Study on Data Clustering and Intelligent Decision Algorithm of Indoor Localization

Zexi Liu

1College of Computer Science, Chongqing University, Chongqing, 400044, China

Email: 345442314@qq.com

Abstract. Indoor positioning technology enables the human beings to have the ability of positional perception in architectural space, and there is a shortage of single network coverage and the problem of location data redundancy. So this article puts forward the indoor positioning data clustering algorithm and intelligent decision-making research, design the basic ideas of multi-source indoor positioning technology, analyzes the fingerprint localization algorithm based on distance measurement, position and orientation of inertial device integration. By optimizing the clustering processing of massive indoor location data, the data normalization pretreatment, multi-dimensional controllable clustering center and multi-factor clustering are realized, and the redundancy of locating data is reduced. In addition, the path is proposed based on neural network inference and decision, design the sparse data input layer, the dynamic feedback hidden layer and output layer, low dimensional results improve the intelligent navigation path planning.

1. Introduction

Indoor positioning technology extends the relation of human perception space geography from outdoor to indoor, so that human beings have the ability of positional perception in architectural space.

Unlike satellite localization technology, due to the complexity of indoor environment, signal transmission attenuation, indoor location technology to conquer signal real-time tracking, multipath signals more effective compensation, and the key technology of network coordination [1].

United States proposed the E911 program to develop the indoor localization technology. Apple has introduced beacon’s indoor localization technology to raise indoor positioning accuracy from 30 meters to about 10 meters. Ericsson used 4G base station technology to optimize the positioning accuracy of wide-area to about 30. Japan has also proposed more advanced indoor targeting technology. EU has proposed a law to promote the development of indoor localization technology. In 2009, China developed a technology called “XiHe”. It used TC - OFDM to improve the ability of indoor and outdoor seamless positioning. According to the survey statistics, indoor localization technology will usher in nearly one trillion in the market space by 2020, and by using base stations, terminals, platforms and services, it will form a new industrial chain. Indoor positioning technology will be applied to fire rescue, emergency management and other social services.

The current indoor localization technology is mainly based on distance measurement, fingerprint and inertial navigation method. These three methods have different advantages and disadvantages in the range, precision and complexity of indoor positioning. But it still has the following problems:
The coverage of single network is insufficient, and multi-source network coordination is required. The localization algorithm based on distance measurement depends on the transmission interval of pilot frequency, and it is difficult to achieve high precision in one room [2]. However, the indoor localization technology based on fingerprint needs dense nodes to realize the high precision of local area. Therefore, multi-network association has become an important idea of indoor localization.

(2) Localization data is redundant, data processing is very complex and location services need precise planning for the navigation path. In the condition of multi-network co-variation, the location data involved is huge [2]. At the same time, the indoor scene is complex and needs to be effectively perceived, only in the way can navigation be more reliable.

To solve the problems of lacking of precision and data redundancy in indoor positioning, this paper put forward an algorithm based on the data clustering and intelligent decision research of indoor localization. By optimizing the clustering process of massive indoor localization data, the data normalization pretreatment, multi-dimensional controllable clustering center and multi-factor clustering are realized, and the redundancy of locating data is reduced. In the end, to make the path navigation more intelligent, we used neural network inference and decision, designed the sparse data input layer, the dynamic feedback hidden layer and low dimensional output layer.

2. Basic knowledge of multisource indoor localization

2.1 Localization based on distance measurement
Using the principle of geometry to calculate the location of the target, it can determine the location of the target by the channel arrival time or the arrival angle.

Wang Qin from University of Science and Technology Beijing proposed a TOA localization algorithm of ranging error classification. Based on the RSSI and RITEM values in the TOA ranging process, we can estimate the error level. By using the maximum likelihood method, we may choose the label position of maximum probability as positioning results to improve the positioning precision of the algorithm[3].

Jing Hao from University of Nottingham put forward a wireless signal distance weighted indoor co-localization method. He used particle filter integrating with course-calculated data, WiFi data, distance between the users, indoor maps and other information to eliminate the error when signal is not stable or database dose not update on time.

Xu Peipei from Southeast University put forward a scheme of TOA without time synchronization process. By using bi-directional ranging method to realize positioning devices. She loaded the transmission time and receiving time to the data frame, to complete the process of ranging between two points. This scheme reduced the consumption of the time synchronization module [4], at the same time, this improved localization algorithm made the whole positioning system more accurate and robust.

2.2 Localization based on fingerprint
By using the distribution characteristics in different spaces, comparing the characteristics of the actual current environment to the measured feature, we can determine the current relative position of the target. Fingerprint localization generally uses signal strength as the characteristic, the localization accuracy depends on the distribution density of the node.

Tao Zheng from Dalian University of technology put forward a WLAN location fingerprinting positioning algorithm based on the chi-square distance. In this method, we can use fast K-means clustering algorithm to cluster reference point of location area. This method enhanced the robustness of data processing. By using the method of sensitivity and chi-square distance as the distance measure function in position fingerprint localization algorithm [5], we can reduce the effects of environmental noise on the positioning accuracy.

