Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 9927800, 15 pages
https://doi.org/10.1155/2023/9927800

Research Article

MA-STS-Based Social Intimacy Analysis Algorithm Using Real Campus Network Data

Yifu Zeng,[1] **Xiangshu Qi**,[1] **Weiping Yang**,[2] **Jiatao Li**,[1] **Nian Pan**,[3] **and Guo Chen**[3]

[1]*College of Computer Science and Engineering, Changsha University, Changsha 410022, Hunan, China*
[2]*Center for Information of Education Management, Beijing 100816, China*
[3]*College of Computer Science and Electronic Engineering, Hunan University, Changsha 410012, Hunan, China*

Correspondence should be addressed to Guo Chen; guochen@hnu.edu.cn

Received 27 February 2023; Revised 7 April 2023; Accepted 1 June 2023; Published 15 June 2023

Academic Editor: Vittorio Memmolo

In recent years, the widespread availability of Wi-Fi in various settings, including universities, enterprises, and large shopping centers, has become increasingly prevalent. The user's time and location information embedded in wireless network systems can reveal individual and group social relationships, which indirectly reflect each person's psychological well-being. However, due to challenges in obtaining complete data, the high complexity of related data, and the absence of suitable data analysis models, few studies have analyzed student social behavior using data from university campus networks. This paper employs real-world data from a renowned Chinese university's wireless campus network for in-depth analysis and introduces a novel multiangle semantic trajectory similarity (MA-STS) algorithm to infer the intimacy and relationship types (such as teacher-student, friends, classmates, or romantic partners) between users. The experiments demonstrate that the proposed algorithm achieves an accuracy of over 95%.

1. Introduction

In recent years, frequent incidents of violence or self-injury on campus have occurred due to academic pressure, emotional difficulties, family problems, social and economic downturns, and other reasons, resulting in excessive psychological pressure. As a result, college students' mental health and counseling have gained attention. However, due to resource shortages, a lack of understanding of psychological problems, and communication gaps between teachers and students, mental health education at many colleges and universities is still in its early stages. If we can actively detect student social situations in the data, it can aid education managers in targeted mental health work.

In the prevailing university system, the campus network is a vital component of daily life. Figure 1 presents a statistical analysis of wireless AP usage across various locations in a renowned Chinese institution, demonstrating significant utilization of network equipment. Consequently, the system enables user mobility tracking and positioning, as well as a basic representation of users' social contexts and statuses based on their interactions at different times and locations. This information serves as a foundation for mental health and social relationship analyses.

Wireless AP data have advantages over GPS data in terms of user privacy and data authorization. Users are often unwilling to share GPS data for personal reasons, while wireless AP data provide convenient client information for data collection, enabling research and applications based on these data.

There has been some early work on exploring mobile and social behavior on university campuses using Wi-Fi tracking data. Kotz and Essien [1] studied the user's Wi-Fi usage behavior and used the quantified data of the number and total duration of the user's access to the AP to determine whether the user had a "residence location (home location)"and obtained the user's mobility. Kim and Kotz [2] established a mobile model by collecting the real mobile trajectories in Wi-Fi data AP on the Dartmouth campus and then analyzed the overlap degree of the trajectories by matching the mobile models among users, thus determining the intimacy relationship. However, much of the research is

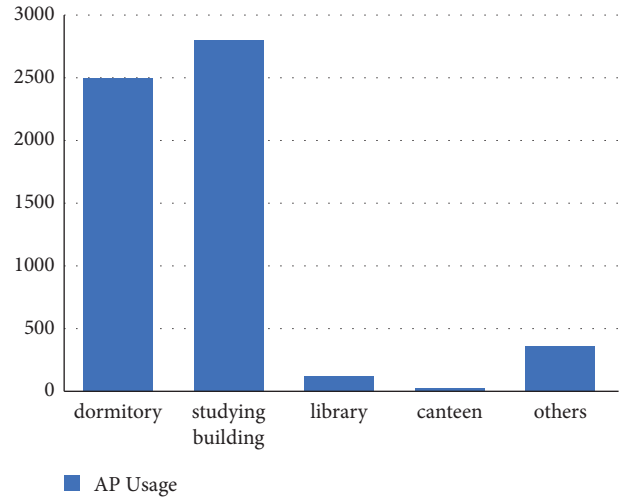

Figure 1: AP usage.

based on syslog messages. Fang and Hong [3] proposed to divide the behavior trajectory into two categories: learning-work (SW) and entertainment-entertainment (ER) and used two visual analysis algorithms (ST path and flow map) to speculate whether there was social relevance in people's behavior analysis in different time and space. Compared with syslog information, user-centric behavior trajectory tracking is easier to obtain. However, from the perspective of these research methods, we found that the analysis of data considers the similarity of users "mobile behavior from the geographical characteristics, Wang et al. [4] proposed two semantic trajectory algorithms, FA-STS and FP-STS, suitable for different crowd scenarios to measure the similarity between users. These two algorithms can get more realistic user relationships by considering both time and space dimensions. However, according to the scenario of the use of a university campus network, we found that the distribution of AP access point data in some geographical locations is not very clear. The distribution of aps is also not accurate to a certain classroom. Therefore, when looking for a classroom corresponding to a user who is staying in the hospital building to use the campus network, the house number of the hospital building cannot be accurately reached, and the data only use the name of the hospital building as the location of the access point. FA-STS and FP-STS algorithms do not consider the coarse and fine granularity of geographical location, which greatly reduces the accuracy of the results when judging the trajectory similarity of users.

In view of the existing usage records of campus networks in colleges and universities, as well as the shortcomings of existing related work, this paper designs a semantic trajectory similarity algorithm with higher accuracy to judge the intimacy between students and classifies the data reasonably through the classification algorithm of machine learning. Moreover, based on the Wi-Fi track of universities, students' social intimacy is mined to establish social networks and understand students' interpersonal relations and mental health development. Due to confidentiality concerns, the original data from 2020 were used in the research, and the data will not be released until 2022.

The main work of this paper includes: (1) using the real data used in the university campus network, the original data are collected, cleaned, and the semantic trajectory is extracted; (2) a trajectory similarity algorithm suitable for university Wi-Fi data (MA-STS) is designed to analyze the intimacy relationship between students. After the algorithm is classified and verified by the binary classification algorithm, the accuracy is greatly improved.

2. Related Work

2.1. Status of Research on Traditional Trajectory Similarity Algorithms. There are a variety of traditional models that consider similarity based on the trajectory of GPS data, and the overall classification is shown in Figure 2.

(1) Based on the Euclidean distance [5, 6], which is easy to understand and relatively simple to calculate, but is not suitable for comparing trajectories with different lengths and is sensitive to noise points.

(2) The calculation based on dynamic time warping (DTW, dynamic time warping) [7, 8] can solve the problem that the Euclidean distance cannot solve the trajectory length, and its main idea is: on two trajectories composed by time series, actively pick the nearest point on the other trajectory on the time axis and align with each other, so that the trajectory shape is as much as possible the same so as to get the maximum. The DTW algorithm is flexible, does not require the length of the trajectory, and measures trajectory similarity better. However, the algorithm does not consider noise points, and off-center points are also considered, resulting in a reduction in accuracy.

(3) The algorithm based on the longest common subsequence (LCSS) [9, 10] can effectively deal with the interference of noisy points by setting the threshold value and excluding the points in the deviated trajectory. However, the problem is that this algorithm does not define the minimum distance threshold well, and if it is not well defined, it may return trajectories that are not similar.

(4) The model based on the Hausdorff distance [11, 12] is also less effective against noisy points, and in most cases, it is used for classification and less often for predicting similarity matching problems.

