Behavior Analysis and Individual Labeling Using Data from Wi-Fi IPS

Yuming Lin
School of Architecture, Tsinghua University

Weixin Huang
School of Architecture, Tsinghua University

ABSTRACT
It is fairly important for architects and urban designers to understand how different people interact with the environment. However, traditional investigation methods for studying environmental behavior are quite limited in their coverage of samples and regions, which are not sufficient to delve into the behavioral differences of people. Only recently, the development of indoor positioning systems (IPS) and data-mining techniques has made it possible to collect full-time, full-coverage data for behavioral difference research and individualized identification.

In our research, the Wi-Fi IPS system is chosen among the various IPS systems as the data source due to its extensive applicability and acceptable cost. In this paper, we analyzed a 60-day anonymized dataset from a ski resort, collected by a Wi-Fi IPS system with 110 Wi-Fi access points. Combining this with mobile phone data and questionnaires, we revealed some interesting characteristics of tourists from different origins through spatial-temporal behavioral data, and further conducted individual labeling through supervised learning.

Through this case study, temporal-spatial behavioral data from an IPS system exhibited great potential in revealing individual characteristics besides exploring group differences, shedding light on the prospect of architectural space personalization.
INTRODUCTION

Understanding the complex interaction between varieties of environmental factors and different types of people is one of the main focuses in environmental behavior study, but an effective method for collecting and analyzing related data remains to be explored.

In the past few decades, on-site observations, questionnaires, interviews, cognitive maps and a few other methods have been the main techniques for engaging in the study of environmental behavior. However, these traditional methods made it hard to access comprehensive and accurate information about human behaviors. Apart from requiring enormous work of the investigators, on-site observations often cause subjective bias and observer effect; questionnaires or interviews might be low-cost and individual, but not ideal in covering detailed information quantitatively. Nowadays, indoor positioning systems (IPS) and data-mining techniques have made it possible to collect a batch of full-time, full-coverage data, providing promising new possibilities for environmental behavior research, especially for comparative research on differences in humans’ spatial-temporal behaviors in built environments.

Under the limitations of traditional investigation methods, the data analysis process for environmental behavior study often appears as qualitative interpretation or regression models based on small samples, making it harder for the case studies to reach comprehensive and generalized conclusions. However, considering the growing availability of data, there might be another solution through the machine-learning algorithm. With the help of data mining and supervised learning, behavioral analysis on individuals can be conducted on a large scale, finally achieving statistical conclusions. The research methods for individuals are not only promising in the commercial field, but also in architectural space personalization.

Based on the huge amount of spatial and temporal behavioral data collected by IPS technology, the main focus of this paper is to validate the reliability of this data source, to conduct a comparative study of environmental behavior differences among different groups of people, and further to label individuals by their spatial-temporal behavioral characteristics. Data mining and machine-learning techniques were employed in the process of data cleaning, data compression, data analysis and feature prediction. The main data processing was based on the Python programming language, while the gradient boosting decision tree (GBDT) algorithm (Friedman 2001) was used for supervised learning, due to its high efficiency and excellent performance.

PREVIOUS RESEARCH

There are many technologies that can be used in IPS, but most of them have their own deficiencies. GPS tracking technology cannot be used inside buildings because of the accuracy limitation (Nirjon et al. 2014). Radio frequency identification or ultra wideband require the subjects to wear specific devices, leading to the sample quantity limitation and observer effects (Ni et al. 2004, Gezici et al. 2005). Bluetooth IPS works well, but the usage of Bluetooth is relatively limited, resulting in possible sampling bias (Feldmann et al. 2003, Rida et al. 2015). In our research, the Wi-Fi positioning technology is chosen due to its balance between applicability and accuracy (Cypriani et al. 2009). Wi-Fi is the most widely used wireless network access technology. According to the IEEE 802.11 protocol, the Wi-Fi monitor is able to record activities of the mobile device unobtrusively as long as its Wi-Fi is on (Liu et al. 2007). Therefore, we could collect an amount of long-term, full-coverage information without disturbing the observer, which is essential for big data analysis.

