Prediction of the crowd behavior in campus based on time series model

Liangwei Chen¹, Jingjing Yang¹,²,³, Zhigang Chen and Ming Huang¹,²

¹School of Information Science and Engineering, Wireless Innovation Lab, Yunnan University, Kunming 650091, China
²Radio Monitoring Center of Yunnan Province, Kunming 650228, China
³Email: yangjingjing@ynu.edu.cn

Abstract. Recognizing and understanding the behavior of the campus crowd is of great significance for improving the comprehensive management capabilities of the campus. The increasingly abundant of mobile terminal WIFI signals provide a unique data source for studying the behavior of the campus crowd. A Wifi-based personnel activity monitoring system is developed in this paper. Through monitoring and analyzing the MAC address (Media Access Control Address), RSSI (Received Signal Strength Indication) and Wireless AP (Access Point) in and around the Science Museum, Huaizhou building of Yunnan University, the association between WIFI signal and the crowd behavior is studied. Time series analysis is used to make predictions, and some valuable information about the gathering and activity status of campus crowd behavior is obtained. It provides an experimental platform for studying the dynamic behavior of people in campus or other crowd areas.

1. Introduction
The mobile Internet has changed people's lifestyle. At present, people can't live without mobile terminals. Since the mobile terminal position and movement trajectory data can reveal the regularity of the individual's movement and the similarity between the movement patterns [1-2], and can identify the surrounding environment and human behavior characteristics by measuring WIFI channel state information [3-6], how to perceive human behavior patterns through radio waves has become a current research hotspot [7]. However, CSI is the channel state information (Channel State Information) for measuring channel conditions. It belongs to the PHY layer and comes from the decoded subcarriers in the OFDM system. CSI is fine-grained physical information and is more sensitive to the environment, so it can be used in motion recognition, gesture recognition, key recognition, tracking and other fields. The purpose of this article is to monitor group events on campus, rather than group events, not individual behavior. Therefore, in this experiment, we monitor the RSS (Received Signal Strength) of mobile devices. Unlike video-based mobile crowd monitoring methods, WIFI signal-based monitoring systems do not involve personal privacy, and are not affected by factors such as heavy rain, fog, and poor light at night. Many scholars at home and abroad have studied this. In 2016, Fuxjaeger P et al. [8] conducted a road traffic analysis study based on WIFI signals. The experimental results show that by deploying WIFI nodes on the highway, traffic can be monitored; Traunmueller M W et al. [9] prove that WIFI signals have great potential for understanding the behavior and movement patterns of people in cities; Zhu L J et al. [10] experiments have shown that monitoring WIFI signals can estimate passenger’s traffic with an accuracy better than 90%; Ding X B et al. [11] pass Image recognition and
WIFI monitoring have monitored the passenger flow in urban rail transit stations, and the results show that the monitoring system can identify the status of passenger flow in time and provide intelligent early warning based on the recognition results. Xu J Y et al. [12] designed and implemented a WiFi user behavior analysis system based on the k-means algorithm, built a multi-dimensional analysis mathematical model of WIFI user behavior, and clustered and analyzed user behavior habits and activity, work and rest habits, and traffic distributed.

In summary, there are many reports about identifying the surrounding environment and human behavior characteristics based on the measurement of WIFI channel state information. Based on this, the study of the behavior of mobile people is combined with the information of the mobile terminal's Mac address, RSSI, and wireless AP because there are large Social and economic benefits have become the current hotspot, however, research work on data prediction combined with time series analysis has not been reported. This paper establishes a WIFI-based campus mobile crowd behavior research experimental system, and digs out some valuable WIFI signals around buildings and their vicinity. The correlation information between the movement rules of mobile crowds provides an experimental platform for studying the dynamic behavior of mobile crowds in campuses or other sensitive areas, and provides new technical means and solutions for management departments.

2. System design and prediction algorithm

2.1. System design

The experimental system structure is shown in Figure 1. The data acquisition module in the figure includes a WIFI probe and an embedded PC. The WIFI probe acquires the MAC address, RSSI, and wireless AP information of the WIFI user by capturing the Probe frame. WIMAX technology connects the data source to the server; the data storage uses the non-relational database MongoDB; the data processing and calculation uses the python scientific computing module; the data prediction uses XGBoost (eXtreme Gradient Boosting), Arima (Autoregressive Integrated Moving Average model) and LSTM (Long Short-Term Memory); the data display module is developed using the web framework Django.

