WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 5, Number 2, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v5n2.13
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Research on Analysis and Recognition of Car User Portrait based on Big Data of Vehicle Network

Jinxi Pang [1, 2], Jin Lu [1, 2, *]

[1] China Automotive Technology and Research Center Co., Ltd. Tianjin, China
[2] China Auto Information Technology (Tianjin) Co., Ltd. Tianjin, China
* Corresponding Author: Jin Lu (Email: m13502037247_3@163.com)

ABSTRACT

With the rapid development of vehicle networking technology, big data of vehicle networking has become one of the core resources of intelligent transportation system. As the key technology of the application of big data in the networking of vehicles, the analysis and identification of user portraits play an important role in realizing the landing of user portraits and precise advertising, improving the efficiency of traffic management, ensuring driving safety and optimizing the user experience. This paper aims to explore the research on user portrait analysis and recognition based on vehicle network big data, and realize the accurate identification of user portrait through comprehensive analysis of vehicle driving data, user behavior characteristics and other information. First of all, the basic data fields of the big data of the Internet of vehicles are introduced, and then the data is structured, including data preprocessing, feature extraction, feature recombination, and finally the unsupervised model K-Means++ is used for model construction. Through experimental verification, the method proposed in this paper has achieved good results in the recognition and analysis of user portrait, and can realize the landing of user portrait and precise advertisement delivery scene, which has high practical application value.

KEYWORDS

Big Data of the Internet of Vehicles; User Portrait Recognition; Data Preprocessing; Feature Extraction; Model Construction.

1. INTRODUCTION

With the rapid development of the Internet of Things technology, the Internet of vehicles has become the core component of the intelligent transportation field, realizing the close connection between vehicles and the Internet. By gathering and analyzing massive information such as vehicle driving data and user behavior patterns, the Internet of vehicles has brought significant help to improve traffic management efficiency, ensure driving safety and optimize user experience. However, in the wide application scenario of big data in the Internet of vehicles, how to accurately identify car users and carry out precision marketing based on car user behavior data is still facing severe challenges. Traditional user identification methods mainly rely on physical characteristics or identity information provided by users, and these methods have certain limitations in practical applications. Therefore, it is of great significance to study the research methods of user portrait analysis and recognition based on the big data of vehicle network.

The research of user portrait analysis and recognition has been concerned by researchers in all walks of life, and has achieved good results. Researchers in the literature skillfully use the data resources of the Internet of vehicles that are closely related to users, and extract the characteristic information that

can represent the user's identity [1]. After feature reduction and careful selection, they used advanced similarity measurement techniques to accurately match these features with similarity. Finally, the system output the matching results with the highest similarity as the conclusion of user identification. The experimental results show that this method significantly improves the recognition accuracy. In addition, the method also shows good adaptability in other similar application scenarios of user identification, which proves its wide versatility and practicability. The authors of the literature introduce an innovative means of user identification, which focuses on analyzing the daily behavior traffic of users to extract features[2]. The method deeply dissects the user's behavior patterns in different applications and time periods, subdividing these behaviors into multiple dimensions. Through in-depth mining of the active users' behavior data set, using the discrete time data analysis technology, the user's feature matrix is constructed, and then a comprehensive user behavior feature database is established. Compared with other authentication methods, this technology not only significantly reduces the sampling cost, but also shows more prominent advantages in recognition effect. In the literature[3], in view of the continuous growth of the number of social network platforms and the number of netizens, a large amount of data information that can deeply reflect the characteristics of users is generated when users are engaged in activities on social platforms with diverse functions. Based on this background, this paper proposes an innovative user identity recognition algorithm -- multi-level network embedding algorithm based on friends' contributions. Through experimental verification on social network data sets, the results show that the algorithm has significant advantages in various performance indexes compared with the existing algorithms. Specifically, its F1 value increased by up to 11.1%, and even the smallest improvement reached 2.5%. In addition, the algorithm also shows obvious advantages in terms of time performance. It can be seen that the user identity identification has a good performance in the field of engineering, at the same time, based on the vehicle network big data for vehicle user identification and automotive precision marketing has great practical application value.

To sum up, this paper puts forward the method of "Analysis and identification research of vehicle user portrait based on vehicle network big data", which has important theoretical significance and practical application value. By analyzing vehicle driving data (such as speed, acceleration, steering angle, etc.) and user behavior characteristics (such as driving habits, common routes, parking places, etc.), a more accurate and safe identity recognition model can be built. This method can not only effectively improve the accuracy and reliability of car user identification, but also provide more intelligent solutions for application scenarios such as precision marketing, personalized service and car user experience for car companies under the premise of protecting user privacy. In addition, with the continuous progress of artificial intelligence and machine learning technology, the user identification method based on big data of the Internet of vehicles is expected to achieve a higher level of automation and wider application prospects in the future.