Dan Yu from Taiyuan University of science and technology proposed an algorithm based on technology of SLAM to optimize the fingerprint data quality and to enhance the uniqueness of
fingerprint. This algorithm ensured the quality of fingerprint data and essentially improved the performance of the fingerprint orientation.

Li Wenhao from Liaoning University proposed that by using Matrix Completion (MC) theory we can establish the off-line position fingerprint effectively. He also put forward a new type of algorithm which is called BSA – SVT. In the online positioning stage he put forward a positioning algorithm based on the DE - BCS, further improving the positioning accuracy.

2.3 Localization based on inertial device

Inertial navigation measures the acceleration of the carrier in the inertial reference system, by integrating it with time and converting it into the navigation system, we can get information in the navigation system such as speed, yaw angle and position.

Gong-min Yan of Northwestern Polytechnic University put forward the compass initial alignment algorithm based on position error [6]. He combined it with the strap down inertial navigation attitude algorithm and get the Positioning damping strap down inertial navigation algorithm. This algorithm is easy to calculate and it’s very stable, but compared with Kalman filter for integrated navigation, there’re still lots of shortcomings in the application conditions.

Xi Wen et al. from Nanchang University obtained the inertial navigation device data on the basis of the Kinect visual range. Using Kalman filter, he optimized the error of position and posture as observed quantity to fix the error. This method improve the positioning accuracy and stability of indoor mobile robot.

Lei Wang from Tongji University used UWB module installed on the mobile node to produce the estimated value of position. This method calibrated the position of moving nodes obtained by inertial navigation algorithm. It has higher indoor positioning accuracy.

3. Clustering processing of large amount of indoor location data

Because in the multi-source indoor positioning there are base station, Wi-Fi, inertial navigation and infrared and other infrastructure, data collected including images, voice, text, etc. are heterogeneous. There are great difference between size of different expression data, and at different time and location, there exists great information redundancy. For this reason, we introduce clustering process of large amount of indoor positioning data. In this process, we set the Euclidean distance as the criteria for the classification of elements, set the simplified location information as the aim of clustering processing. We iterate this classification process until the location information has no obvious change.

3.1 Euclidean similarity measure

Assume a given location data set \( X = \{ x_m | m = 1, 2, ..., n \} \), we use \( d \) successive attributes \( A_1, A_2 ... A_d \) to describe samples in \( X \). \( x_i = (x_{i1}, x_{i2}, ..., x_{id}) \) describes all attributes of \( i \)th sample.

We usually use \( d(x_i, x_j) \) as distance to describe the similarity between sample \( x_i \) and \( x_j \), the greater the distance, the greater the difference. According to Euclidean distance formula, we can calculate the similarity:

\[
d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}
\]  

3.2 Error sum of squares criterion and function clustering criterion

In this paper, we use error sum of squares function as clustering performance criteria function. Assuming data set \( X \) includes \( k \) cluster subsets \( X_1, X_2, ..., X_k \) and the sample number of this data set are \( n_1, n_2, ..., n_k \); the center of clustering subsets are \( m_1, m_2, ..., m_k \), the formula of error sum of squares function can be presented as:

\[
E = \sum_{i=1}^{k} \sum_{p \in X_i} ||p - m_i||^2
\]  

Cluster average value of similarity

There are four main steps:
1. Allocate all of the objects to $k$ clusters which are not empty.
2. Calculate the average of each cluster and use that average to represent the cluster.
3. Allocate each object according to the distance between the object and the cluster center.
4. Turn to step 2, recalculate each cluster’s average value. Iterate this process until it satisfies some criteria function.

In the process of clustering, assume that the location data meets the characteristics of the checkerboard data set. 486 positive data is used, and the data is transformed to $[-1, 1]$. The distribution is shown in figure 1.

![Figure 1. initial distribution of the data](image1)

![Figure 2. clustering distribution of data](image2)

Through the iterative process, the location data set is divided into different categories, which makes the estimation of the location information of the clustering performance to be optimal. The results are shown in figure 2.

4. Path reasoning and decision making based on neural network

Each unit in artificial neural network has the extensive connection, parallel-distributed information storage and processing ability, adaptive learning ability, etc. Thus, it can provide better fault tolerance, strong adaptability and fusion processing ability for the path reasoning and decision making of indoor positioning [7]. We will design sparse data input layer, dynamic feedback implicit layer and
low-dimensional result output layer for the path reasoning and decision making based on neural network.

4.1 Data input layer

Indoor positioning data are mass and redundant. Through clustering algorithm, the data category will be simplified. But the number of same type of data is still large and its correlation with path planning is still low, which makes it difficult to process fast.

Therefore, at first, we should input the clustered data concurrently. Scale transformation and pretreatment are required after the most important input data are determined. The input vector of \( n \) dimensions are defined as \( M \). After the scale transformation, data will be transformed into \( \frac{M}{|\sum M|} \). We must check if there are any exceptions before scaling.

