open sourced under the LGPL license. *Funf* framework can collect the data from many of sensor of the smartphone such as location, movement, communication and usage, social proximity, and many more. Details about *Funf* architecture and data format was not described in this paper. More details about *Funf* architecture and data format can be seen in the main site of *Funf*[1] and also *Funf* developer site.[2]

Our application follows opportunistic sensing method. To do that, we have to define the time (interval and duration) first in our application. Interval means how many times in a second, the application will send the data request to the smartphone. An example, we set interval 300 s or equal to 5 min so the application will request and store the data for every 5 min. Duration is the measure of the continuance of any object or event in time. Duration is used in sensor's data because its impact to the size of data that will collected by the application. An example of duration setting, when we set interval 120 s or equal to two minutes and duration 0.07 s means the application will send data request to the smartphone for every 2 min and recording the data during 0.07 s. The details of interval and duration for recording the data from all sensors can be seen on Table 1. Those values are not random values, it means those values are based on our research. The default setting from Funf Library is users can change the interval and duration for collecting the data by themselves. As we explained in previous, we do not want to bother the users, our approach is opportunistic sensing. So, we defined the interval and duration values and tested. We find that those values are optimum one, it means with those values we still can get the valuable information from all of sensors data which we defined and also the data size which generated by this application is not too big. An example is, we used the value of magnetic field sensors in second (5 s) for the duration collection. When we use this setting, the size of data which collected probably more than 1 GB for every day. It will be a problem for the user who uses this application.

Moreover, to make easy to understand, we classify the data that we collected to three of categorization, are:

1. On request data (Current Data).
2. Historical data (Saved in Android database system).
3. Continuous data (Sensors data).

On request data means we ask the current values (information) from an android system such as location, battery, nearby Bluetooth and etc. Historical data means the data

[1] http://www.funf.org/.

[2] https://code.google.com/p/funf-open-sensing-framework/.

Table 1 The application's setting for data collection

No.	Probes	Interval, duration (s)
1.	Location	300
2.	Wi-Fi	300
3.	Bluetooth	300
4.	Battery	300
5.	Call log	86,400
6.	SMS log	86,400
7.	Applications installed	86,400
8.	Hardware info	86,400
9.	Contacts	86,400
10.	Browser search log	86,400
11.	Browser bookmark	86,400
12.	Light sensor	120,0.07
13.	Proximity	120,0.07
14.	Temperature	120,0.07
15.	Magnetic field	120,0.07
16.	Pressure	120,0.07
17.	Activity log	120,0.07
18.	Screen status	120,0.07
19.	Running application	120,0.07

that stored in android database system so we only need to access and copy those data from an android database system to our application. The example of historical data such as contact, call log, SMS log, and etc. Continuous data means we can get those data continuously such as sensor data (accelerometer, gyroscope, magnetic field, and etc.). The duration that we used to collect on request data is 300 s, 1 days (86,400 s) for historical data, and 120 s interval and 0,07 s duration for the continuous data (sensors data).

The list of all sensors data which collected and the explanation for each sensor can be seen on Table 2. From the 19 kinds of sensors data, the total dataset that used is 9 probes/sensors data. The data that we used in this research marked with "X". The total of students who participated in this research are 47 students but not all data are fully available. Some students do not have SMS's log, or other data. The reason they do not have SMS data probably they prefer to use application messenger, such as Kakao, Whatsapp, etc., instead of SMS application. In this research, we use data from 37 students which all the data are available during around less than 2 months. The total size of data from all the students is around 28 GB.

About the number of samples, we sure that 37 students are enough for this research. In this research, we have checked the data from all samples, we sure that dataset that we used are valid and reliable. The procedure that we use for selecting the samples as follows:

1. We do not know details the distribution of samples/subjects, such as the age, sex, weight, height, and another additional information. So, we could not explain the distribution of samples. In this research, our main focus is to discover whether personal data can be used for identification or not.

Table 2 List of data sensors

No.	Name of probes	Explanation	Used
On request data			
1.	SimpleLocationProbe	GPS data (user location)	X
2.	WifiProbe	Nearby Wi-Fi signals	X
3.	BluetoothProbe	Nearby Bluetooth signals	X
4.	BatteryProbe	Battery status	X
Historical data			
1.	CallLogProbe	User call log	X
2.	SmsProbe	User SMS log	X
3.	ApplicationsProbe	List of application installed	
4.	HardwareInfoProbe	User's smartphone hardware info	
5.	BrowserBookmarksProbe	User Bookmarks	
6.	BrowserSearchesProbe	User Browser log	
7.	ContactProbe	User contact (phonebook)	
Continuous data			
1.	LightSensorProbe	Measures the ambient light level (illumination) in lx	
2.	ProximitySensorProbe	Measures the proximity of an object in cm relative to the view screen of a device	
3.	TemperatureSensorProbe	Measures the temperature of the device in degrees Celsius (°C)	
4.	MagneticFieldSensorProbe	Measures the ambient geomagnetic field (x, y, z) in μT	
5.	PressureSensorProbe	Measures the ambient air pressure in hPa or mbar.	
6.	ScreenProbe	Screen phone (on and off)	X
7.	RunningApplicationsProbe	List of running applications	X
8.	ActivityProbe	User activity log based on accelerometer sensor (none, low, and high activity)	X