![Figure 1. Overall system block diagram.](image)

The data collection node has the edge computing function. The data is stored in MongoDB, and real-time processing, prediction and web publishing are performed simultaneously. When the number of real-time users exceeds the prediction threshold, the data collection node status is abnormal. The server regularly collects abnormal information from the data collection nodes, and performs data mining and analysis to achieve the purpose of monitoring the gathering and activity status of personnel through WIFI signals.
2.2. Prediction algorithm

There are five main types of time series predictions: (1) traditional time series modeling methods, linear models such as ARMA and ARIMA. The ARMA model is suitable for the analysis of stationary time series, and the ARIMA model can be used for the analysis of homogeneous non-stationary time series. (2) Use xgboost / LSTM model / temporal convolutional network to start with feature engineering, and use time sliding window to change the data organization method to make predictions. (3) The time series is transformed into a supervised learning dataset, using XGboost / LSTM model / time convolutional network. (4) In 2017, Facebook open-sourced a time series algorithm, Fbprophet, which can process outliers and missing values in the data and automatically predict future trends. (5) Deep learning network, using CNN + RNN + Attention to perform time series prediction. In this paper we use typical ARIMA models, XGboost models, and deep learning models LSTM to predict time series analysis.

2.2.1. ARIMA model. In the autoregressive quadrature moving average model Arima (p, d, q), AR is "autoregressive" and MA is "moving average"; p is the number of autoregressive terms, q is the number of moving average terms, and d is the number of differences made by stationary sequences. The ARIMA (p, d, q) model can be expressed as [13]:

\[ (1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^{q} \theta_i L^i) \varepsilon_t \]  

(2.1)

In the above formula, L is a lag operator, \( d \in \mathbb{Z}, d > 0 \).

Among them, AR describes the relationship between the current value and the historical value, and uses the historical time data of the variable itself to predict itself. Formula definition for the p-order autoregressive process:

\[ y_t = u + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t \]

In the above formula, \( y_t \) is the current value, u is the constant term, p is the order, \( \phi_i \) is the autocorrelation coefficient, \( \varepsilon_t \) is the error.

MA focuses on the accumulation of error terms in the autoregressive model, it can effectively eliminate random fluctuations in prediction. The formula definition of the q-order autoregressive process:

\[ y_t = u + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \]

ARMA is a combination of autoregressive model and moving average model:

\[ y_t = u + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t \]

In the ARIMA model, I is the difference model, and ARIMA is the ARMA model after the difference, which ensures the stability of the data. The ARIMA (5,1,0) model is used in this experiment.

2.2.2. LSTM. The long-term and short-term memory model LSTM consists of an input gate \( i_t \), a forget gate \( f_t \), an output gate \( o_t \), a memory cell \( c_t \), and a hidden state \( h_t \): the forget gate determines what information is removed from the cell state; the input gate memory information at the current time; the output gate controls whether the information at the current time is output to the hidden state \( h_t \). Finally, combining the current time with past information, the LSTM unit has the ability to save, read, reset, and update long-distance historical information [14]. The model formula is as follows:

\[ f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \]  

(2.2)

\[ i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \]  

(2.3)

\[ o_t = \text{tanh}(W_o \times [h_{t-1}, x_t] + b_o) \]  

(2.4)

\[ u_t = \sigma(W_o \times [h_{t-1}, x_t] + b_u) \]  

(2.5)
among them:

\[ h_t = O_t \times \tanh(C_t) \]  \hspace{1cm} (2.6)

\[ C_t = f_t \times C_{t-1} + i_t \times u_t \]  \hspace{1cm} (2.7)

In the above formula, \( x_t \) represents the input time series, \( W \) represents the connection weight, and \( \sigma \) (sigmoid) and \( \tanh \) are activation functions. In this experiment, the training model of LSTM has 5 layers, for the first 3 layers, each layer has 200 neurons, then there is a dropout layer, the dropout rate is 0.3, and finally a fully connected layer outputs prediction data. The training process has a total of 250 epochs. Using the Adam optimizer, the initial learning rate is \( lr = 0.001 \), and the learning rate is reduced by 20% for every 100 epochs.