Then the data transformed will be controlled at input layer. Any neuron of a certain layer \( j \)'s input is \( \text{net}_j(\frac{M}{|\sum M|}) \), output is \( y_j \); The output of any neuron in an adjacent lower layer \( i \) is \( y_i \). At this time, the interface between the input layer and the implicit layer can be obtained as follows:

\[
\text{net}_j(\frac{M}{|\sum M|}) = \sum_i w_{ij} y_i \tag{3}
\]

\[
y_i = f(\text{net}_j) \tag{4}
\]

\( w_{ij} \) refers to the weight of the neuron input unit. \( f(\)\ refer to the output function of the neuron network. The sparse data input layer is shown in figure 3.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{sparse_data_input_layer.png}
\caption{Sparse data input layer}
\end{figure}

4.2 Dynamic feedback implicit layer

Learning is the most important characteristics of neural network. Sample input and output mode of the same training set are acted on the network repeatedly [8]. The network automatically adjusts the connection strength or topology structure between neurons according to certain training rules, so that the actual output meets the expected requirements or tends to be stable.

We can set threshold function as below:

\[
y = \text{sgn}(\sum_{i=1}^{n} w_{ij} y_i - \theta) \tag{5}
\]

\( \theta = -w_{n+1} \). The formula (5) can be further written as the dot product relationship, \( y = \text{sgn}(W^T X) \), \( \eta \) is the learning factor and represents the proportionality constant of learning speed. Assume that \( y_i(t) \) and \( y_j(t) \) represent the state (output) of ith and jth neuron. We can get the formula:

\[
w_{ij}(t + 1) = w_{ij}(t) + \eta[y_j(t)x_j(t)] \tag{6}
\]

The learning process can be designed as below:

(1) Select a set of initial weights \( w_{ij}(1) \)

(2) Calculate the error between the actual output and the expected output result.

(3) Refresh the weights. The threshold value can be considered as a input constant (-1);
(4) Return to step (2), until all output of the network under all training mode satisfy the requirement.

4.3 Low dimensional result output layer

Through the dynamic feedback implicit layer, we can output the post-processing data into the neural network. Output unit sums up all input values, and the output function of threshold model produces a set of output modes.

In this paper, we used Matlab 7.0 to build and test the experiment environment. Neural Networks Toolbox for Matlab was mainly used.

(1) Select the number of neurons: There is no direct expression control of the number of hidden layer neurons. When the number of hidden layer is 20 x 10 and 8 x 4, the training result of the navigation path is shown in figure 4. When there are too few neurons in the hidden layer, it is not possible to deduce the correct navigation path from the location data; when the number of hidden layer neurons is too much [9], the location data study time will be too long, and it can hardly identify the samples that were not directly received before.

![Figure 4. Training error curve of the same hidden layer number of different nodes](image)

(2) Because the learning process of neural network uses gradient descent method, we use the first derivative information of the error to guide the direction of next step to get the minimum error. In order to guarantee the convergence of the navigation path algorithm, learning rate $\eta$ has to be less than a certain limit. We usually set $0 < \eta < 1$. The closer it gets to the minimum, the convergence of the algorithm gets slower as the gradient change is going to zero. Accuracy of path prediction under different learning factors is shown in figure 5.
Figure 5. Accuracy of path prediction under different learning factors

5. Conclusion
In this article, through indoor positioning data clustering algorithm and intelligent decision-making research study. We designed massive indoor location data clustering process and path reasoning and decision making based on neural network. The problem of insufficient accuracy and data redundancy in indoor positioning is solved effectively. The intelligence of navigation path planning was improved.

Reference
[1] Hatamlou. Black hole: A new heuristic optimization approach for data clustering. Information Sciences, 2013, 222(3):175-184.
[2] W. Lin, Z. Deng. Dimensional functional differential convergence for Cramer-Rao lower bound. Journal of difference equations and applications. 2017, 23(1-2):249-257.
[3] Niknam T, Fard E T, Pourjafarian N, et al. An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering. Engineering Applications of Artificial Intelligence, 2011, 24(2):306-317.
[4] Yan X, Zhu Y, Zou W, et al. A new approach for data clustering using hybrid artificial bee colony algorithm. Neurocomputing, 2012, 97(1):241-250.
[5] Haykin S S. Kalman Filtering and Neural Networks. Adaptive & Learning Systems for Signal Processing Communications & Control, 2014:170 - 174.
[6] Schmidhuber J. Deep Learning in neural networks: An overview.[J]. Neural Networks the Official Journal of the International Neural Network Society, 2014, 61:85–117.
[7] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors. Computer Science, 2012, 3(4): 212-223.
[8] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks. 2013, 38(2003):6645-6649.
[9] Lin W, Deng Z, Fang Q, et al. A new satellite communication bandwidth allocation combined services model and network performance optimization. International Journal of Satellite Communications & Networking, 2016,35(3): 263-277