2. All the subjects are undergraduate students in the same semester at Chonnam National University, Korea.
3. To make sure that all the data which used in this research is reliable, we have checked it. As we explained before, the total students who participated in this data collection are 47 students. We defined many of variables to said that the data is reliable or not such as is all the data from sensors available, is there any errors in their data, and etc., and the final number of sample that we got is 37 students.
4. The duration for data collection is 2 months but not all students follow the rules, some of them do not start to collect their data when they should to start and also some of them stop their data collection not even 2 months. To overcome this problem, we used data in 1 month 20 days, we use same starting and stopping point.
5. Previous research which done by Thang [11] about human gait recognition, they made summary about the number of subjects from many of researches (Table 1), most of them, they used subjects/samples less than 36 subjects, even some that only used 6, 11 subjects. So, we think that 37 students is enough and obviously we have checked that the data that we used is reliable.

Data pre-processing

Funf library has a problem in historical data collection. Historical data is the data which has been stored in an android database system such as contact, SMS log, call log, and etc. We use 86,400 s interval so it means the application copy those data from an android database system to our application database once every day. It makes duplication in our database and we have to care about it. Another problem is system does not always work well. Sometimes something wrong happened and the user's smartphone return value such as NA, error, or/and has no value. We use R programming language to create a module which can remove this duplication and clean the noisy data.

As we mentioned in previous that the size of all of the data is around 28 GB. When we load all of those data in the same time it will spend computer resource especially RAM. It happened because to process data, R environment system load all the data that will be processed in RAM. To handle that problem, we have to define what kind of data that we want to use and store those data to another file (temporary file). In this case, we use CSV file as a temporary file. We have three kind of preprocessing modules and each module will store new data to CSV file. Figure 1 shows the preprocessing process and dataset transformation from preprocessing I until behavior modeling module. Preprocessing I will load all the raw data, removing duplication data, cleansing data, and select the most important data that have been defined. Preprocessing I will store the result data to the CSV I database. Preprocessing II will load the CSV I data. In this preprocessing II, we will apply features extraction. The result of Preprocessing II stored in CSV II. Preprocessing III load the CSV II data and transform the data to the fit format before creating behavior model applied. Those ways will reduce time processing and computer resource's usage.

Human behavior identification

Features are functions of the original measurement variables that are useful for classification or pattern recognition. Feature extraction is the process of defining a set of features, which will most efficiently or meaningfully represent the information that is important for analysis and classification. In this stage, before we extracted the features, we have to define first what the features that we want to use. To extract the features, we have to know first what the human behavior is. In this research, we define that human behavior is human daily activities which carried out continuously. As we mentioned in an introduction section, about the Bob's daily activities from he wakes up until he arrives in his lab room in working day. We call that Bob's activities are Bob's behavior because that activities carried out continuously by Bob in his working day.

In terms of human daily activities, we have to consider about four important things as follows:

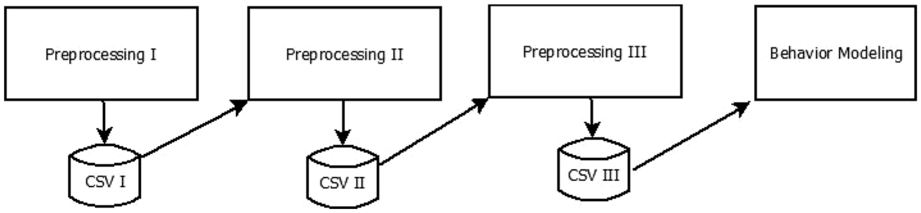


Fig. 1 Data preprocessing flows

- What kind of human activity (e.g., meeting, studying).

Our application follows opportunistic method to collect user personal data, so we do not have activity label in our dataset. We only have activity status (*none*, *low*, and *high*). These status based on accelerometer sensor activity. We use a sum of variance to detect the user activity. If the variance sum more than or equal to 10 float, it will return high activity. If the variance sum value between 3 float and less than 10 float it will be returned low activity and else is none activity. We use this data to define the user activity, even though we do not know the name of activity (activity label). With this activity labels, we still now the user activity pattern (*none*, *low*, and *high*) and can be used to detect user behavior.

- When the activity happened (e.g., 9 AM).

Every value in our dataset has timestamp value. The timestamp value following UNIX timestamp. We have to transform the time to human time. Date and time are used as features in this research.

- Where the location is (e.g., Lab's room).

Rather than living in time domain we also live in place domain (location). In this research, we use three of features to define the human location such as GPS, nearby Wi-Fi, and nearby Bluetooth. GPS is used for defining the user location in outside while nearby Wi-Fi and nearby Bluetooth can be used to define user location inside building.

- Interaction with (user interaction).