(5) The size of the weighting ratio of the algorithm based on LIP (locality in-between polylines) [13] is proportional to the size of the overlapping area region. When the area is 0, it means that there is no overlap between trajectories and the LIP distance value is also 0. When the area weighted sum is larger, it means that the overlap area of trajectories is larger and the LIP value is also larger. This method works well for GPS data processing, but has some limitations in the study of Wi-Fi data application.

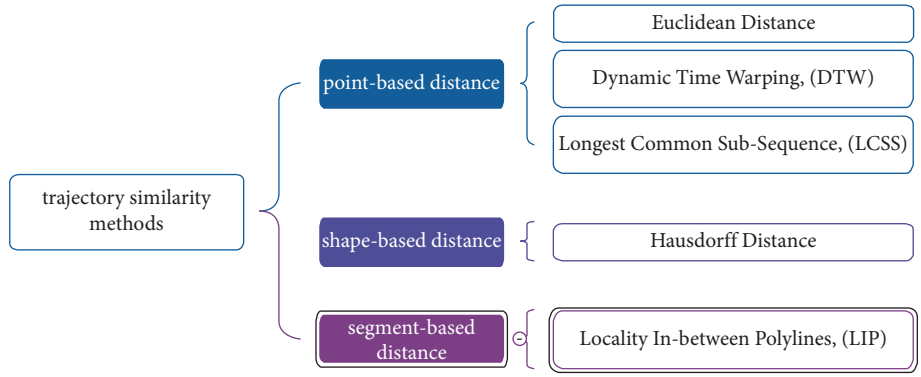

Figure 2: Classification of trajectory similarity methods.

As the above methods are based on GPS track data, every second of the user's movement will be recorded, while Wi-Fi data are dependent on the AP deployment location before the user's data record exists, and the user's movement between specific AP locations cannot be detected. Therefore, it is not very accurate to measure Wi-Fi data in terms of trajectory shape and distance. Therefore, based on the limitations of traditional similarity algorithms, this paper attempts to measure the similarity of trajectories between users using semantic trajectory algorithms.

2.2. Current State of Research on Semantic Trajectory Similarity Algorithms. The paper adopts the concept of semantic trajectories introduced by Spaccapietra et al. [14], where a semantic trajectory consists of a sequence of access points, with each trajectory point recording information, including coordinates, time, and the name of the access point. Two semantic trajectory algorithms [4], FA-STS and FP-STS, suitable for different crowd scenarios, are proposed to measure the similarity between users. The two algorithms mainly calculate the longest common subsequence between trajectories, and for friends who often travel together, resulting in more consecutive access points, the more similar the two trajectories are, but this algorithm ignores the similarity between users who often appear in the same location at the same time. Xiao et al. [15] proposed to estimate the similarity between users based on their physical location history and to achieve friend and location recommendations through potential social connections between users. The paper proposes the maximum move matching algorithm (MTM) to compare the similarity between trajectories. Geographic overlap, semantic overlap, and location order are used to infer whether users have similar points of interest. Such methods do not consider the time information of users arriving at the access point and are suitable for comparing trajectories with a wide range of user activities. Since most trajectory sequences of students on campus consist of dormitory -> academic building -> cafeteria, comparing mobile trajectories between students on campus through location sequences only will increase the similarity ratio to a large extent.

Therefore, this paper proposes a new trajectory similarity metric algorithm that considers the problem from multiple perspectives, calculating the similarity of trajectories using the degree of similarity in different situations, such as users appearing at the same access point consecutively at the same time, the frequency of users appearing at the same access point at the same time, and the duration of time. The algorithm does not only consider similarity matching in one dimension, but matches trajectory points with the same dimension in all dimensions.

2.3. Current State of Social Networking Research. Hou et al. [16] proposed to combine psychology and big data methods to explore the social patterns of university students and proposed the SEASON framework based on theories related to statistics and network science to predict the social network location of university students in different social scenarios, and the article also proposed the link-based prediction framework DEFINE to embed students' facial cognitive features, psychological features, and network structure features in social Althoff et al. [17] proposed an entropy-based model to predict social relationships between users and estimated the relationships between users by using the fixed slice method to calculate the frequency of cooccurrence. Baker and White [18] used student one-card campus access card data to mining social relationships.

The social network analysis data in this paper focuses on students' campus network usage data. In the research related to social relationships, complex social network algorithms are not applied to analyze the social networks among students, but rather the importance indicators between nodes based on multiangle similarity algorithms to measure the students' activity in the crowd.

3. Data Acquisition and Cleaning

Most of the current trajectory similarity algorithms do not consider the two dimensions of time and geographic space at the same time, and the accuracy of the prediction similarity algorithm needs to be improved. Based on these practical problems, this chapter collects the campus network usage

data of a famous university in China in 2020, cleans the data, and explains it in detail.

The campus wireless network of this university adopts a framework controlled by two major service manufacturers, Huazan and Ruijie, with 5800 wireless access points (aps). The provider server remembers the username and device Mac address through the user's student login record. When the user's phone opens the "Wi-Fi connection" operation, the router will detect the Wi-Fi request signal transmitted by the phone in its detection range and record the Mac address of the AP that the phone is connected to at this time and the Mac address of the device, the online time, the offline time, and the timestamp. The data collection system is shown in Figure 3. The raw data collected are shown in Tables 1–3.

The wireless AP log data in the above table are also provided by the two manufacturers of wireless network deployment in the campus network and is updated at a fixed time every day and stored on the local server. The data obtained from the study are the data obtained from March 2020 to December 2020. The two manufacturers set the time node to obtain the log file in hours and store it in the form of csv file. Because the data are on the server, in order to avoid the problems of transmission and data analysis efficiency caused by frequent access to the server, this design uses the FTP (file transfer protocol) class defined in the ftplib module of python to implement a simple ftp client, which is used to download the source log files on the ftp server at a time and store them in the local server for analysis and research.

3.1. Data Cleaning Is Mainly Divided into Two Parts

3.1.1. Number of Wireless Device Access. The processing and visualization of all time series are shown in Figure 4. The wireless AP source data truly reflect the situation of equipment access to the campus network wireless network of 2020 and accurately reflects the impact of force majeure such as COVID-19 Government Prevention Policy, COVID-19 University Prevention Policy, and Public Holidays.

According to the characteristics of the data obtained after data cleaning, the source data are adjusted and the existing source data are selected to a certain extent. For example, the missing period is removed because the original data are difficult to be obtained due to the equipment/manufacturer problems. Due to the epidemic in the first half of 2020, the log file reflected the real network usage, but most of the wireless network resources were in sufficient state, and most of the student users did not return to school and could not use the campus network. Therefore, there were no mobile trajectory data, so the data in this period were not analyzed.

3.1.2. Remove Invalid Mac. There is still a lot of noise and redundancy in the Wi-Fi data pulled from the server. Many Mac addresses collected in the original data set cannot find the corresponding student number information. Through analysis, it is found that the reason for this phenomenon is that there are pseudo-MAC addresses in these Mac addresses, that is, illegal or nonexistent Mac addresses. Therefore, we will clear the Mac address of the device that cannot match the student number information. Secondly, we found that there were many Mac addresses without moving behavior in the data. Such Mac addresses were generally detected by AP in fixed places, similar to the computer equipment equipped with Wi-Fi detection used by students in the laboratory. These Mac addresses are of no research value for us to analyze the movement trajectory, so they are also eliminated. In addition, the building information corresponding to many aps is not well improved, but through analysis, it is found that most of the building information with a large number of people has corresponding information, so we remove the Mac address of the AP that cannot find the corresponding building information.

4. User Trajectory Similarity Measurement

After the Wi-Fi data are cleaned and sorted, the trajectory between two users can be generated, as shown in Figure 5. From the semantic trajectory, we can consider the temporal and spatial information to compose a sequence of time and place. This paper compares the semantic trajectories between users to determine whether there is a similarity in behavioral movement patterns in the same time and proposes that the MA-STS algorithm comprehensively considers whether the trajectories between users are similar from more perspectives in semantic trajectory time (compared with FA-STS proposed by Wang et al. [4]). On this basis, the binary classification problem of machine learning can be used to train the weights of each algorithm to determine whether there is an intimate relationship between users.