On the other hand, human dynamics study has made a fruitful contribution on revealing the statistical characteristics and dynamic mechanism of various human behaviors. Barabási (2005) showed that human trajectories showed high spatial-temporal regularity using mobile phone data. Also based on mobile phone data, Song et al. (2010) indicated 93% potential predictability in human mobility, showing the scientifically grounded possibility of a predictive model. Furthermore, Sekara, Stopczynski, and Lehmann (2016) revealed a dynamic social network structure in a campus through a mobile Bluetooth application that can search for people nearby (Sekara, Stopczynski, and Lehmann 2016). At a city scale, Sapiezynski et al. (2015) succeeded in tracking people using Wi-Fi signals (Sapiezynski et al. 2015). These studies provide reliable references and valuable guidance for our analysis of spatial-temporal behaviors.

METHODOLOGY

The Wi-Fi IPS technology is widely used in public, commercial or urban space (Zeng, Pathak, and Mohapatra 2015). The system works as follows: according to the IEEE 802.11 protocol, a Wi-Fi-enabled device will broadcast a probe-request signal to the surrounding access point (AP) for a short timed interval. Monitoring APs will unobtrusively record the signal information, consisting of time, MAC address of device and received signal strength. The received signal strength varies with the distance between device and AP, obeying the propagation loss formula (Hata 1980). Hence, given the location of APs (Zhu and Feng 2013), we may locate the device through the trilateral positioning algorithm. Based on the unique MAC address, we are able to track one’s trajectory continuously. As for the complex
The basic system of the Wi-Fi indoor position system.

Aerial view of Vanke Songhua Lake Resort.

The exponential distribution of present location counts.

The temporal distribution of Wi-Fi positioning data.

The cyclical fluctuation in human flow.

After the necessary data cleaning and processing, the Wi-Fi positioning data is ready to provide abundant and valuable information on human trajectories for environmental behavioral study. The analytical method for the information will be further introduced in the case study of Vanke Songhua Lake Resort.

CASE INTRODUCTION

Vanke Songhua Lake Resort is located in Jilin City, northeast of China. There is a 250 m long commercial pedestrian street in the center of resort, providing a family hotel, catering, ski rental and ticketing from north to south. The mountain and ski slopes are located further south, and are open from 8:00 to 17:00 during the day. The study revolves around the tourists’ spatial-temporal trajectories on the pedestrian street.

Data used in this paper is mainly collected through the Wi-Fi IPS. There are 110 monitor APs covering the whole commercial street, collecting 200 million anonymous records in 60 days (from January to March, 2015). The dataset contains 1,607,171 unique locations from 466,065 different MAC addresses after compression. Table 1 shows some examples of raw data.

<table>
<thead>
<tr>
<th>Time</th>
<th>MAC address of device</th>
<th>X coordinate</th>
<th>Y coordinate</th>
<th>MAC address of AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-01-31 12:29:48</td>
<td>D4970B****8E</td>
<td>2,220,000</td>
<td>2,123,000</td>
<td>70BAEFAF32D0</td>
</tr>
<tr>
<td>2015-01-31 12:29:48</td>
<td>D8878F******EA</td>
<td>2,221,000</td>
<td>2,135,000</td>
<td>70BAEFAF4480</td>
</tr>
<tr>
<td>2015-01-31 12:29:48</td>
<td>E6191D******F1</td>
<td>2,072,000</td>
<td>2,837,000</td>
<td>70BAEFAF44A0</td>
</tr>
</tbody>
</table>

In addition, we obtained a set of anonymous data as the baseline reference, including the number of ticket sales, the number of ski rentals and ropeway card records. Besides, the most important data came from people who anonymously submitted their phone numbers for Wi-Fi certification. Table 2 shows some examples of Wi-Fi certification data.