We divide user's interaction to two types of interactions. First is an interaction between users and their smartphone, and second is an interaction between users and other users (between human). Interaction between user and their smartphone can be identified by some of sensors such as a battery, screen status, and running applications. Based on battery data, we can know when the user usually charging their batteries. Smartphone screen data can be used as base information about user's smartphone usage. Running applications data stores the list of current applications that used by the user. To know interaction between human and another human, in this research, we use SMS and Call log sensor.

The output of preprocessing II (features extraction) is the data from 9 sensors with some features values that can be seen on Table 3.

Another important thing is we have to realize that machine format is different with the human format in terms of time. A machine can calculates and shows the exactly time such as 00:22:44:34 (millisecond) but a human could not do that. As a human, usually when we want to do an activity in term of time we said on hour and minutes. An example is when we have an agreement with someone, usually we said "OK, we have a meeting at 9.30 AM". We never say: "OK, we have a meeting at 09:30:00:00 (until millisecond)". In this research, we transform machine time format to human time format. We create the module to transform machine time format to human time format in module Pre-processing III.

The main function of preprocessing III is to make the data fit enough before applying the behavior modeling. The details process in the Pre-processing III module as follows:

1. Converting machine time format to human time format. In this research, to convert machine time format to human time format, we round time with the setting: If min-

Table 3 List of sensors dan feature values

No.	Name of probes	Value 1	Value 2	Value 3
1.	ActivityProbe	Status ("none", "low", and "high")		
2.	LocationProbe	Latitude	Longitude	
3.	WifiProbe	List of nearby SSID	MAC	Signal strength
4.	BluetoothProbe	List of nearby Bluetooth devices		
5.	BatteryProbe	Status ("discharging", "full", and "charging")		
6.	ScreenProbe	ON/OFF		
7.	RunningApps	Apps name	Duration	
8.	CallLogProbe	Number	Types	Duration
9.	SmsProbe	Number	Types	Text length

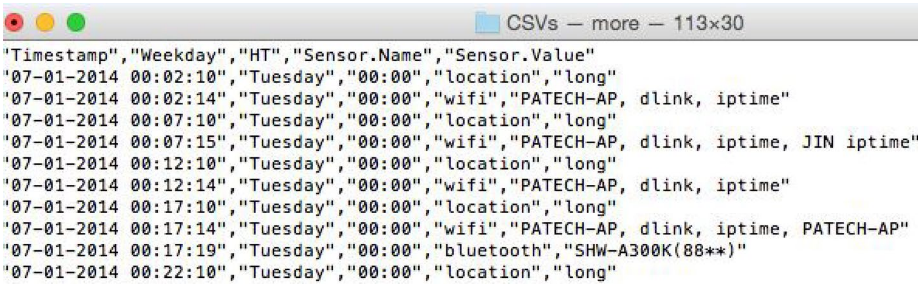
ute less than 30 min will be round down; If minute more than or equal to 30 min will be round up.

2. Changing GPS location value. In this research, we want to find the similarity behavior pattern to build a behavior model. So we change the value of the GPS to "moving status" that value filled by "*same*", "*little*", or "*long*". *Note: 0.0001 degree = 11.1132 m.*

   a. If the previous value of GPS location not change, it means no movement. So the value filled by "*same*".
   b. If the moving distance between 0.0001 and 0.0005, it means little movement. So the value filled by "*little*".
   c. If the moving distance more than 0.0005, it means long movement. So the value filled by "*long*".
   d. We have to decide optimal value that can be used to decided long movement or not. Based on our experiment, 0.0005 value is the optimal one that can distinguish the long movement and little movement in our experiment.

3. Aggregating the values of Wi-Fi and Bluetooth. The data from Wi-Fi and Bluetooth sensors in same time for every value of Wi-Fi stored in one row, and also for the Bluetooth. In this module, if the time is same the sensor values will be aggregated in one row.
4. Aggregating the values of Call Log and SMS log. In this preprocessing, we combine two of values from call log and SMS log into one column. The values of call log and SMS log that used are "type and number". An example of value of call log *"incoming 1bae527e84708183049d8e892a1c959a492ee6a9"*. Even the number was hashed but if the number is same, it has same hash value so we still have pattern information.
5. Removing values such as text length and duration from SMS log and call log, duration from running applications probe, MAC and signal strength from nearby Wi-Fi probe. The reason why we did not use these features because our purpose is to find the similarity.