2.2.3. **XGBoost.** XGBoost is a machine learning model developed by Chen Tianqi and others. The model won the Kaggle competition. XGBoost is an integrated model. Its core is to continuously perform feature splitting (greedy algorithm that enumerates all different tree structures) to add a decision tree, and use the new function \( f(x) \) learned to fit the last predicted residual Poor, the superposition process is as follows [15]:

\[ \hat{y}_t^0 = 0 \]  \hspace{1cm} (2.8)

\[ \hat{y}_t^1 = f_t(x_t) = \hat{y}_t^0 + f_t(x_t) \]  \hspace{1cm} (2.9)

\[ \hat{y}_t^2 = f_t(x_t) + f_2(x_t) = \hat{y}_t^1 + f_2(x_t) \]  \hspace{1cm} (2.10)

\[ \hat{y}_t^l = f_t(x_t) + f_2(x_t) + \ldots + f_l(x_t) = \hat{y}_t^{l-1} + f_l(x_t) \]  \hspace{1cm} (2.11)

In the formula, \( \hat{y}_t^l \) is the prediction of the \( t \)-th round model of the \( i \)-th tree, \( \hat{y}_t^{l-1} \) is the prediction of the previous \( t-1 \) round of the model, and \( f_t(x_t) \) is the newly learned function, and then The results of \( k \) trees are summed to get the predicted value of the sample. The objective function is as follows:

\[ \text{obj}^{(t)} = \sum_{i=1}^{n} l \left( y_i, \left( \hat{y}_t^{(t-1)} + f_t(x_i) \right) \right) + \Omega(f_t) + \text{constant} \]  \hspace{1cm} (2.12)

Among them, \( l \left( y_i, \left( \hat{y}_t^{(t-1)} + f_t(x_i) \right) \right) \) is a loss function, and \( \Omega(f_t) \) is regularization. This experiment uses python's machine learning library XGBoost. The initial parameters are selected as subsample = 0.8, colsample_bytree = 0.85, eta = 0.1, max_depth = 10, seed = 42. Then use the “cv” function of XGBoost to determine the learning rate under Number of optimal trees: max_depth = 7, best number of iterations: n_estimators = 171. During the training process, select the error function rmse, early_stop = 50.

3. Results and discussion

3.1. Evaluation standards

The evaluation standards include \( R^2 \) (R-squared), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). The calculation formulas are:

\[ R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \]  \hspace{1cm} (3.1)

\[ RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \]  \hspace{1cm} (3.2)

\[ MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \]  \hspace{1cm} (3.3)

In the above formula, \( y_i \) is the real value, \( \hat{y}_i \) is the predicted value, and \( \bar{y}_i \) is the average value. The value of \( R^2 \) is between 0 and 1. A larger value indicates that the predicted value is closer to the true value; a smaller RMSE and MAE indicates that the predicted value is closer to the true value.

3.2. Experimental results

3.2.1. **Laboratory devices.** We deployed two monitoring nodes and a back-end server on the East Road Campus of Yunnan University. Node1 and the server are installed in the Huai Zhou Building as shown in Figure 2, while Node2 is installed in the Science Museum, as shown in Figure 3.
3.2.2. **Science Museum Node.** Figure 4 shows the raw data and the prediction results of different models of the WIFI traffic flow of the Dongluyuan Science Museum of Yunnan University from December 17, 2019 to December 25, 2019. It can be seen that the flow of people in the vicinity of the Science Museum changes periodically with the date. The daily change pattern is related to the activities of nearby personnel. The change pattern is basically the same from Monday to Thursday, and the other three days have different characteristics. All three prediction algorithms are able to predict the overall situation of the flow of people.
Figure 4. The changes and predictions of the number of people around Science Museum.

Figure 5 (a) shows the flow of people at the Science Museum of Yunnan University on Thursday. It can be seen that there are 11 characteristic peaks at 4:30-24:00. Experiments have shown that these characteristic peaks and the departure time of the school bus, the Science Museum meeting and The opening hours of the opposite library are closely related. 7:00, 9:00, 12:30, 14:30, 17:40 are the departure times from the Cuihu Campus to Chenggong Campus, 8:30, 10:30, 13:00, 16:00, 17:30, 18:00 and 21:00 are the time from the Chenggong Campus to the school bus of the Donglu Campus. Experiments have confirmed that characteristic peaks will appear near these points, and the characteristic peaks at 12:30 and 17:40 are the highest, indicating that more people The choice of school buses for these two periods coincides with the activities of the campus crowd. Comparing Figure 5 (b), Figure 5 (c), and Figure 5 (d), we can see that the overall number of riders decreased on the weekend, and the number of people who arrived at the Donglu Campus at 9pm on the evening was minimized on Friday. There is a meeting at the Science Museum at 4 pm, and there will be a sudden increase in traffic at about 4 pm. The minimum flow of people in a day is after 22:20, which is the closing time of the library. Figure 5 (b) shows the flow of people on Friday. It can be seen that there are 11 characteristic peaks from 4:30 to 24:00, which proves that the day and Thursday have the same change pattern, but the number of people in each time period is greater than that of the week and the period of time when the minimum flow of people at the lowest point in advance. Figure 5 (c) shows the change in the flow of people on Saturday. It can be seen that the overall flow of people is about half of that on Thursday. There are 17 characteristic peaks at 4:30-24:00, subjected to off-campus activities. Random effects of personnel entering campus. Figure 5 (d) shows the change in the flow of people on Sunday. It can be seen that there are also 11 characteristic peaks at 4:30-24:00, which indicates that the activities of the personnel have returned to normal on Sunday. The number of people increased, and arrangements were made for the next week’s teaching plan; the number of people returning to the Donglu campus at 9 pm also returned to normal.
Figure 5. Daily number of people around Science Museum. (a) Thursday traffic variation; (b) Friday traffic variation; (c) Saturday traffic variation; (d) Sunday traffic variation.

