User Attribute Classification Method Based on Trajectory Patterns with Active Scanning Devices

Kenji Takayanagi[1(✉)], Kazuya Murao[1], Masahiro Mochizuki[2], and Nobuhiko Nishio[1]

[1] College of Information Science and Engineering, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan
gibson@ubi.cs.ritsumei.ac.jp, {murao,nishio}@cs.ritsumei.ac.jp
[2] Research Organization of Science and Technology, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan
moma@ubi.cs.ritsumei.ac.jp

Abstract. Technologies for grasping the distribution and flow of people are required for urban planning, traffic planning, evacuation, rescue activities in case of disaster, and marketing. In order to grasp what kind of attribute the distribution and flow of people are formed, this paper proposes a method that estimates the attributes of users. As a method of estimating user attributes, we utilize probe request frame of Wi-Fi that smartphones are emitting. Probe request frame includes MAC address, enabling us to acquire the movement trajectory of a user by tracking the MAC address. By using the feature values obtained from the movement trajectory of the user, users are roughly classified into several types. In this paper, we focus on the user attribute estimation in underground city comprising of stations, shops, restaurants and so on. Through the practical experiment at Osaka underground city, we confirmed that the proposed method can classify the users into commuter or not by using the intervals between probe request frames.

Keywords: People flow analysis · Attribute estimation
Spatiotemporal data · Probe request frame

1 Introduction

There is a demand for technology to grasp the distribution and flow of people for urban planning, traffic planning, evacuation, rescue activities in case of disaster, and marketing. As a related service, there is a store congestion status notification service provided by Google[1]. This is based on the anonymous data aggregated from the users who enabled the location history of Google to determine crowded time and staying time. When deploying such services, information on time and

[1] https://blog.google/products/search/know-you-go-google/.

K. Murao et al. (Eds.): MobiCASE 2018, LNICST 240, pp. 282–288, 2018.
https://doi.org/10.1007/978-3-319-90740-6_24

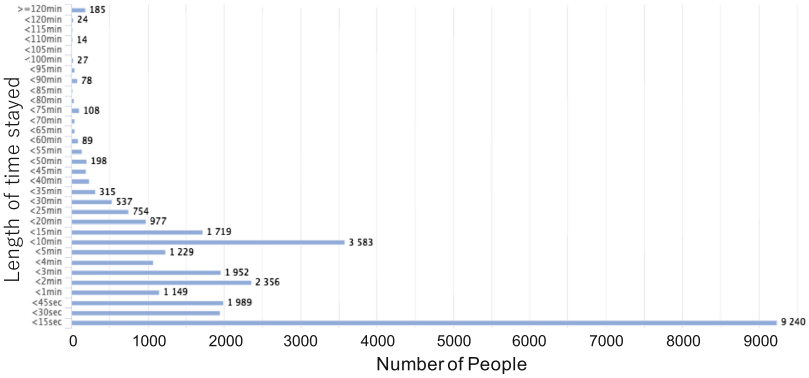

Fig. 1. Distribution of the length of time users stayed near a lunchbox store in the station.

place is required for fine granularity even in an indoor space such as a commercial facility. We focused on devices with Wi-Fi and have proposed a system that can inexpensively grasp the distribution and flow of people [1] thus far. With this system, it is possible to grasp the flow of people and the number people who are not moving. However, there is a problem that it is not easy to grasp what kind of attribute influences the distribution of the number of stayers. Actually, we investigated the distribution of time the users stayed from probe request frame obtained from sensors installed near a lunchbox store in a station as shown in Fig. 1. Figure 1 shows that the peaks appear at 15 s, 2 min, and 10 min. It can be inferred that the peak at 15 s probably passed through the lunch box store. On the other hand, the peak at 2 min and the peak at 10 min probably are those who dropped in near the lunch box store comprising of regular customers buying the boxed lunch quickly and first-visit customers taking long time to decide what to buy. In other words, our interest is shifting to grasp what attributes of people affect the distribution of the number of stayers. Therefore, in this paper, our goal is to find actions such as walkthroughs, commuters, shoppers, and facilities stakeholders.

2 Method

This section explains how to collect probe request frames, how to deal with MAC-randomized device, how to generate a moving path and how to cluster moving paths in order to find user attributes such as walkthrough, commuters, shoppers, and facility workers.

2.1 How to Collect Probe Request Frames

The overall system configuration diagram is shown in Fig. 2. A device such as a smartphone accompanied by a Wi-Fi function periodically transmits a probe