The example of the final output of preprocessing III can be seen in Fig. 2.



```
"Timestamp","Weekday","HT","Sensor.Name","Sensor.Value"
"07-01-2014 00:02:10","Tuesday","00:00","location","long"
"07-01-2014 00:02:14","Tuesday","00:00","wifi","PATECH-AP, dlink, iptime"
"07-01-2014 00:07:10","Tuesday","00:00","location","long"
"07-01-2014 00:07:15","Tuesday","00:00","wifi","PATECH-AP, dlink, iptime, JIN iptime"
"07-01-2014 00:12:10","Tuesday","00:00","location","long"
"07-01-2014 00:12:14","Tuesday","00:00","wifi","PATECH-AP, dlink, iptime"
"07-01-2014 00:17:10","Tuesday","00:00","location","long"
"07-01-2014 00:17:14","Tuesday","00:00","wifi","PATECH-AP, dlink, iptime, PATECH-AP"
"07-01-2014 00:17:19","Tuesday","00:00","bluetooth","SHW-A300K(88**)"
"07-01-2014 00:22:10","Tuesday","00:00","location","long"
```

Fig. 2 An example output of preprocessing III

Human behaviors modeling

Figure 3 shows the data visualization example in the same day for 4 days from 2 students. Look at the different pattern from both of the users and if we observe the result of the plot for more than 1 weeks we will see the pattern obviously. Based on our observation, we sure that the data features in user personal data log can be used for many purposes such as user identification and classification, recommendation, and etc. In this section, we explain about how we discover human behavior based on their data and building the human behavior model.

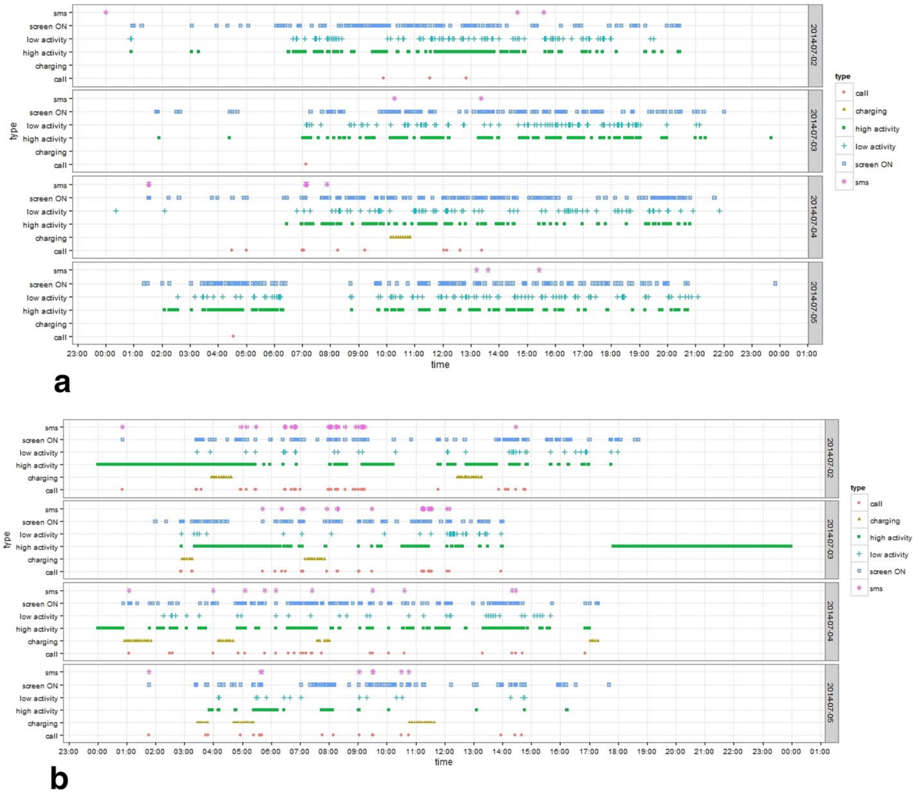
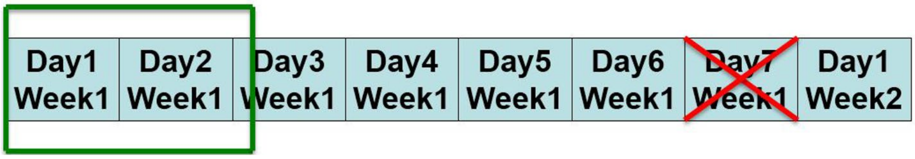


Fig. 3 Example data visualization from two students in the same day for four days. **a** Data visualization from student A. **b** Data visualization from student B

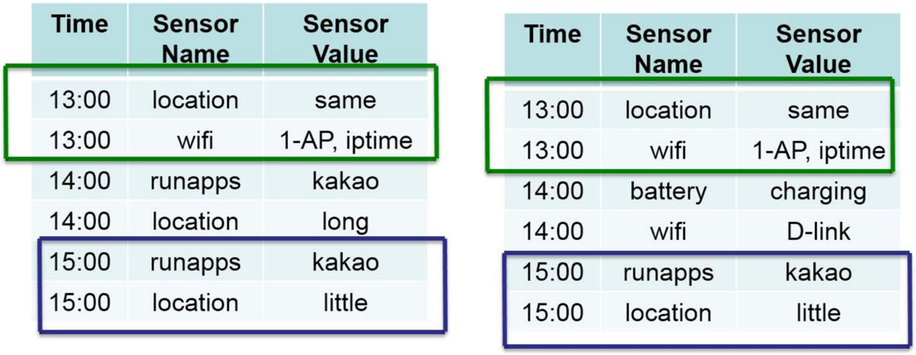
We have dataset around one month and 20 days (7 weeks). We use one month dataset to build user behavior model and then use the remaining data to testing our approach performance. In this approach, we tried to find similar data pattern between days. First, we define the window size. In this research, the window size that we use is two, means two days. We remove the last day of weekday (Sunday) because when the window size is two and the first day start from Monday, so the days in one window is "Monday-Tuesday","Wednesday-Thursday", "Friday-Saturday" the remaining is "Sunday", so we remove it, the illustrated can be seen on Fig. 4. Then we want to discover the same pattern between two days inside the window.