3.2.3. Huaizhou Lou Node. Figure 6 shows the raw data and the prediction results of different models of the traffic flow in Huaizhou Building, Yunnan University Donglu Campus from December 21, 2019 to December 29, 2019.

Figure 6. The changes and predictions of the number of people around Huai Zhou Lou.
Figure 7 (a) shows the traffic flow of HuaiZhou lou on Thursday. It can be seen that there are 11 characteristic peaks at 5:30-24:00. Experiments have shown that these characteristic peaks are closely related to the teacher's work time and the student's class time. At 8:00, the flow of people began to increase. Teachers and students entered the Huaizhou Tower, and the flow of people did not change until 11:30. After 11:30, the flow of people began to decrease, and a characteristic peak appeared until 2pm. Students start classes at 14:00, but the number will fall and rise in a short period of time. This is because some students leave the classroom during the class and return soon. After class at 4:30 in the afternoon, the traffic flow began to decrease again. There will be evening self-study from 19:00 to 21:00, and the flow of people will increase, but during this time, the flow of people will fluctuate greatly, resulting in multiple characteristic peaks, which is closely related to the quality of evening self-study. After 23:00, the traffic flow was reduced to a minimum, and there was basically no one in Huaizhou Tower. Figure 7 (b) shows the flow of people on Huaizhou Tower on Friday. It can be seen that there are 27 characteristic peaks at 5:30-24:00, which indicates that the personnel activity is high and the activity rules are complicated on Friday. From 7 to 11 o'clock, the flow of people increased in a wavy manner, and there was no rapid increase on Thursday. At 11:30, the reduction in staff still showed a wavy decline; the punctuality rate of students who attended class at 2:30 pm Decreased on Friday; after 4:30 a.m., the traffic flow showed a wave-like decrease. Figure 7 (c) and Figure 7 (d) are the changes in the flow of people on Huai Zhou Lou on Saturday and Sunday. It can be seen that there are 24 and 27 characteristic peaks at 5:30-24:00, respectively. Half of the four, the change pattern is the same as that on Friday, people enter and exit frequently, and the flow of people changes randomly.

Figure 7. Daily number of people around Huai Zhou Lou. (a) Thursday traffic variation; (b) Friday traffic variation; (c) Saturday traffic variation; (d) Sunday traffic variation.
3.2.4. Analysis of prediction results. In this experiment, the WIFI probe collects data every ten minutes. We take one month of data for experiments, of which 22 days of data is used as the training set, and 8 days of data is used as the test set. Based on the ARIMA model, LSTM model, and XGBoost model, the prediction results of the Science Museum and Huai Zhou Tower are shown in Table 1 and Table 2. From the tables, it can be seen that the three models all predict the change pattern of human flow very well. The best data prediction models for science museum data prediction are LSTM model and ARIMA model. In comparison, the XGBoost model has poor prediction results in two places.

| Table 1. Comparison of model performance based on the data of science museum. |
|-----------------------------|-----------------------------|-----------------------------|
|                             | (MAE)          | (R squared)     | (RMSE)         |
| ARIMA                       | 15.7158        | 0.9188          | 21.6609        |
| LSTM                        | 13.1661        | 0.9256          | 20.7340        |
| XGBoost                     | 14.5162        | 0.7188          | 25.2974        |

| Table 2. Comparison of model performance based on the data of Huai Zhou Lou. |
|-----------------------------|-----------------------------|-----------------------------|
|                             | (MAE)          | (R squared)     | (RMSE)         |
| ARIMA                       | 3.4674         | 0.9085          | 5.5312         |
| LSTM                        | 3.9115         | 0.8857          | 6.1834         |
| XGBoost                     | 12.0461        | 0.1295          | 17.0641        |