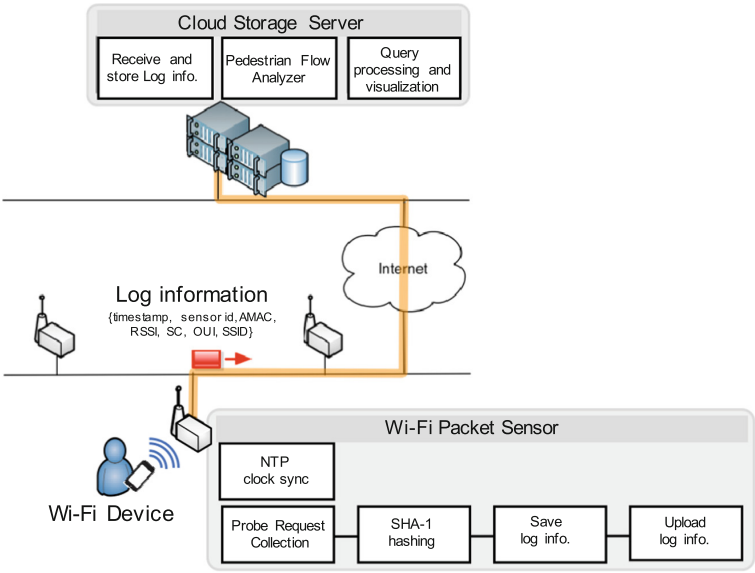

Fig. 2. System overview.

request frame to connect to a Wi-Fi access point. Distributively arrange sensors that can collect probe request frames in the environment where the station and the commercial facility are close. The MAC address of the collected information is anonymized by SHA-1 algorithm for the consideration of privacy. This anonymized MAC address is called an AMAC address.

Our method uploads these pieces of information to the cloud storage server and utilize those accumulated information to estimate attributes.

2.2 How to Deal with MAC-Randomized Device

This section describes the influence of the MAC-randomized device. MAC-randomized device indicates a device that has a dummy MAC address randomly changed. The reason why such a dummy MAC address is transmitted from the device is to protect the privacy.

A device with a dummy MAC address can be specified by checking the second least significant bit of the first octet of the first 24-bit product vendor code which is called OUI (Organizationally Unique Identifier) as shown in Fig. 3 [2].

The ratio of the probe request frame including randomized MAC address and the ratio of the MAC-randomized device among the probe request frames observed from February 18, 2017 to March 18, 2017 are shown in Fig. 4. From the Fig. 4, the average randomization rate (blue line) of all probe request frames observed on a day is about 13 % on average. It is considered that this is because the number of probe request frames transmitted from MAC-non-randomized devices is frequent. The average randomization rate (orange line) of all devices

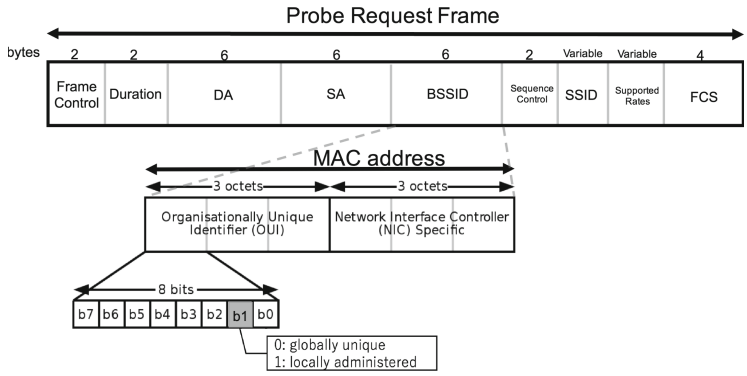

Fig. 3. Randomize bit of MAC address.

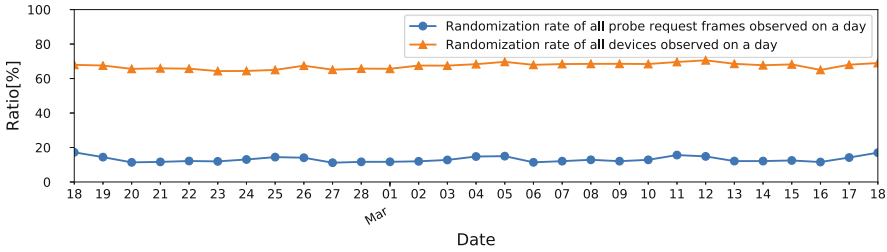

Fig. 4. Ratios of probe request frames and devices with randomized MAC address. (Color figure online)

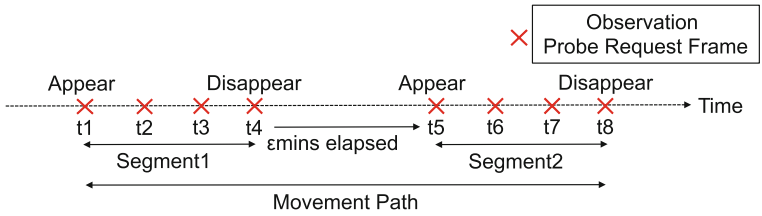

Fig. 5. Example of movement path generation.

observed on a day is about 67% on average. More specifically, it is difficult to identify two different frames emitted from a same device, therefore devices with randomized MAC have to be eliminated in the user attribute estimation.

2.3 Movement Path Generation

When there are observed probe request frames arranged in chronological order as shown in Fig. 5, probe request frames whose intervals are within ϵ seconds are grouped as a segment, and one or more segment groups are called movement paths. This movement path is generated on a daily basis. In addition, a probe

request frame observed after ϵ-second interval is regarded as appearance, and the last probe request frame in the segment is regarded as disappearance. It can be said that the segment composed of appearance and disappearance is within the observation area.