Figure 5 shows the way that we used to find similar data patterns. On that figure, we have two of days in one window. First data is the data of the first day and the second data is the data of the second day and both of data have six rows. We want to find the similar data between first data and second data. Based on an example in that figure, we have two groups of data which similar. The first group in the green rectangle and the second group in the purple rectangle. To know the similarity between data in rows, we use simple strings matching method. The output of the strings matching method is *true* when the string is same/match and *false* when the string is not match. We have used *Levenshtein* distance also to measure the similarity score between two strings in rows to anticipate the data which not match but actually similar. We have mentioned that we applied aggregate function among strings in our dataset. We can imagine, when we use string matching, strings "D-Link AP" and "D-Link AP" is not matched because the



Day1 Week1	Day2 Week1	Day3 Week1	Day4 Week1	Day5 Week1	Day6 Week1	~~Day7 Week1~~	Day1 Week2

Fig. 4 Finding similar pattern in different days and same week (the window size is 2 days)



Time	Sensor Name	Sensor Value
13:00	location	same
13:00	wifi	1-AP, iptime
14:00	runapps	kakao
14:00	location	long
15:00	runapps	kakao
15:00	location	little

Time	Sensor Name	Sensor Value
13:00	location	same
13:00	wifi	1-AP, iptime
14:00	battery	charging
14:00	wifi	D-link
15:00	runapps	kakao
15:00	location	little

Group-1 = 13:00,location,same | 13:00,wifi,1-AP,iptime
Group-1 = 13:00,location,same | 13:00,wifi,1-AP,iptime
Group-2 = 15:00,runapps,kakao | 15:00, location, little
Group-2 = 15:00,runapps,kakao | 15:00, location, little

Fig. 5 Find similar patterns algorithm overview

second string has *"space"* in the end of the word. By using *Levenshtein,* we can handle those problems.

We applied our method in all data with looping function, we collected the all of same data and grouped in groups, one group means the set same user activities. The details of the algorithm can be seen on Algorithm 1. We collect all of intersection data between groups and mark those data as the human behavior model/profile.

Algorithm 1.

```
Data : D, w
Result : All Detected Group in a Window
grpAll, grpTemp, grpPrevious <- NULL
dataValue, dataValueNext <- NULL
while (D in w) for all of D do
     dataValue <- D.current.day
     dataValueNext <- D.next.day
     grpTemp <- findingSimilarPatterns(dataValue,dataValueNext)
     if (grpTemp in grpPrevious)then
      grpNew <- merge(grpPrevious, grpTemp)
      grpAll <- add(grpNew)
     else
      grpAll <- add(grpTemp)
```

Experiment and results

In this section, we explain about our research result and analysis. The goals of our research are to discover human behavior from the user smartphone life log data and based on those behavior data we want to build behavior model which can be used for user identification. This section consists of two of subsections which are behavior identification and performance evaluation.

Behavior identification

Before we explain the result, the details of our experiment as follows:

1. The dataset that we used is around 1 month 20 days, not fully two months. We divide the dataset to two parts.

   a. First month for creating model (first dataset).
   b. Remaining dataset for testing performance (second dataset).

2. Modeling user behavior based on the first dataset (first month dataset). We applied our approach to our first dataset and build human behavior model/profile. We call that profile is B1 data.
3. Extracting and processing the second dataset.

   a. Applying similarity detection to the second dataset with the same setting as that used in building behavior model.
   b. We called the result from this process is B2 data.

4. Is the all of new behavior (B2) identified by behavior model (B1)?.

Table 4 Result of user identification

		TEST					
		ENFP_0719	ENFP_2012	INTJ_5498	ISTJ_3052	ESTJ_5190	ESFP_4634
MODEL	ENFP_0719	**67.922**	0	0.4	2.187	0	1.943
	ENFP_2012	0	**83.582**	0	0	0	0
	INTJ_5498	2.178	0	**75.977**	2.087	0	3.401
	ISTJ_3052	2.289	0	0.4	**93.439**	8.232	1.943
	ESTJ_5190	0	0	0	0.099	**22.866**	0
	ESFP_4634	2.289	0	0.977	2.087	0	**89.686**

a. How many groups of activities (B2) which identified by behavior model (B1)?

b. Calculate the percentage of groups of activities (behavior) which identified.