Let tra_1 and tra_2 represent two semantic tracks, respectively; (stop_n, t_n) is a trajectory point on the semantic trajectory, where stop_n represents a dwell position, that is, the AP position; and t_n is the time when the mobile user appears in the dwell position stop_n, then the two trajectories tra_1 and tra_2 are expressed as

$$\begin{aligned} \mathrm{tra}_1 &= \{(\mathrm{stop}_1, \mathrm{t}_1), (\mathrm{stop}_2, \mathrm{t}_2) \ldots (\mathrm{stop}_n, t_n)\}, \\ \mathrm{tra}_2 &= \{(\mathrm{stop}_1, \mathrm{t}_1), (\mathrm{stop}_2, \mathrm{t}_2) \ldots (\mathrm{stop}_m, t_m)\}, \end{aligned} \tag{1}$$

$|\mathrm{tra}_1|$ and $|\mathrm{tra}_2|$ are the total length of time at the trajectory dwelling point, and the FA-STS [4] algorithm is expressed as

$$\mathrm{FA-STS}\left(\mathrm{tra}_1, \mathrm{tra}_2\right) = \frac{\mathrm{LCSS}_{\Delta T}\left(\mathrm{tra}_1, \mathrm{tra}_2\right)}{\left(\left|\mathrm{tra}_1\right| + \left|\mathrm{tra}_2\right|\right)/2}, \tag{2}$$

where LCSS is the longest common subsequence of tra_1 and tra_2 trajectories calculated. ΔT is the time difference between trajectory stopping points to determine whether the trajectories are similar to each other, requiring ΔT to be less than a certain threshold value set. The setting of the threshold value can greatly affect the accuracy of user similarity. In this case, the problem can be solved by using recursive computation, dividing the problem into smaller subproblems (i.e., solving for similarity between "prefix" pairs in a sequence) and using dynamic programming.

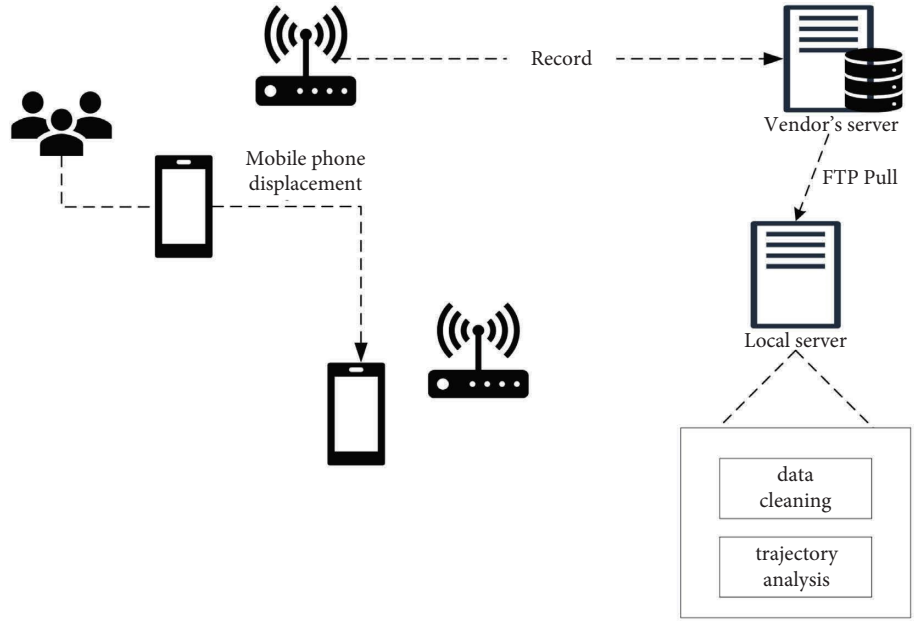

Figure 3: Wi-Fi data collection system.

The prefix is defined as follows:

For a sequence $Z = \{z_1, z_2, \ldots, z_m\}$, for $i = 0, 1, \ldots, m$, the ith prefix defining Z is $Z = \{z_1, z_2, \ldots, z_i\}$. For example, if $Z = \{C, M, A, O, C\}$, then $Z_4 = \{C, M, A, O\}$, if $X = < x_1, x_2, x_3, x_4, \ldots, x_m >$, $Y = < y_1, y_2, y_3, y_4, \ldots, y_n >$ is two sequences and $Z = < z_1, z_2, z_3, z_4, \ldots, z_k >$ is the common subsequence of the sequence, the analysis can be obtained as follows:

(1) If $X_m = Y_n$, then $Z_k = X_m = Y_n$, and Z_{k-1} is an LCSS of X_{m-1} and Y_{n-1}

(2) If $X_m ! = Y_n$, and $Z_k ! = X_m$, then Z is an LCSS of X_{m-1} and Y

(3) If $X_m ! = Y_n$, and $Z_k ! = Y_n$, then Z is an LCSS of X and Y_{n-1}

Therefore, if a two-dimensional array c is used to represent the LCSS length of the first i and first j characters of the sequence corresponding to X and Y, we can get the following:

$$c[i, j] = \begin{cases} 0, & \text{if } i = 0 \text{ or } j = 0, \\ c[i-1, j-1] + 1, & \text{if } i, j > 0 \text{ and } x_i = y_j, \\ \max(c[i, j-1], c[i-1, j]), & \text{if } i, j > 0 \text{ and } x_i \neq y_j, \end{cases} \tag{3}$$

In the FA-STS algorithm, if two semantic trajectory sequences $tra_1 = \{(\text{stop}_1, t_1), (\text{stop}_2, t_2) \ldots (\text{stop}_n, t_n)\}$, $\text{tra}_2 = \{(\text{stop}_1, t_1), (\text{stop}_2, t_2) \ldots (\text{stop}_m, t_m)\}$ are entered, the algorithm saves the value of $L[i, j]$ in an $n * m$ matrix L,, first calculating the first row of L from left to right, then calculating the second row, and so on, calculating the table entries in the main order. If ΔT is set to three min, given two semantic trajectories A and B, $A = \{$(apartment 1, 08 : 00), (canteen 1, 08 : 30), (comprehensive building, 09 : 15), (research building, 09 : 55), (canteen 1, 11 : 40)$\}$, $B = \{$(canteen 1, 08 : 20), (laboratory building, 08 : 37), (comprehensive building, 09 : 14), (scientific research building, 09 : 54), (canteen 1, 11 : 40)$\}$, the longest common subsequence is $\{$(comprehensive building, 09 : 14), (research building, 09 : 54), (canteen 1, 11 : 40)$\}$, LCS $S_{\Delta T}(A, B) = 3$, A and B obtained a trajectory similarity value of 0.6 by the FA-STS algorithm.

To a certain extent, the FA-STS algorithm can measure users with the same movement trajectory pattern at the same time, but after our analysis, we find that the FA-STS algorithm cannot consider the campus network users in the same time period. In fact, users who do not arrive at the same point in time and visit the same point in time, but have the same period of stay, often have an intimate relationship. An example is shown in Figure 6. (C1, C4) terminal access on AP1 at T1 moment. When T1 changes to T2, (C1, C2, C3, C4) terminal access on AP1. If scenarios similar to this frequently occur among (C1, C2, C3, C4), we consider it probable that a certain degree of intimacy exists between them. This is often overlooked in the way the FA-STS algorithm is calculated. Therefore, in order to achieve more accurate intimate relationship judgment, this section proposes another algorithm for calculating similarity.