Then, a method of deciding the threshold ϵ when creating a segment, but since the probe request frame transmission interval differs for each model of the device, a histogram of probe request frame transmission intervals is investigated for each vender that can be known from the MAC address. The third quartile of the histogram created by each vendor is set to ϵ. The reason why we set the third quartile to ϵ is that the segment breaks more than necessary and effective feature values cannot be obtained if the mode is set to ϵ.

2.4 Movement Path Clustering

The proposed method extracts feature values from each segment. Four kinds of feature values are employed: average path length, average speed, variance of speed, and time stayed. Using these features, we estimate user attributes: walkthroughs, shoppers, and facility workers each segment. In addition, focusing on segments, there are intervals in which user device is not found. Definition of these features are as follows:

- Average Path Length: Calculating the length at each sensor hop in the segment.

$$\bar{l} = \frac{\sum_{k=1}^{n} l_k}{n}$$

- Average Speed: Calculating the speed at each sensor hop in the segment.

$$\bar{v} = \frac{\sum_{k=1}^{n} v_k}{n}$$

- Variance of Speed: Calculating the speed at each sensor hop in the segment.

$$\sigma^2 = \frac{\sum_{k=1}^{n} (v_k - \bar{v})^2}{n}$$

- Time Stayed: Calculating the time difference between the appearance time and the disappearance time in the segment.
- Idle Time: Calculating idle time difference between the disappearance time of a segment and the appearance time of the next segment.

In this section, we describe a method for estimating the attributes of commuters by using the feature value of idle time between segments. This feature value is obtained from a movement path composed of two or more segments. As a result of investigating the number of segments per movement path on a weekday, 77% of a movement path consist of one path and 23% of movement paths consist of two or more paths. Figure 6 shows the clustering result of idle time using K-means++ clustering method (K = 2) between segments. Two clusters colored in blue and orange shows the first peak and second peak, respectively. Figure 7

shows a histogram of devices that appeared first time on a day at each hour. Colors in the bars show two clusters obtained in Fig. 6. From the Fig. 7, high peak appears at 8 a.m., which is the morning commuting time zone. Another peak appears at 6 p.m., which shows workers going back home. On the other hand, there can be seen people who spent 2 to 3 h around this area evenly in the evening. From these results, we have confirmed that commuter attribute.

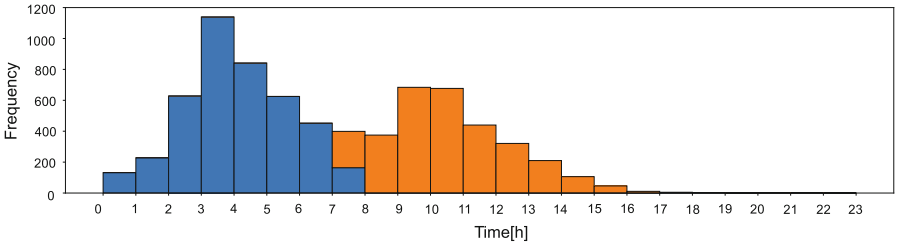

Fig. 6. Clustering result of idle time between segments (K = 2). (Color figure online)

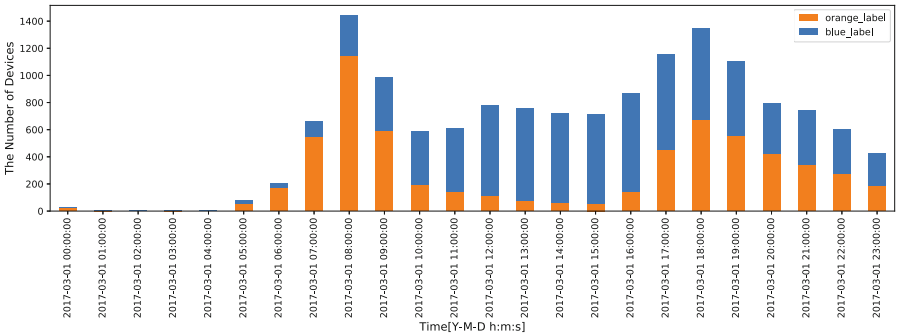

Fig. 7. Time distribution of idle time between segments of the day. (Color figure online)

3 Conclusion

In this paper, we estimated the user attribute by using the feature amount obtained from the movement trajectory generated by using probe request frame. In fact, we estimated the commuter attribute from the idle time between segments, which is the feature values of clustering, using the data on the first day of the weekday. As a result, a characteristic mountain was seen when commuting in the morning and returning home in the evening.

References

1. Fukuzaki, Y., Mochizuki, M., Murao, K., Nishio, N.: Statistical analysis of actual number of pedestrians for Wi-Fi packet-based pedestrian flow sensing. In: Proceedings of the 1st International Workshop on Smart Cities, pp. 1519–1526 (2015)
2. Robyns, P., Bonné, B., Quax, P., Lamotte, W.: Non-cooperative 802.11 MAC layer fingerprinting and tracking of mobile devices. Secur. Commun. Netw. **2017**, 1–26 (2017)