5. Applying to all students data and observing the result.

Table 4 shows the result of user identification. We applied to all student's data which are 37 students but that table only shows the data from 6 students. In this paper, we do not show all of result because space is not enough to show that. That table is not confusion matrix table, it just looks like confusion matrix table. The value means the percentage of B2 (behavior data from test dataset) which is successfully identified by B1 (behavior model). We can see that our proposed features and our approach can be used for identification. Based on the result and our observation, our approach achieved good enough accuracy even some users has a bad accuracy (under 30 %). The reason why some students have a bad accuracy is because of theirs dataset. For example is the data from *"ESTJ_5190",* the dataset from that user after preprocessing III and splitting to two datasets (model and test), the size of the model dataset is 64 KB. The number of rows less than 500 rows, whereas another data from students who have good accuracy, those data have numbers of rows around more than 50,000 rows. It means the problem is theirs dataset were not enough for creating their behavior model. We also tried to plot the activities from one user who has bad accuracy which can be seen on Fig. 6. The users who have bad accuracy, besides in some days they have few activities, they also have different behavior almost in every day which our approach could not handle it.

Despite some users have a bad accuracy (under 30 %) means only around 30 % behavior data in test dataset which identified in behavior model, but the value is the highest one than other values. We can see from student who has ID "*ESTJ_5190*" only 22.866 % B2 which are identified by B1 (model), but this value is the highest than another values in the horizontal (same row) and vertical (same column), see appendix for full result. It means our approach still can be used for identification.

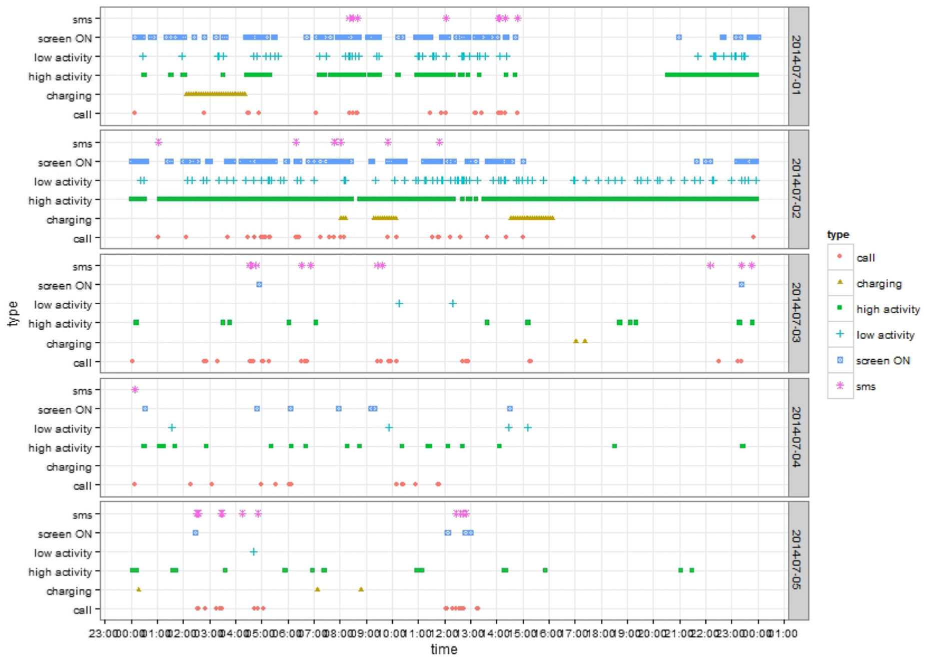


Fig. 6 An example plot of data from student who has bad accuracy

In previous, we have mentioned that we also use *Levenshtein* distance to measure the similarity score between two strings in rows. The reason why we used *Levenshtein* is to anticipate the data which not match but actually similar. Finally, we only use string matching method to find similarity data patterns. We did not use *Levenshtein* distance because whether use it or not, it does not affect the accuracy but only increasing time processing.

The key lesson from this work is that this work is the proof that our personal data can be used for identification system. Even we can say that but we have many limitations in this work. In this work, we used the static window size, it is 2 days. We compare between two days, it will be generated different results when we change the number of days to more than 2 days. The comparison method that we used is a horizontal method, it means we compare between a previous day with a current day. It will have a different result when we compare the days in vertical, it means we try to compare same days but in a different week. In this research, we only use one-time precision, it is one hour. We round the time in 1-hour precision, of course, it will be different when we change the precision to 10, 15, 30 min.

We have challenges to improve this research such as that we mentioned before to change the number of window size, using a vertical method instead of a horizontal method to compare the days, and using different precision time. The other is about a model itself. In this research, we use one month data for building the model and the remaining dataset (20 days) for the testing. It is possible that human can change their behavior, so it will be good if we can update the knowledge inside the model continuously. It is the biggest challenge that we have.

Evaluating performance by removing some features

When we doing research in this field and want to collect personal user data, we cannot said that all the users have same smartphone brand which have same sensors. We have to realize that some sensors probably were not supported by users smartphone or probably user does not have any data in one of sensor such as user does not have SMS and call log. Based on our result, our approach is good enough for user identification. However, we try to answer the question about data quality if we remove some features or sensors data. We want our approach can dealing well with realistic data.