2. RESEARCH METHODS

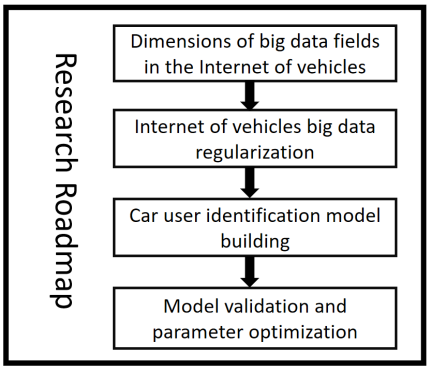

Fig 1. Research roadmap of this paper

By searching a large number of references and related reports to determine the research method of this paper, the first step to determine the field dimension of the big data of the Internet of vehicles; The second step is to regularize the vehicle data, including data preprocessing and feature extraction; The third step is to build the vehicle user identification model based on the unsupervised clustering model; The fourth step is model validation and discussion of the parameter optimization method. The research roadmap of this paper is shown in Figure 1 below.

2.1. Data Field Dimensions of Big Data of Internet of Vehicles

The field dimensions of the collected iov data are shown in Table 1 below, including simple personal information of the car owner, as well as the number and nature of apps used in the driving process.

Table 1. shows the field dimensions of the collected connected vehicle data

FIELDS	FIELDS	FIELDS
Sex	Work in an industry	Frequency of banking app use
Age	Whether you have children	Frequency of stock app use
Recent behavior	Car purchase purposes	Investment app usage frequency
Life stage	City of residence level	Frequency of use of bookkeeping app

As shown in Table 1, the field dimensions of the big data of the Internet of vehicles collected in this paper are 12, which basically cover the basic information of car users, as well as the time, frequency and other behaviors of using intelligent network connected devices at ordinary times. Based on these fields, they play an important role in realizing the landing of user portraits and precise advertising scenarios, improving traffic management efficiency, ensuring driving safety and optimizing user experience.

2.2. Network of Vehicles Big Data Regulation

Data normalization refers to the process of data preprocessing, feature extraction, feature combination, etc., on the original data to make it suitable for analysis and modeling, the purpose of which is to improve data quality and ensure consistency, accuracy and integrity in order to better support decision making and analysis.

2.2.1. Data Preprocessing

Data preprocessing is the first step in user identification, aiming to improve data quality and availability. According to the characteristics of the data of the Internet of vehicles, the pre-processing process mainly includes three steps: data cleaning, data integration and data transformation. These pre-processing steps lay a solid foundation for the subsequent feature extraction and model training. Data preprocessing mainly includes the following three steps.

Data cleaning: aims to remove duplicates, process missing values, outliers and noisy data, ensure data accuracy and consistency, and improve data quality.

Data integration: To unify and integrate data from different sources, eliminate redundancy and contradictions, and form a complete dataset. The data integration formula is shown in formula (1) below.

$$D = \begin{cases} D \cup d_i \ if \quad d_i \in d \\ D \varnothing d_i \ if \quad d_i \notin d \end{cases} \tag{1}$$

Data transformation: normalization, standardization and other processing of the data to improve the comparability and analyzability of the data, so that the data is more suitable for subsequent analysis. According to the type of data, this paper adopts max-min standardization, and max-min standardization is shown in formula (2).

$$x^* = \frac{x - \max}{\max - \min} \tag{2}$$

Where, max is the maximum value of the original data, min is the minimum value of the sample data, max-min is the range, deviation standardization retains the relationship existing in the original data.

2.2.2. Feature Extraction

Feature extraction is a process of extracting feature information that is significantly distinguishable for user portrait analysis and recognition from original data, aiming to build a representation model of user behavior through data-driven analysis. In the user identification research based on big data of the Internet of vehicles, feature extraction mainly covers the following three dimensions [4-5].

(1) Driving characteristics: Key features such as speed, acceleration, average acceleration, fuel consumption, rapid acceleration, rapid deceleration and energy consumption are extracted from vehicle driving data to reflect the user's driving habits, vehicle performance and driving stability index. For example, the driving smoothness index can be measured by the variance of acceleration. Smaller variance indicates a smoother ride. The formula for calculating the driving smoothness index is shown in formula (3).

$$P = \frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})^2 \tag{3}$$

Where, in formula (3) a_i is the acceleration at the ith time point, $\bar{a}$ is the average acceleration, and N is the total number of time points.