<table>
<thead>
<tr>
<th>Phone number</th>
<th>Log in</th>
<th>Log out</th>
<th>Duration</th>
<th>IP address</th>
<th>MAC address</th>
</tr>
</thead>
<tbody>
<tr>
<td>13331****61</td>
<td>2014-12-29 09:54:57</td>
<td>2014-12-29 12:45:58</td>
<td>222</td>
<td>172.100.24.86</td>
<td>889FFA****37</td>
</tr>
<tr>
<td>13331****61</td>
<td>2014-12-30 17:27:31</td>
<td>2015-01-01 00:04:40</td>
<td>1842</td>
<td>172.100.24.86</td>
<td>889FFA****37</td>
</tr>
</tbody>
</table>

In addition, on-site observations and questionnaires are fairly important supplements, providing a solid foundation for explanations of results.

SPATIAL-TEMPORAL DISTRIBUTION

Figure 1 is the general heat map of the accumulated spatial distribution of people in the resort. It can be observed that people are mainly distributed along the pedestrian street, with right part being a hotel beyond the coverage area. The darker...
Beijing, Shanghai and Guangzhou. An OLS regression shows that the number of tourists from different provinces is proportional (0.0005, significant at 0.01 level) to the per capita GDP and inversely proportional (-0.0008, significant at 0.05 level) to the travel distance. Since local Jilin tourists accounted for nearly half, and tourists outside Jilin showed no significant difference in behavior from one another, the crowd is divided into two categories, "local tourists" for tourists from Jilin Province, and "non-local tourists" for tourists outside Jilin.

Temporal Difference
There are many interesting differences between local and non-local tourists. It is shown in Figure 8 that the number of people from different groups follows the same tendency but different volatility. It can be calculated that the correlation coefficient between the two is high (0.6816), but the standard error differs 2.48 times (38.8/96.6).

The distribution of days present for different groups also differ (Figure 9). For local tourists, the majority of people only visit for 1 day, showing the characteristics of a power-law distribution again. The power-law exponent is in the interval of (1,2) which is consistent with many empirical studies before. For the non-local tourists, there seems to be a lognormal distribution with the mode at 2. Since the resort is only 30 kilometers away from Jilin City, local people may prefer a short trip, while the tourists from afar are more willing to stay for more days. This kind of time preference can be further analyzed intuitively.

CLASSIFICATION STUDY
In addition to the simple description, we can also conduct a comparative study of tourists from different origins based on mobile phone data. It can be a reasonable presumption that short-distance travelers and long-distance travelers are different in behavior, and the origin information in the phone number can verify this.

The phone data contains about 43,000 records, nearly 6,000 of which are valid mobile phone numbers. From Figure 7, it can be indicated that about a half of people are from local Jilin City, while a quarter is from the 3 most developed cities in China:
last time, so that all people can be plotted in the two-dimensional density map. (Fig. 10). As can be seen from Figure 10, the behavioral patterns of local and non-local tourists are very different. The local tourists tend to come to the resort around 9:00 and return home before 18:00; non-local tourists will still appear in the resort until about 21:00. Those who “arrive” at midnight should have spent the night in the resort, which are more common for the non-local tourists.

**Spatial Difference**

In addition to the temporal behavioral difference, the spatial behaviors diverge. Figure 11 shows that local tourists appear more often in parking lots, fast-food restaurants and the ski rental shop. In contrast, non-local tourists appear more often at the bus stop, full-service restaurants and resort hotels, showing a discrepancy in transportation and consumption habits. For non-local tourists, staying and eating in a resort may be a more convenient option, while local tourists may prefer skiing itself.

Concerning the number of locations visited by everyone (Figure 12), this is subject to exponential distribution or lognormal distribution, suggesting a connection between spatial behaviors and temporal behaviors.