To answer that question, we tried to remove one and more features from our dataset and then we compare the result with the previous result which is using all features. The cases that we tried are:

- Without GPS sensor data.
- Without Wi-Fi sensor data.
- Without Activity data.
- Without Current running applications data.
- Without Battery sensor data.

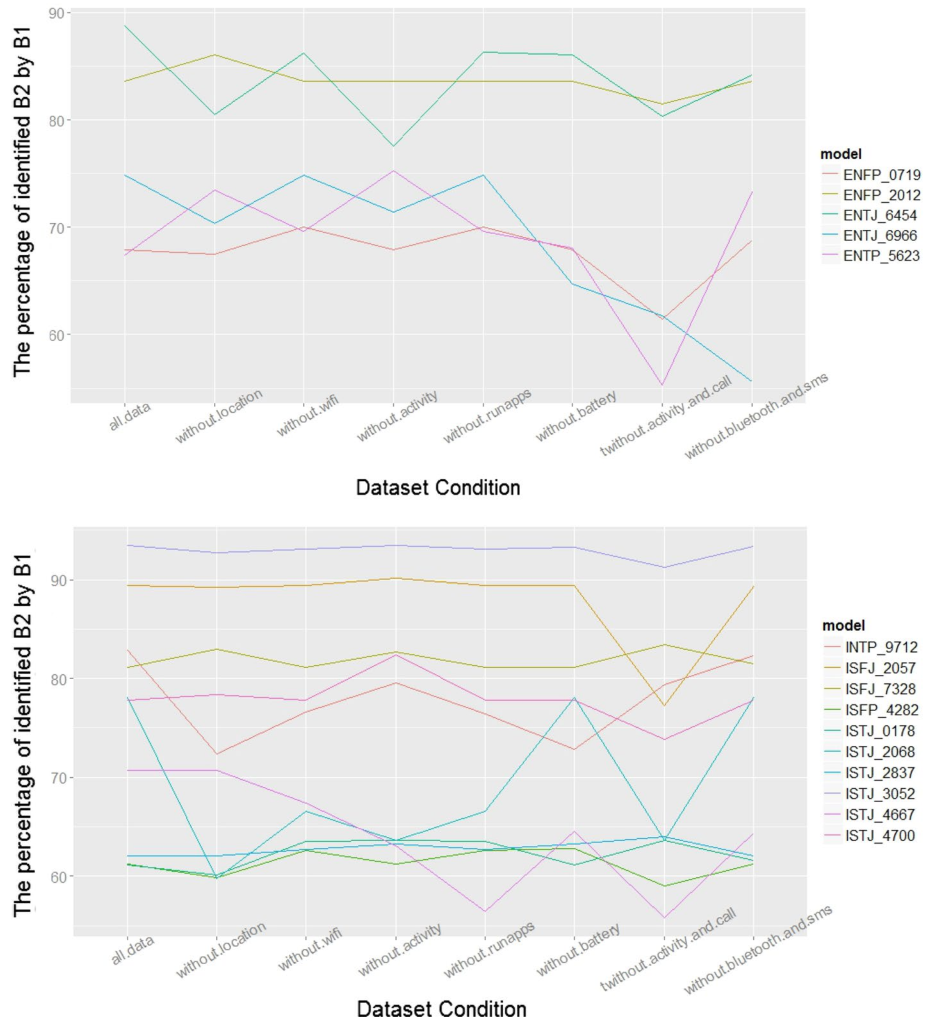


Fig. 7 The percentage of identified B2 by B1 in different dataset condition

- Without Activity data and Call log data.
- Without Bluetooth sensors data and SMS log data.

The result of our cases implementation can be seen on Fig. 7. That figure only shows data from five students, we cannot display all the data due to space constraints. When we see and observe the result, we can conclude that by removing one or two features our approach still well enough for user identification. It means by using our approach, we can handle the realistic data which sometimes the data from one or more sensor does not available.

Conclusion

In this paper, we proposed an approach that can be used for user identification by building human behavior model. We use and combine of many sensors instead only focus on one sensor because we realize that sometimes the users not have data from one or more sensors. Based on our result, we can see that our approach is good enough for user identification. We have tried also to remove one or more features and then observe the accuracy values. The result shows that even one or more features have been removed but our system still can be used for identification. It means our system can handle the problem if one or more data sensors from users smartphone not available. Some of result from our system can achieve up to more than 80 % accuracy but we have four students who have less than 30 % accuracy. In this paper, we have explained also why four students have bad accuracy. The reasons are students who have bad accuracy, their datasets are too small and they have different behavior for almost each day which our approach does not capable to handle it. Despite some of the accuracy values are under 30 % but those values still can be used for identification because those values are the highest one compared to others. It means that our approach still good enough for identification system.

Authors' contributions

RM designed and performed experiments, analysed data and wrote the paper; IGDN reviewed and fixed grammar and English errors; DC supervised the project. All authors read and approved the final manuscript.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2012R1A1A2007014).

Competing interests

The authors declare that they have no competing interests.