If two different MAC terminal devices are detected by the same AP at the same time period, we call it time coincide, and the more time coincidence, the higher the correlation. The online time of the user access to the AP is arranged in chronological order to form the semantic trajectory of the user's movement track. Each user corresponds to a series of

Table 1: Sample wireless access point log data table.

	Device_Mac	AP_Mac	Up_time	Down_time	Duration
(Record according to time series)	Mac address of the device that connects to the wireless AP	Mac address of the AP	Device start to online time	Equipment down time	Equipment online time

Table 2: Geographic location information of wireless access points.

	AP_Mac	Location
(Record according to AP sequence)	Mac address corresponding to AP	AP specific geographical location

Table 3: Student number and equipment information.

	User_id	Device_Mac
(According to the online time record)	Student number	Mac address of the device connected to the wireless AP

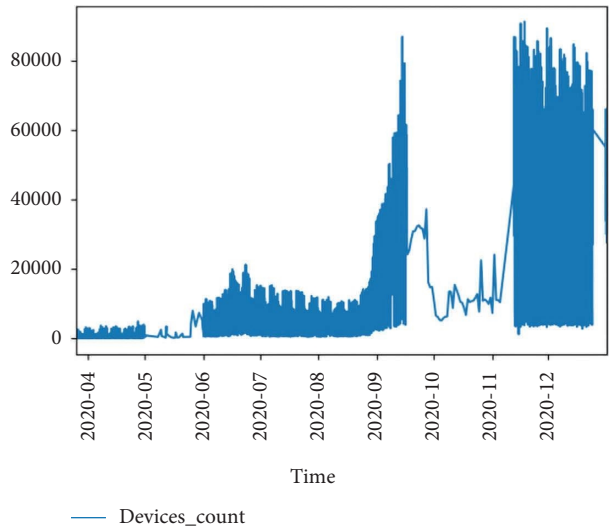

Figure 4: The number of campus network access.

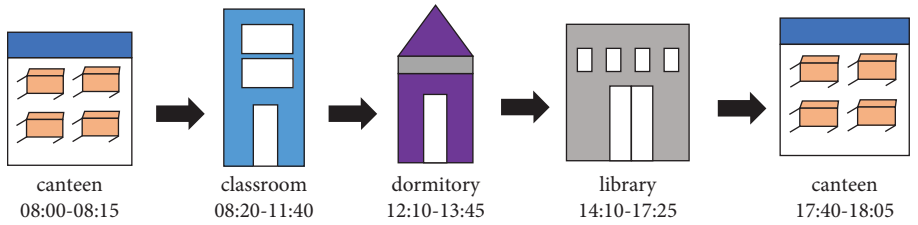

Figure 5: Example diagram of a semantic trajectory.

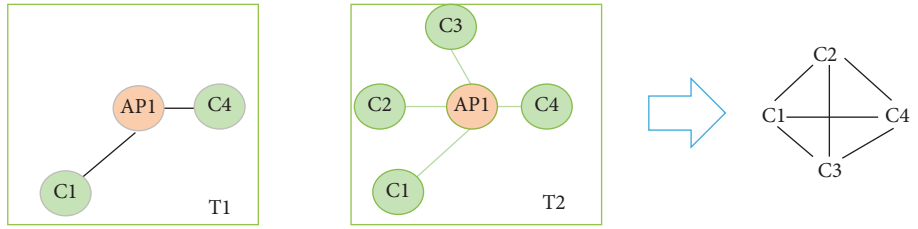

Figure 6: Location-based intimacy.

track records containing $((ST_i, ET_i), ap_i)$. The trajectory tra_1 is shown in

$$tra_1 = \{((ST_1, ET_1), ap_1), ((ST_2, ET_2), ap_2), \ldots, ((ST_n, ET_n), ap_n)\}, \quad (4)$$

where ap_i is the location where the terminal is connected to the MAC, and (ST_i, ET_i) is the start time and end time of the terminal in ap_i connection. This section measures similarity by the proportion of time periods that coincides in the two trajectories to the total duration. The similarity algorithm between the two terminals is expressed as

$$\text{sim1}_{i,j} = \frac{\sum_{ap}\text{len}\left(\text{overlap}\left(I\left(i\right), I\left(j\right)\right)\right)}{\left(\sum_{ap}\text{len}\left(I\left(i\right)\right) + \text{len}\left(I\left(j\right)\right)\right)/2}. \tag{5}$$

Among them, overlap () is used to calculate the time coincidence interval in the same access point, len () is the long coincident time, $I(i), I(j)$ is the online time of the two devices on the access point, in order to better show the similarity relationship between users, the algorithm also only believes that there may be an intimate relationship if the threshold exceeds the set value. If ΔT is set to 6 min, given two semantic locusals P and Q, $P = \{((08:00,\ 08:10),$ apartment 1), ((08 : 30, 09 : 10), canteen 1), ((09 : 15, 09 : 40, comprehensive building), ((09 : 55, 11 : 37), research building), ((11 : 40, 12 : 10), canteen1)}, $Q = \{((08:22,\ 08:37),$ canteen 1), ((08 : 40, 09 : 10), laboratory uilding), ((09 : 40, 09 : 19), comprehensive building), ((09 : 54, 11 : 35), scientific research building), ((11 : 40, 12 : 10), canteen 1)}, the coincident time is {((09 : 54, 11 : 35), scientific research building), ((11 : 46, 12 : 10), canteen 1)}, overlap $(P, Q) = 125$ (min), P and Q through the position correlation algorithm to obtain the trajectory similarity value of 0.78, see algorithm 1 for detailed steps. To calculate the coincident time period algorithm between the two trajectories, first iterate the trajectories sorted in chronological order, determine whether the end time of the tra2 time interval is less than the start time of tra1, whether the start time of tra2 time interval is greater than the end time of tra1, and if the conditions are met, the corresponding trajectory with a small start time value takes the time interval later; If there is a coincident interval, the time difference between the coincident interval is taken.

5. User Social Relationship Analysis and Model Evaluation

Since there is some ambiguity in our Wi-Fi data, they ignore the number of common items between moving objects in the coincident time period. As shown in Table 4, FA-STS (A, C) = 0.03, Sim1 (A, C) = 0.44, they may get high similarity even with few common items. In fact, A and C have only one identical stay interval, and there is no close relationship. For example, Sim1 (A, B) = 0.17, the similarity value is very small, but FA-STS (A, B) = 0.35, the common items are relatively large consistent with the expected close relationship judgment.

In order to solve this problem, we propose to use (6) to consider the similarity of common terms of nonrepeating locations in the coincident interval.

$$\text{sim2}_{i,j} = \frac{\sum_{ap}\text{count}\left(\text{overlap}\left(I\left(i\right), I\left(j\right)\right)\right)}{\left(\sum_{ap}\text{unique}\left(I\left(i\right)\right) + \text{unique}\left(I\left(j\right)\right)\right)/2}. \tag{6}$$

In this algorithm, count () calculates the number of overlapping interval locations between two users. It should be noted that in order to better measure the similarity accuracy between fuzzy data, AP needs to remove the duplicate value, and unique () is to remove the duplicate value of the user's stay area. That is, the ratio of the total number of nonrepeating aps in the coincidence interval and the average sum of nonrepeating aps in the two tracks.

Therefore, based on the above three similarity algorithms, MA-STS (multiangle algorithm for semantic trajectory) algorithm can be defined in the following:

$$\begin{aligned}\text{MA} - \text{STS} = {} & w_1 \times \text{FA} - \text{STS}\left(\text{tra}_1, \text{tra}_2\right) + w_2 \times \text{sim1}_{i,j} \\ & + w_3 \times \text{sim2}_{i,j},\end{aligned} \tag{7}$$

where w denotes the weight proportion of the three different similarity algorithms, and the training process of the weights will be introduced in detail in the next section. After three different similarity algorithms train different weight values, the similarity score ratio will be obtained to predict whether there is an intimate relationship between users.