The APs in the resort cover the major shops and public spaces in the pedestrian street, thus we can analyze the traffic in different areas and find different behavioral patterns among various types of public spaces or shops. Figure 13 shows the standardized number of populations in several typical areas. Local tourists seem more active during the day, appearing frequently at the ticket office or ski rental shop. As for the Chinese restaurant, the proportion of non-local tourists during dinner is higher than that during lunch, and the same is true for the hotel’s breakfast. These discrepancies can be attributed to the difference in the choice of accommodation. There is an abnormal flow peak in ski rentals at night, which remains to be further explored.

**INDIVIDUAL LABELING**

Although classification-based research has provided us with useful revelations, the number of labeled samples is limited and may have sample errors. Therefore, it is necessary to generalize the conclusions based on a small number of labeled samples, which is individual labeling. From the perspective of environmental behavior research, the supervised learning of the individual label essentially establishes the link between personal behavior and personal attributes.

Nevertheless, this type of link is embedded in high-dimensional data, requiring efficient learning algorithms to mine it out. This paper uses the gradient boosting decision tree (GBDT)
algorithm as the core supervised learning algorithm for several reasons. First, it can be flexible to deal with various types of data, including continuous and discrete values; second, it can achieve relatively high accuracy in a short time; and third, the robustness of to the abnormal value is very strong. These are fairly important when we are exploring a new source of data.

According to behavioral characteristics and expert experience, each individual trajectory record is converted into 39 independent features, which can mainly be divided into two aspects. Temporal information includes the length of stay, the number of present days, arrival time, departure time, etc., whereas spatial information includes the number of locations, location range, points of interest, moving distance and so on. These features are standardized for individual labels, which are binary values of origins (local or non-local).

In order to verify the effectiveness of supervised learning, the 6,000 labeled samples were randomly divided into an 80% training set and a 20% test set, and logistic regression was used as the baseline model. The null model refers to random predictions according to the proportion of each label. The results are shown in Table 3. It can be found that the GBDT algorithm can use the trajectory data to predict the individual labels at about 83% accuracy, which is significantly better than the baseline model.

In order to test the idea further, we altered the label to 3 categories (from Jilin City; from Beijing, Shanghai or Guangzhou; from other cities). The training results are shown in Table 4. With the increase in categories, the accuracy of the classification declined, but GBDT still maintained the best performance at about 63% accuracy. The results suggest that it is feasible to carry out individual labeling using trajectory features. There might be some over-fitting problems due to the lack of samples or invalid features.

Table 3: The result of supervised learning on origin (two-category).

<table>
<thead>
<tr>
<th>Evaluating Indicator</th>
<th>Baseline Model</th>
<th>Logistic Regression</th>
<th>GBDT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training set</td>
<td>Test set</td>
<td>Training set</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.5083</td>
<td>0.7921</td>
<td>0.7868</td>
</tr>
<tr>
<td>AUC Score</td>
<td>0.5000</td>
<td>0.8612</td>
<td>0.8507</td>
</tr>
</tbody>
</table>

Table 4: The result of supervised learning on origin (three-category).

<table>
<thead>
<tr>
<th>Evaluating Indicator</th>
<th>Baseline Model</th>
<th>Logistic Regression</th>
<th>GBDT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training set</td>
<td>Test set</td>
<td>Training set</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.3889</td>
<td>0.5685</td>
<td>0.5606</td>
</tr>
<tr>
<td>AUC Score</td>
<td>0.5000</td>
<td>0.7515</td>
<td>0.7458</td>
</tr>
</tbody>
</table>

Supervised learning provides us with a new means of understanding this type of link. Figure 14 shows the importance of different features in the prediction process, suggesting which
aspects of behavior have a more significant difference in labeling. In fact, the most important features are about the arrival-departure time, the distance traveled in the resort and whether to eat at the restaurant, which has been confirmed in our classification study before.