Received: 15 June 2015 Accepted: 4 October 2015
Published online: 12 October 2015

References

1. Faetti T, Paradiso R (2013) A novel wearable system for elderly monitoring. Adv Sci Technol 85:17–22
2. Pierleoni P, Pernini L, Belli A, Palma L (2014) An android-based heart monitoring system for the elderly and for patients with heart disease. Int J Telemed Appl 2014:11
3. Tong L, Song Q, Ge Y, Liu M. HMM-based human fall detection and prediction method using tri-axial accelerometer. IEEE Sens J. vol. 13, no. 5
4. Aziza O, Parkc EJ, Morid G, Robinovitch SN (2014) Distinguishing the causes of falls in humans using an array of wearable tri-axial accelerometers. Gait Posture 39:506–512
5. Zhou P, Zheng Y, Li M. How long to wait?: predicting bus arrival time with mobile phone based participatory sensing. In: MobiSys '12 Proceedings of the 10th international conference on Mobile systems, applications, and services
6. Bogomolov A, Lepri B, Pianesi F (2013) Happiness recognition from mobile phone data. In: BioMedCom 2013
7. LiKamWa R, Liu Y, Lane ND, Zhong L (2011) Can your smartphone infer your mood?. In: PhoneSense workshop

8. Chittaranjan G, Blom J, Gatica-Perez D (2013) Mining large-scale smartphone data for personality studies. Personal Ubiquitous Comput 17(3):433–450
9. Singh VK, Freeman L, Lepri B, Pentland A (2013) Predicting spending behavior using socio-mobile features. In: BioMedCom 2013
10. Maisonneuve N, Stevens M, Niessen ME, Steels L (2009) NoiseTube: Measuring and mapping noise pollution with mobile phones. In: Information technologies in environmental engineering
11. Hoang T, Nguyen T, Luong C, Do S, Deokjai C (2013) Adaptive cross-device gait recognition using a mobile accelerometer. J Inf Process Syst 9(2):333
12. Ayu M, Mantoro T, Fariadi A, Basamh S (2011) Recognizing user activity based on accelerometer data from a mobile phone. In: 2011 IEEE symposium on computers & informatics (ISCI), Kuala Lumpur
13. Galvan-Tejada C, Carrasco-Jimenez J, Branea R (2013) Location identification using a magnetic-field-based FFT signature. In: The 4th international conference on ambient systems, networks and technologies (ANT 2013)
14. Min JK, Wiese J, Hong JI, Zimmerman J (2013) Mining smartphone data to classify life-facets of social relationships. CSCW '13 Proceedings of the 2013 conference on Computer supported cooperative work, pp. 285–294
15. Min JK, Cho SB (2011) Mobile human network management and recommendation by probabilistic social mining. IEEE Trans Syst Man Cybern—Part B: Cybern 41(3):761–771
16. Singh VK, Freeman L, Lepri B, Pentland A (2013) Predicting spending behavior using socio-mobile features. BioMedCom 2013
17. Meng L, Liu S, Striegel A (2014). Analyzing the impact of proximity, location, and personality on smartphone usage. 2014 IEEE INFOCOM workshop on dynamic social networks
18. Montoliu R, Blom J, Gatica-Perez D (2013) Discovering places of interest in everyday life from smartphone data. J Multimed Tools Appl 62(1):179–207
19. Seneviratne S, Seneviratne A, Mohapatra P, Mahanti A (2014) Predicting user traits from a snapshot of apps installed on a smartphone. ACM SIGMOBILE Mob Comput Commun Rev 18(2):1–8
20. Sutanto J, Phang CW, Tan CH, Lu X (2011) Dr. Jekyll vis-a`- vis Mr. Hyde: personality variation between virtual and real worlds. J Inf Manag 19–26
21. Bogomolov A, Lepri B, Pianesi F (2013) Happiness recognition from mobile phone data. BioMedCom 2013
22. Bogomolov A, Lepri B, Ferron M, Pianesi F, Pentland A (2014) Pervasive stress recognition for sustainable living. The Third IEEE international workshop on social implications of pervasive computing
23. LiKamWa R, Liu Y, Lane N, Zhong L (2011) Can your smartphone infer your mood? PhoneSense workshop
24. Antila V, Polet J, Lämsä A, Liikka J (2012) RoutineMaker: towards end-user automation of daily routines using smartphones. PerCom 2012. Lugano
25. De oliveira R, Cherubini M, Oliver N (2013) Influence of personality on satisfaction with mobile phone services. ACM transactions on computer-human interaction, Vol. 20, No. 2, Article 10
26. de Montjoye YA, Quoidbach J, Robic F, Pentland A (2013) Predicting people personality using novel mobile phone-based metrics. Soc Comput Behav-Cult Model Predict (2013)
27. Zhang H, Yan Z, Yang J, Munguia Tapia E, Crandall D (2014) mFingerprint: privacy-preserving user modeling with multimodal mobile device footprints. Soc Comput Behav-Cult Model Predict Lecture Notes Comput Sci 8393:195–203