Each two semantic trajectory will be evaluated by the above three similarity algorithms to obtain the similarity scores of different angles. As shown in Figures 7 and 8, the X-axis represents the ratio of the longest common subsequence obtained by FA-STS algorithm, the Y-axis represents the time cumulative ratio of the coincidence interval, and the Z-axis represents the ratio of the nonrepeating AP common item of the coincidence interval. The blue triangle is the presence of intimacy, and the red dot is the relationship between strangers. As can be seen from the figure, blue graphics are mostly distributed in spaces with small values of the X-axis, while red patterns are distributed in different spaces according to different values of X, Y and therefore, we propose a binary problem training model to carry out weighted classification of these similarity algorithms.

Binary classification problems are often used to solve problems where the output is 0 or 1, in our case, close relationship. Among them, we provide 1500 pairs of known close relationships of users through volunteers, which are divided into 1300 pairs of training sets and 200 pairs of testing sets. Each pair of users can obtain similarity scores of three different values through the above similarity algorithm. Therefore, we take the three-in-one-out neuron learner shown in Figure 9 to process the data, and the network includes an input layer, a hidden layer, and an output layer.

The input layer of the training set has three characteristics as

$$T = \left\{\left(x_1, y_1, z_2\right), \left(x_2, y_2, z_2\right), \ldots \left(x_m, y_m, z_m\right)\right\}. \tag{8}$$

Since the input layer has three features, the length of W is $1 * 3$, and the weight matrix is expressed as

```
Enter trajectory tra1, trajectory tra2
Output the ratio of the total time spent on Mac online during the overlapping time period of tra1 and tra2
(1)  m = len (Time (tra1)), n = len (Time (tra2)), l = k = 0
(2)  if m and n:
(3)    while True:
(4)      if l >= m && k >= n then
(5)        break
(6)      if tra2 [k].time [1] <= tra1 [l].time [0] then
(7)        k ++
(8)        continue
(9)      elif tra2 [k].time [0] >= tra1 [l].time [1] || tra2 [k].time [1] >= tra1 [l].time [1] then
(10)       l ++
(11)     if tra1.pos = = tra2.pos then
(12)  time_stamp←max (tra2 [k].time [0], tra1 [l].time [0])−min (tra2 [k].time [1], tra1 [l].time [1]).total_seconds ()
(13)       associated_time + = time_stamp
(14)     end if
(15)   end while
(16)   return (associated_time/(m + n/2))
```

Algorithm 1: Calculate the coincident time period algorithm of the trajectory.

Table 4: Examples of different algorithms and common entries for users.

Moving objects	FA-STS	Sim1	Public items	Intimate relations
(*A*, *B*)	0.17	0.35	3	1
(*B*, *C*)	0.12	0.43	1	0
(*C*, *A*)	0.03	0.44	1	0
(*D*, *A*)	0.46	0.24	4	1

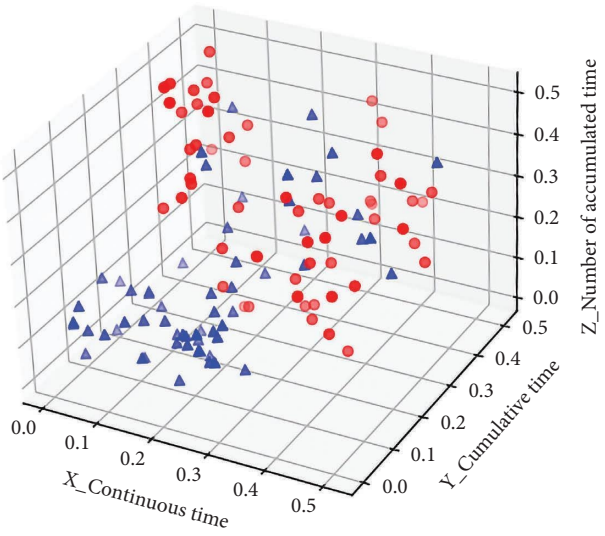

Figure 7: 3D scatter plot of similarity scores.

$$W = \left(w_{1,1}, w_{1,2}, w_{1,3}\right),$$
$$w_{1,i} = \frac{1}{N}, i = 1, 2, 3. \tag{9}$$

B has length $1 * 1$, the number of rows is always the same as W, and the number of columns is 1, which is expressed by

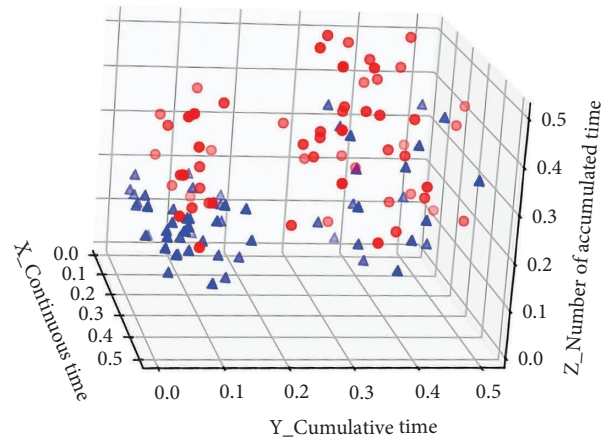

Figure 8: Rotated 45° similarity score 3D scatter plot.

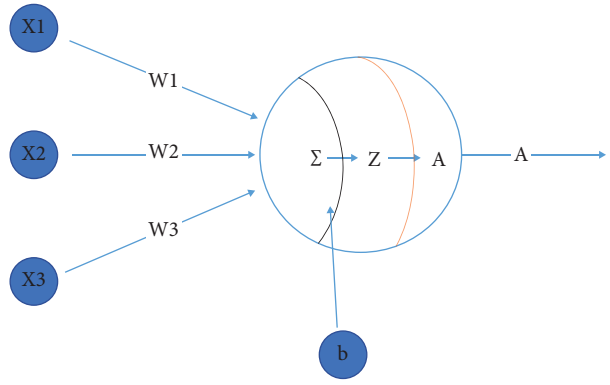

Figure 9: Binary classification problem-three in and one out neurons.

$$B = \left(b_{1,1}\right). \tag{10}$$

The expression of the output layer is

$$\begin{aligned} Z &= W_{1,1} * X_1 + W_{1,2} * X_2 + +W_{1,3} * X_3 + B, \\ A &= \text{Sigmoid}(Z). \end{aligned} \tag{11}$$

The binary cross entropy loss function is defined as (12), and the way of classification is determined by the threshold of A

$$J = -[Y \ln A + (1 - Y) \ln (1 - A)]. \tag{12}$$

In the case of binary classification in this paper, positive examples are marked as 1 while negative examples are marked as 0 when we label samples as

$$\mathrm{Y} = \left(\mathrm{Y}_1\, Y_2\, \ldots\, Y_m\right) = (1\,1\,0 \ldots 1). \tag{13}$$

We find the optimal weight solution through the way of back propagation decline of the neural network, so the gradient of W is derived as

$$\frac{\partial \mathrm{J}}{\partial W} = \frac{\partial \mathrm{J}}{\partial A} \frac{\partial \mathrm{A}}{\partial Z} \frac{\partial \mathrm{Z}}{\partial W} = \frac{A - Y}{A(1 - A)} \,.\, A(1 - A).X^T = (\mathrm{A} - \mathrm{Y})X^T. \tag{14}$$

The detailed work flow of training classification model algorithm combined with AdaBoo algorithm is shown in Algorithm 2:

After randomly initializing the connection weight and threshold value in the network, the weighted error of each round is calculated in the fifth line. If the error is too large and unsatisfactory, the weight value is reset. After the continuous iterative update of the data connection weight, the loss rate and accuracy of the data operation result are shown in Figures 10 and 11, respectively. In the process of training, due to the excessively small data set, underfitting occurred at the beginning, so we tried to increase the data quantity of training set and adjust the data quantity batches of test set to improve the test effect. The evaluation results of binary AUC model are shown in Figure 12. As can be seen from the effect in the figure, our training accuracy reaches 0.95, indicating that the similarity algorithm is very feasible for judging the intimacy between two users.