Basing on the labeled samples, the trained classifier can now label all captured data, which is shown in Figure 15. With brief analysis it can be found that the previous results basically remain intact. There are still some subtle changes, such as the blurring of some features. In general, it suggests that the mobile data samples are not biased and behavioral features are effective. Except for the verification of previous work, it demonstrates the feasibility of obtaining statistical conclusions from large-scale behavioral analysis conducted on individuals, which requires further study in the future.

DISCUSSION

Architectural space and individuals inside it together compose a complex system, especially for large public facilities. Revealing people’s behavioral patterns in such complex systems has great value for design, management and daily life. Contemporary IPS technologies and big data analysis methods have greatly extended our ability to understand people’s behaviors.

Compared to traditional environmental behavior investigation methods, the IPS system is capable of fully covering the investigated area, continuously working over a long period, and collecting data from a large number of people. Its advantage lies in that, first, since the data has full spatial and temporal extension, it can re-visualize the spatial-temporal trajectory of individuals, and reveal the spatial function network and its evolution; second, since the data contains trajectory information of a huge number of people, it has the potential to reveal different patterns and variations of behavior; third, if we consider the multitude of information in different dimensions such as time, space and different groups of people, there could be more in-depth analysis regarding people’s behavior patterns.

In this paper, based on the 60-day Wi-Fi IPS data of a ski resort, we first tried to depict the general spatial and temporal pattern of people’s flow, and then explored more details on the behaviors of different groups of people, including temporal, spatial and compounded distribution. Furthermore, a supervised learning algorithm was used in generalizing conclusions, and its results revealed the importance of different features and uncovered some behavioral characteristics. These analyses demonstrated that Wi-Fi IPS data contained abundant information of people’s behavior, and could become a good source for environmental behavioral research.

Machine-learning algorithms have proven their effectiveness in descending dimensions and extracting underlying behavioral patterns, also have shown a great potential in predicting individual attributes. In this case study, mobile data provides us with the basis for forecasting individual origins, and other data sources could be taken into account in the future. Besides origins, other personal attributes could be predicted in the same way when effective data sources are provided. This kind of ability to combine data from different sources is quite critical in the era of big data.

It should be pointed out that although Wi-Fi IPS data contains abundant information from people, it could still be biased. The dataset may be contaminated by noise from other devices in the environment, and there may be systematic deviation in the process of using mobile devices to represent people, including the fact that some people may not turn on their Wi-Fi, may not have a smart phone, may carry multiple devices and other possible situations. Furthermore, IPS data does not contain information of behavior type, and we can only infer it by the location. It is even harder to know what people feel and think, which may be inferred by other data sources such as social media data. In many situations, on-site observation and interviews are still very important.

In addition to environmental behavioral research, large-scale positioning data provides unlimited possibilities for generative...
architectural design. Furthermore, with the help of machine learning, architects may learn about the characteristics of each individual, and ultimately form statistical conclusions rather than vice versa. In this way, the architectural design can be carried out based on the individual’s characteristics rather than statistical conclusions. Thus personalized architectural designs like interactive buildings or intelligent homes may be able to subvert the basic logic of architecture.

ACKNOWLEDGEMENTS
This research project is supported by the National Natural Science Foundation of China (NSFC) (Project No. 51578299).

REFERENCES


Yuming Lin is a Master Candidate in the School of Architecture, Tsinghua University. He holds a Bachelor of Architecture degree from Tsinghua University. His research focuses on the application of big data technology in behavior analysis.

Weixin Huang is an Associate Professor in the School of Architecture, Tsinghua University. He is the Associate Director of Digital Architectural Technology Education Committee of China, committee member of Computer Aided Architectural Design Research in Asia (CAADRIA), and one of the founders of Digital Architectural Design Association (DADA) of the Architectural Society of China (ASC). He received his Ph.D. from Kyoto University, Japan. His research focuses on digital design of architectural & structural integrated system, big data spatio-temporal behavior analysis, and design cognition.