(2) Behavioral characteristics: By analyzing the user's driving behavior, the behavioral characteristics such as lane change frequency, braking frequency, braking strength, and the number of sudden braking are extracted to reflect the user's driving style and preference. The calculation formula of behavior characteristic style index is shown in the following formula (4).

$$T = \sum_{i=1}^{n} w_i \cdot f_i \tag{4}$$

Where, the i in formula (4) refers to the rank of the eigenmatrix of the behavior feature and f_i is the eigenvalue of the eigenmatrix.

(3) Spatiotemporal feature extraction: Combining the vehicle location and time information, the user's travel rules, frequent destinations, travel distance distribution, travel duration distribution, frequent destinations, travel rules and other spatiotemporal features are extracted to reflect the user's lifestyle and activity habits. For example, frequently-visited locations scoring can evaluate the importance of a certain location by comprehensively considering the visit frequency and stay time. The scoring formula for frequently-visited locations is shown in formula (5).

$$S_i = w_1 \cdot f_i + w_2 \cdot T_{dwell,i} \tag{5}$$

Where, in formula (5) S_i is the score of place i, f_i is the visit frequency of place i, w_1 and w_2 is the weight size, $T_{dwell,i}$ is the average stay time of place i.

Feature extraction is the core link of user identification research in big data of the Internet of Vehicles. By extracting driving characteristics, behavior characteristics and temporal and spatial characteristics, a representative model of user behavior can be constructed, so as to achieve accurate user identity recognition and behavior analysis. These features not only have a significant degree of differentiation, but also provide rich data support for Internet of vehicles applications.

2.2.3. Feature Combination

In section 2.2.2, driving features, behavior features and spatiotemporal features are extracted. The above three features are recombined in this section, and the combined feature vector is expressed as formula (6).

$$F = [P, T, S] \tag{6}$$

Where, P is the driving feature vector, including speed, acceleration, fuel consumption; T is the behavior feature vector, including lane change frequency, braking frequency, sudden braking times, etc.; S is the spatio-temporal feature vector, including frequented locations, travel time rules, travel distance distribution, etc.

2.3. Identification Model Construction

Model construction is the core step of user identification based on big data of vehicle network. In the process of model construction, it is necessary to choose the appropriate algorithm and model structure to realize the accurate identification of user identity. Commonly used model building methods include: In the case of independent user identity labels, K-means clustering, principal component analysis and other clustering, dimensionality reduction and other methods to discover the potential structure and pattern in the data [6]. According to the characteristics and laws of the data, the K-means ++ algorithm is adopted, and the K-Means++ algorithm solves the problem that the K-means algorithm is sensitive to the initialization of the cluster center [7]. The difference between the K-Means++ algorithm and the traditional K-means algorithm lies in the selection of the initial K center points. The K-Means algorithm uses a randomly given way. The steps of K-Means++ algorithm are shown in Figure 2 below.

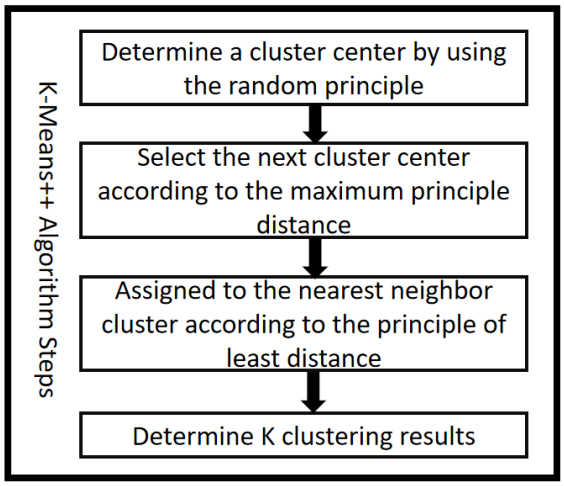

Fig 2. Schematic diagram of K-MEANS ++ algorithm steps

The K-Means++ algorithm consists of four steps. The first step is to determine a cluster center by using the random principle; The second is to select the next cluster center according to the maximum principle distance; The third step is to assigned to the nearest neighbor cluster according to the

principle of least distance; The fourth step is to determine K clustering results. The specific steps are shown below [6-8].

(1) Choose any node from the data set as the first clustering center μ_1;

(2) For each point x_i in the data set, calculate the distance x_i to all existing cluster centers D (x), $D(x) = \arg\min \sum_{r=1}^{k_{selected}} \left\| x_i - \mu_r \right\|_2^2$, that is, select the next cluster center as far away from the nearest cluster center as possible;

(3) To select a new data point as the cluster center, the selection principle is that the point with a larger D(x) is more likely to be selected as the cluster center;

(4) Repeat the process of (2) and (3) until K cluster centroids are selected;

(5) Calculate the distance of each object to each of these k centers and assign it to the nearest cluster according to the principle of minimum distance;

(6) The sample mean in each cluster is used as the new cluster center;

(7) Repeat steps (5) and (6) until the cluster center no longer changes;

(8) The cluster is over and you have k clusters.