6. Experiments

6.1. MA-STS vs. FA-STS. To verify the accuracy of the MA-STS algorithm in measuring user closeness relationships, we compared it with the existing FA-STS algorithm using Wi-Fi data rather than GPS positioning data. The experimental data included 1500 pairs of users with known intimacy relationships provided by volunteers. The algorithm judged whether there was an intimacy relationship between users, and was considered correct if it matched the user's actual intimacy. We varied the threshold settings to influence the results and conducted experiments using different threshold settings. Figure 13 shows the experimental results for both algorithms.

As can be seen from Figure 13, the MA-STS algorithm is significantly more accurate than the FA-STS algorithm under the same sample test. When the threshold is set at about 3 minutes, the accuracy of the algorithm is relatively high, because if the threshold is set too short, the data result will be misjudged, but when the time is set too long, it is equivalent to only considering the influencing factors of geographical location when judging the trajectory, which will also lead to misjudgment of the data result. In summary, although the calculation process of the MA-STS algorithm in this paper is more complicated, it is more accurate than other algorithms in determining whether there is an intimacy relationship between users.

6.2. Social Intimacy Authenticity Test Combined with Decision Tree Model. Using MA-STS, we can detect intimate relationships among campus network users by examining similarity score values across different time periods and places, particularly among classmates, romantic partners, and friends. In this section, we experimentally predict the relationships between nonstranger users as classmates, romantic partners, or friends, and compare them with actual relationships to verify the accuracy of our proposed social intimacy analysis algorithm.

In terms of time characteristics, according to the common time schedule on campus, we can divide the time period into three types: Monday to Friday class time (7 : 00–18 : 00); Monday to Friday break time (18 : 00–24 : 00); and weekend time (7 : 00–24 : 00).

In terms of location characteristics, we categorized locations based on the distribution of access points and campus network Wi-Fi data into four types: teaching buildings, dormitories, research buildings, and libraries. Undergraduates frequently use academic buildings, dormitories, and libraries, while graduate students spend most of their time in research buildings. To improve the accuracy of our experiments, we analyzed the data of undergraduates and graduates separately.

Using the intimacy dataset from the previous chapter, we obtained 1500 pairs of users with known social relationships through volunteer participation. Figures 14–16 display the average similarity score ratio based on time and location characteristics. Our analysis indicates that classmate relationships are more prevalent during class time and in teaching buildings, while couple relationships tend to occur more frequently on weekends and in libraries or research buildings. Friendships occur evenly across different times and places. We used a decision tree model to classify social relationships based on changes in similarity scores across different features.

In order to analyze social relationships based on similarity scores, we propose an information gain ratio C4.5 algorithm based on decision tree classification technology, that is, use information gain ratios to select different features and the next node. Information entropy usually measures the uncertainty probability of a sample, and the more uncertain and chaotic the standard value, the greater the entropy value and its calculation is as follows:

$$\text{Entropy}(D) = \sum_{k=1}^{|y|} p_k \log_2 p_k, \tag{15}$$

Input training set $D = \{(x_k, y_k)\}_{k=1}^{m}$;
learning rate η
process:
(1) Initialize all connection weights and thresholds in the network randomly in the range (0, 1)
(2) **while**
(3) **for all** $(x_k, y_k) \in$ D **do**
(4) The gradient term of neurons in the output layer is calculated according to equation (14) g_i
(5) $\varepsilon_i = \mathbf{1/N}[\sum_j w_j I(D_i(x_j) \neq y_j)]$
(6) **if** $\varepsilon_i > 0.5$ then
(7) $w \leftarrow \{w_j = 1/N | j = 1, 2, \ldots, N\}$
(8) $i = i-1$
(9) break
(10) **end if**
(11) The gradient term of hidden layer neurons was calculated according to equation (14) e_h
(12) **end for**
(13) **until** Meet the stop condition
(14) **Output**: connection weights with thresholds determined by the neural networks

Algorithm 2: Training classification model algorithm.

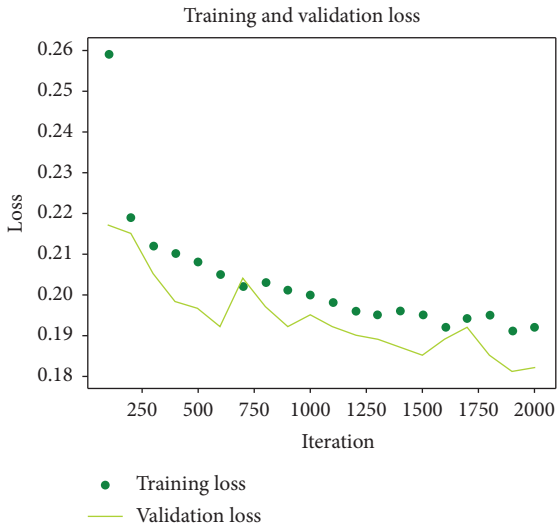

Figure 10: Dataset loss rate.

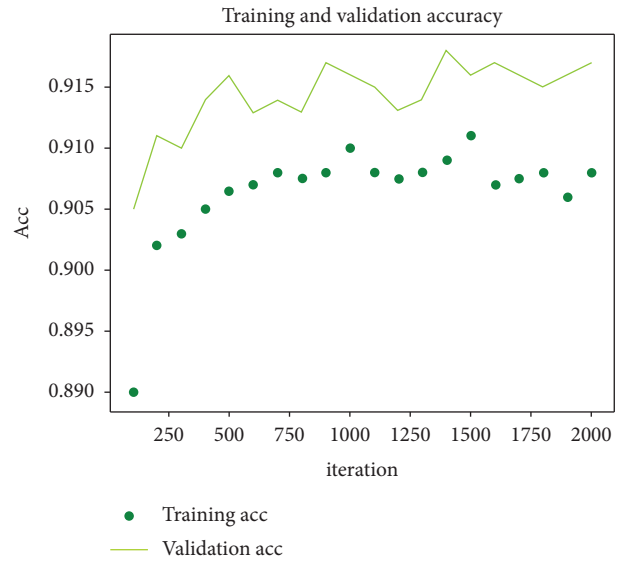

Figure 11: Dataset accuracy.

p_k is the proportion of the current sample to the whole sample. In order to have an ideal decision tree training information effect, it is necessary to reasonably select the features that distinguish the nodes in order to recursively construct the correct decision tree layer by layer. Generally, in the process of construction, the size value of the information gain is used for the selection of features, expressed as the difference between the front and back entropy, i.e., information gain = Entroy (front)-Entroy (back), the formula is expressed as.

$$g(D, A) = E(D) - E(D|A). \tag{16}$$

However, the information gain value tends to prefer features that take more values. The certainty of the sample can be better estimated by multiplying the information gain by a penalty parameter, which is the inverse of the entropy of the data set D using feature A as a random variable as

$$g_R(D, A) = \frac{g(D, A)}{E_A(D)}. \tag{17}$$

Combining the above information, we propose the following algorithm for constructing a decision tree model Algorithm 3:

To construct a decision tree, the algorithm requires input training datasets, feature values, and thresholds. Starting from the root node, the information gain ratio of all features is calculated, and the feature with the maximum information gain ratio is used to create child nodes. The algorithm is called recursively until all feature information gain ratios are small or there are no features to select. The completed decision tree is then used to predict social relationships. Table 5 displays characteristic values for graduate students in relationships as couples, classmates, and friends and measures students' behavior using eight indicators. Gender

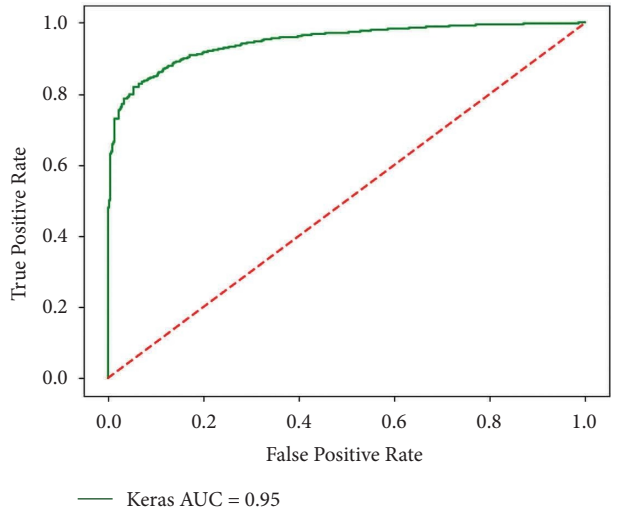

Figure 12: The AUC model evaluates the effect.