The models I chose are the three most basic prediction models, and there is no good or bad between the models. The only difference is the data. In this experiment, the data set is not too large, and the XGBoost model requires a large amount of data to get a good prediction effect, so the prediction effect of XGBoost is the worst in this experiment. Once the LSTM neural network model is well matched, a good prediction result can be obtained. The ARIMA model is a traditional time series prediction model. By adding the current value to the data set each time to predict the next time value, the prediction result is also very good. In future experiments, we will increase the data set to obtain more accurate experimental results.

In order to confirm the function of the system, we did an experiment in Huai Zhou Tower at 5pm on January 10, 2020. First, the model predicts the number of people × (1 ± 10%) to get the threshold for the flow of people, with an upper limit of 29 people and a lower limit of 24 people. Secondly, six students of the organization laboratory are turned on the WIFI function of the mobile phone to test the prediction model. At this time, the system monitored 31 real-time people and issued an alarm prompt, which verified the reliability of the prediction model.

4. Conclusions

This experiment proves that by collecting Mac address, RSSI, wireless AP and other information of the building and its nearby mobile terminals, and performing data mining and time series analysis on this basis, it can reveal the dynamic behavior of mobile people in campus or other sensitive areas. Time series analysis shows that the LSTM prediction algorithm has the best prediction effect on the 24-hour peak of human flow, the Arima model takes the second place, and the XGBoost model performs worse than the previous two algorithms. In order to confirm the function of the system, we turned on the WIFI function of 6 mobile phones at the same time in Huai Zhou Tower, and successfully alerted the abnormal situation of human flow, indicating that the system has application value in public safety and campus management.

References

[1] Marta C, Cesar A and Barabasi A 2008 Understanding individual human mobility patterns *Nature* **453** 779-782
[2] Xu Z X, Wang Y, Si H B and Feng Z M 2011 *Acta Phys.Sin* 60(4) 46-52
[3] Chen Z H, Zhang L, Jiang C Y, Cao Z G and Cui W 2019 WiFi CSI Based Passive Human Activity Recognition Using Attention Based BLSTM *IEEE Transactions on Mobile Computing* 18(11) 2714-2724
[4] Zhou S W, Zhang W, Peng D, Liu Y H, Liao X W and Jiang H B 2019 Adversarial WiFi Sensing for Privacy Preservation of Human Behaviors *IEEE 24(2) 259-263
[5] Cao Y J, Wang F H, Lu X X, Lin N, Zhang B, Liu Z and Stephan S 2020 Contactless Body Movement Recognition during Sleep via WiFi Signals *IEEE Internet of Things Journal*
[6] Wang Z J, Jiang K G, Hou Y H, Dou W W, Zhang C M, Huang Z H and Guo Y J 2019 A Survey on Human Behavior Recognition Using Channel State Information *IEEE Access* 7 155986-156024
[7] Liu J, Liu H B, Chen Y Y, Wang Y and Wang C 2019 Wireless Sensing for Human Activity: A Survey *IEEE Communications Surveys & Tutorials*
[8] Fuxjaeger P, Ruehrup S, Paulin T and Rainer B 2016 Towards Privacy-Preserving Wi-Fi Monitoring for Road Traffic Analysis *IEEE Intelligent Transportation Systems Magazine* 8(3) 63-74
[9] Traunmueller M W, Nicholas J, Awais M, Awais M and Kontokosta C E 2018 Digital footprints: Using WiFi probe and locational data to analyze human mobility trajectories in cities *Computers,Environment and Urban Systems* 72 4-12
[10] Zhu L J, Tong H J, Lou L L and Xiong Y 2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)
[11] Ding X B, Liu Z G and Xu H B 2019 The passenger flow status identification based on image and WiFi detection for urban rail transit stations *J. Vis. Commun. Image R* 58 119–129
[12] Xu J Y 2019 *M.S. Thesis* (Beijing: Beijing University of Posts and Telecommunications)
[13] Zhang G P 2003 Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. *Neurocomputing* 50 159-175
[14] Hochreiter S and Schmidhuber J 1997 Long Short-Term Memory *Neural Computation* 9(8) 1735-1780
[15] Chen T Q and Guestrin C 2016 Publication: KDD '16: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785-794