2.4. Model Verification and Optimization

After the construction of the model, it is necessary to verify and optimize the model to ensure its accuracy and reliability in practical applications.

(1) Model verification

Model verification is a key step to ensure the generalization ability of the model. Common methods include feature selection transformation, cross-validation, regularization, etc. Through these methods, the model performance can be evaluated and optimized effectively.

Feature selection transformation: By evaluating the importance of each feature to vehicle user identification, key features are selected for model training to improve the accuracy and efficiency of the model.

Cross-validation: The data set is divided into a training set and a test set, and the performance of the model is evaluated through multiple cross-validations. Commonly used cross-validation methods include K-fold cross-validation, leave-one cross-validation, etc.

(2) Model optimization

Parameter tuning: Optimize the performance of the model by adjusting the parameter Settings of the model. Commonly used parameter tuning methods include grid search, random search, etc.

The core ideas of grid search are as follows: 1) Define hyperparameter space: it refers to specifying a set of candidate values for each hyperparameter. 2) Traversing all combinations: refers to training the model for each hyperparameter combination and evaluating the performance. 3) Choosing the best combination: It means choosing the hyperparameter combination that performs best on the verification set.

3. CONCLUSION AND PROSPECT

3.1. Conclusion of this Paper

This paper studies the identity identification of vehicle users based on the big data of vehicle network. Through the steps of data field analysis, data normalization (data preprocessing, feature extraction, feature normalization), K-Means++ model construction, verification and optimization, the accurate identification of vehicle users is realized, which can be effectively applied to automobile enterprises to better implement precision marketing. The experimental results show that the proposed method has high accuracy and efficiency. In addition, by analyzing the influence of each feature on user identity recognition, this paper also finds that driving feature and behavior feature have high value in user identity recognition.

3.2. The Prospect of this Paper

Although this paper has made some achievements in the analysis and recognition of user portrait based on big data of vehicle network, there are still some problems and challenges to be solved. Future research directions can include the following aspects:

The first is to optimize the feature extraction algorithm, improve the feature quality and availability, and provide more accurate feature information for user identification. The second is to combine advanced technologies such as deep learning and reinforcement learning to explore new model building methods to improve the accuracy and efficiency of identification. The third is to expand the application scenarios, and apply the user portrait analysis and recognition method to intelligent parking, vehicle tracking, insurance pricing and other fields to give full play to its application value.

This paper attaches great importance to privacy protection: in the process of collecting and processing big data of the Internet of vehicles, strict privacy protection measures have been taken to ensure that users' personal information and sensitive data are not leaked. At the same time, advanced technologies such as differential privacy and federal learning have been introduced to further ensure the security and privacy of user data.

REFERENCES

[1] Wang Lan. User identity recognition based on IPTV big data research [D]. South China university of technology, 2020. DOI: 10.27151/d.cnki.ghnlu.2020.004976.

[2] LI Shuangshuang. Research on Identity Identification Technology Based on Online Traffic Characteristics of Mobile Users [D]. Huazhong University of Science and Technology, 2015.

[3] Xu Neng. Across social network user identity recognition algorithm research [D]. Anhui architecture university, 2022. DOI: 10.27784/d.cnki.gahjz.2022.000373.

[4] TIAN Hengyi, WANG Yu, XIAO Hongbing. Automatic brain tumor segmentation algorithm based on multi-modal feature recombination and scale cross attention mechanism [J]. Chinese Journal of Lasers, 2024, 51 (21): 129-138.

[5] Zhang Fan, Guo Yaxin, Yang Jing, et al. Research on electric energy prediction based on GBDT+ feature engineering method [J]. Electronic Quality, 2020, (01): 1-4.

[6] FANG Xiao, YUAN Xiaofang, Guan Donglin, et al. Research on Abnormal User Behavior Detection based on K-means Algorithm [J]. Network Security Technology and Application, 2025, (02): 23-25.

[7] HAN Xiaocui, Hu Yewei, Wu Qingyan, et al. Abnormal data recognition and automatic processing system of personnel management based on K-means clustering algorithm [J]. Electronic Design Engineering, 2024, 32 (24): 27-31. DOI:10.14022/ j. ssn1674-6236.2024.24.006.

[8] Chen Hongtao. Correlation Analysis of Student Behavior Data and Academic Achievement based on K-means Algorithm [J]. China Science and Technology Information, 2024, (23): 86-88.]