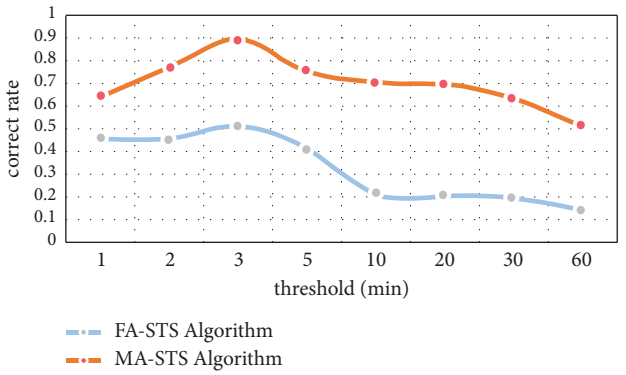

Figure 13: Comparison of accuracy of trajectory similarity algorithms.

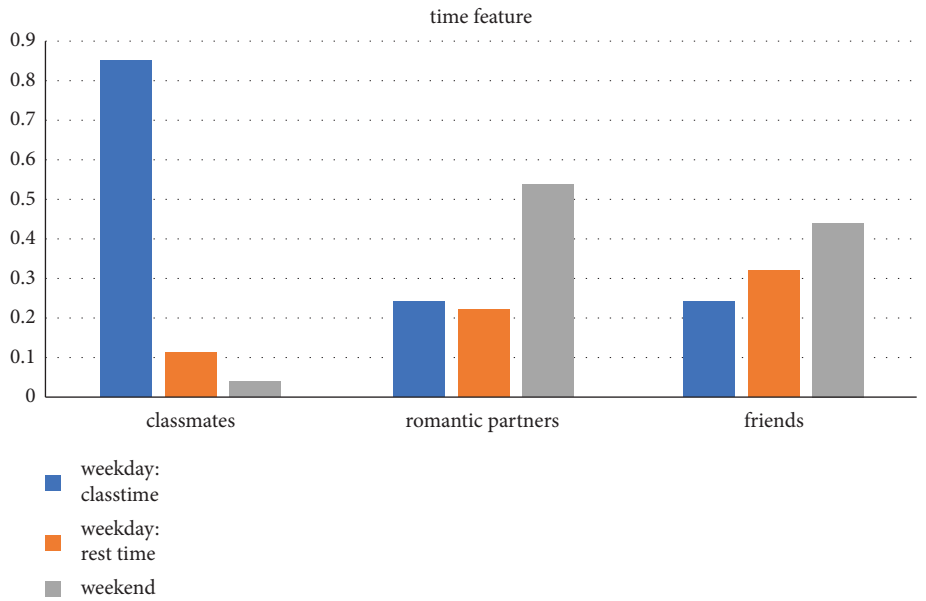

Figure 14: Average similarity scores by time.

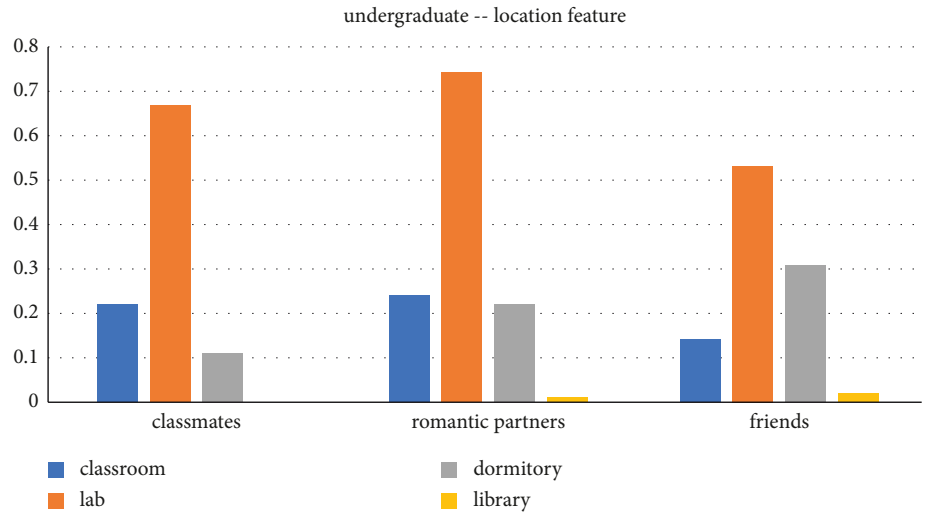

Figure 15: Average similarity scores of undergraduates by location.

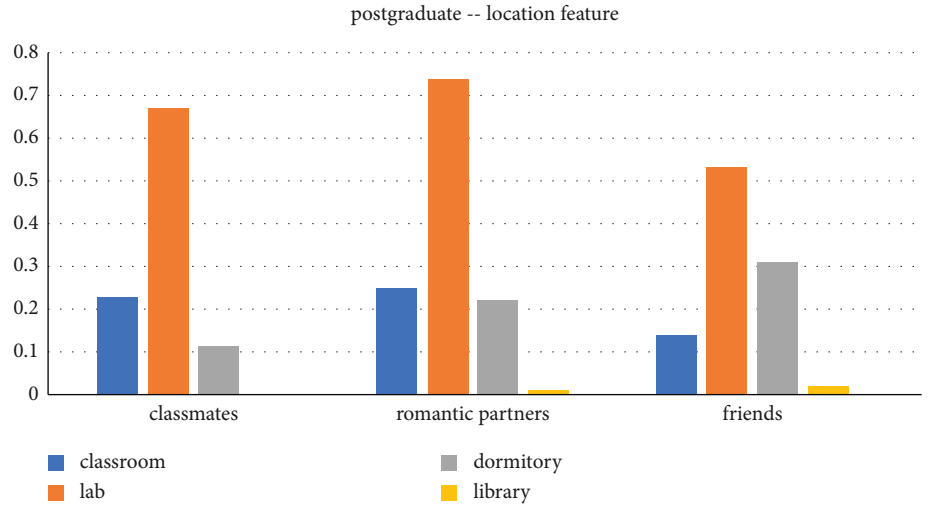

Figure 16: Average similarity scores for graduate students by location.

Enter Training Set D, Eigenvalue A, Threshold ε
Process:
(1) **while** (The decision tree does not classify all the examples correctly)
(2) According to equations (3)-(4), we get the attribute A_t of the instance contained in D that best distinguishes from A
(3) D_t is the dataset that reaches node t
(4) **if** $g_r < \varepsilon$
(5) A_t is marked as leaf nodes for most of class C
(6) **else**
(7) Splitting D_t into smaller subsets, A_t is not a leaf node
(8) **end if**
(9) **end while**
(10) **return:** decision tree

Algorithm 3: Training decision tree model algorithm.

Table 5: Feature values.

Teaching buildings	Research building	Dormitory	Library	Monday to Friday class time	Closed from Monday to Friday	Weekend	Gender
0.154	2.56	0.143	0.078	1.356	1.223	2.89	1
0.145	2.245	0.11	0.0	2.78	0.12	0.06	0
0.85	2.14	1.35	0.02	1.24	1.44	1.89	0

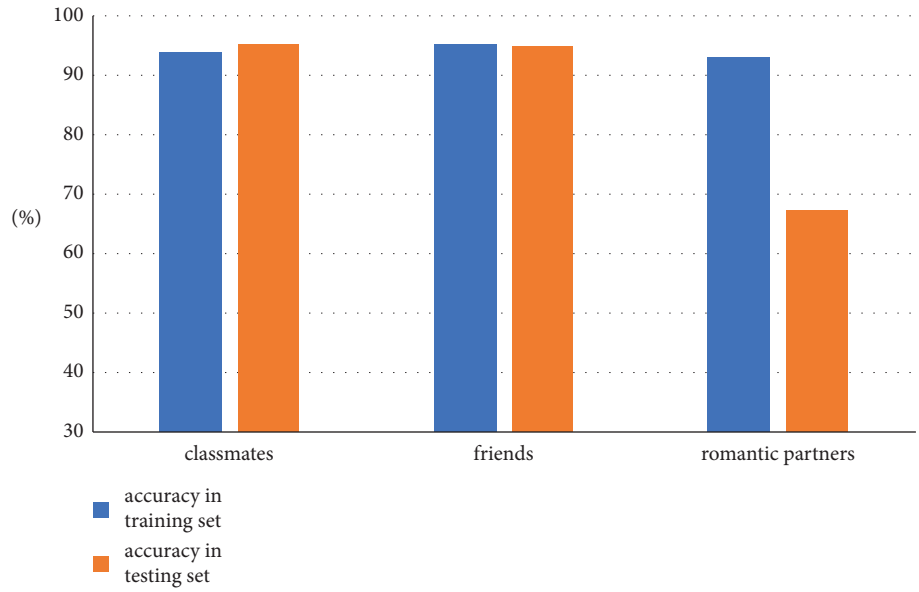

Figure 17: Dataset accuracy.

information is obtained from the student ID in the data system.

In our experiments, the same was divided into a training set of 1300 pairs used to construct the decision tree and a test set of 200 pairs used to evaluate the accuracy of the prediction model. As can be seen from Figure 17, based on the social closeness analysis algorithm proposed in this paper, the decision tree model can effectively determine the social relationship between two students with a correct rate of 94% in the training set and 85.8% in the testing set. The accuracy rate of friends, classmates, and strangers in the training set was above 90%, and in the testing set, the accuracy rate of friends and classmates was 95%, while the accuracy rate of romantic partners was 67.4%. This further confirms the practical value of the social intimacy analysis algorithm proposed in this paper.

7. Conclusions

This paper proposes the multiangle semantic trajectory similarity (MA-STS) algorithm to calculate user intimacy, which consists of three similarity algorithms. The first considers users arriving at the same access point at the same time with high frequency as similar; the second calculates the time period overlap between two users at the same location; and the third calculates the ratio of users visiting different locations to solve geographical ambiguity in campus network data. Experiments show MA-STS outperforms FA-STS in accuracy, and predictions of social relationships (friends, classmates, or romantic partners) using MA-STS are also highly accurate when combined with volunteer survey data. However, due to the complexity of human social behavior, there is still room for improvement in inferring intimacy and relationship types.

Data Availability

The data that support the findings of this study are available from the corresponding author and the first author upon reasonable request for the reason of protecting student privacy.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The research was partially funded by the Natural Science Foundation of Hunan Province (grant no. 2021JJ40635), the Scientific Research Foundation of Hunan Provincial Education Department (grant no. 21A0535), the National Natural Science Foundation of China (grant no. 62006030), and the National Key R&D Program of China (grant no. 2020YFB2104000).

References

[1] D. Kotz and K. Essien, "Analysis of a Analysis of a Campus-Wide Wireless Networkampus-wide wireless network," *Wireless Networks*, vol. 11, no. 1-2, pp. 115–133, 2005.

[2] M. Kim and D. Kotz, "Modeling users' mobility among WiFi access points," in *Proceedings of the Papers Presented at the 2005 Workshop on Wireless Traffic Measurements and Modeling*, pp. 19–24, Berkeley, CA, USA, 2005.

[3] T. Fang and X. Hong, "Discovering meaningful mobility behaviors of campus life from user-centric WiFi traces," in *Proceedings of the Of SouthEast Conference*, pp. 76–80, Kennesaw, GA, USA, April 2017.

[4] F. Wang, X. Zhu, J. Miao, and J. Miao, "Semantic trajectories-based social relationships discovery using WiFi monitors," *Personal and Ubiquitous Computing*, vol. 21, no. 1, pp. 85–96, 2017.

[5] H. Breu, J. Gil, M. Werman, and D. Kirkpatrick, "Linear time Euclidean distance transform algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 529–533, 1995.

[6] B. Wang, X. Liu, B. Yu, R. Jia, and X. Gan, "An improved WiFi positioning method based on fingerprint clustering and signal weighted Euclidean distance," *Sensors*, vol. 19, no. 10, p. 2300, 2019.

[7] M. Müller, *Dynamic Time Warping Information Retrieval For Music And Motion*, Springer, Berlin Heidelberg, 2007.

[8] Y. Xi, D. Huang, Y. Yuan, Z. Liu, and K. Anish, "Improved dynamic time warping algorithm for bus route trajectory curve fitting," *Journal of Transportation Engineering, Part A: Systems*, vol. 147, no. 8, Article ID 04021044, 2021.

[9] J. W. Hunt and T. G. Szymanski, *A fast algorithm for computing longest common subsequences Communications of the ACM*, vol. 20, no. 5, pp. 350–353, 1977.

[10] W. Bian, G. Cui, and X. Wang, "A trajectory collaboration based map matching approach for low-sampling-rate GPS trajectories," *Sensors*, vol. 20, no. 7, p. 2057, 2020.

[11] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[12] C. Wang, S. Zourlidou, J. Golze, and M. Sester, "Trajectory analysis at intersections for traffic rule identification," *Geo-spatial Information Science*, vol. 24, no. 1, pp. 75–84, 2021.

[13] N. Pelekis, I. Kopanakis, G. Marketos, and N. Irene, "Similarity search in trajectory databases," in *Proceedings of the 14th international symposium on temporal representation and reasoning (TIME'07). IEEE*, pp. 129–140, Alicante, Spain, June 2007.

[14] S. Spaccapietra, C. Parent, M. L. Damiani, and J. A. De Macedo, "A conceptual view on trajectories Data & knowledge engineering," *Data \& knowledge engineering*, vol. 65, no. 1, pp. 126–146, 2008.

[15] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Inferring social ties between users with human location history," *Journal of Ambient Intelligence and Humanized Computing*, vol. 5, no. 1, pp. 3–19, 2014.

[16] J. Hou, H. Pan, T. Guo, I. Lee, and X. Kong, "Prediction methods and applications in the science of science: a survey," *Computer Science Review*, vol. 34, Article ID 100197, 2019.

[17] T. Althoff, P. Jindal, and J. Leskovec, "Online actions with offline impact: how online social networks influence online and offline user behavior," in *Proceedings of the Of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 537–546, Cambridge, United Kingdom, February 2017.

[18] R. K. Baker and K. M. White, "Predicting adolescents' use of social networking sites from an extended theory of planned behaviour perspective," *Computers in Human Behavior*, vol. 26, no. 6, pp. 1591–1597, 2010.