International Journal on Semantic Web and Information Systems
Volume 18 • Issue 1

A Path-Clustering Driving Travel-Route Excavation

Can Yang, Hunan Institute of Traffic Engineering, China*

ABSTRACT

The refueling trajectory of self-driving tourists is sparse, and it is difficult to restore the real travel route. A sparse trajectory clustering algorithm is proposed based on semantic representation to mine popular self-driving travel routes. Different from the traditional trajectory clustering algorithm based on trajectory point matching, the semantic relationship between different trajectory points is researched in this algorithm, and the low-dimensional vector representation of the trajectory is learned. First, the neural network language model is used to learn the distributed vector representation of the fueling station; then, the average of all the station vectors in each trajectory is taken as the vector representation of the trajectory. Finally, the classic k-means algorithm is used to cluster the trajectory vectors. The final visualization results show that the proposed algorithm effectively mines two popular self-driving travel routes.

KEYWORDS

Distributed Representation, Self-Driving Tour, Semantic Model, Sparse Trajectory, Tour Route Mining, Trajectory Clustering

INTRODUCTION

Freedom is the "soul" of self-driving travel. For tourists, the self-driving tour itself is for the convenience of travel and can meet their free needs. But for the city, the test is whether the product is sticky. Therefore, the development of self-driving tours must be guided by consumer demand. Retaining self-driving tourists requires not only the transformation and upgrading of scenic spots, but also the overall image of a self-driving destination to attract car owners, and concentrate more product formats for tourists to consume and enjoy, so as to improve tourist satisfaction.

Self-driving tours are organized and planned with self-driving cars as the main means of travel. The rise of self-driving tours conforms to the psychology of the younger generation. They are unwilling to be restrained and pursue personality, the independence and freedom of mind, self-driving tour just fills this demand.

Self-driving tour is a type of self-guided tour, it is a new type of tour that is different from the traditional group tour. Self-driving travel provides tourists with flexible space in terms of object selection, participation procedures, and experience freedom. Self-driving tours, with their inherent characteristics of freedom and individuality, flexibility and comfort, choice and seasonality, are

DOI: 10.4018/IJSWIS.306750 *Corresponding Author

radically different from traditional participation. Compared with the collective way, it has its own characteristics and charm.

The literal meaning of self-driving is that the driver is himself. Vehicles include cars, mainly cars, off-road vehicles, RVs (Recreational Vehicle), motorcycles and bicycles. They are mainly privately owned, these can also be borrowed, leased and other methods. The driving purpose is with diversity and arbitrariness, the final decision lies with the car owner or travel team. It can be seen that tourism is one of self-driving activities. When self-driving is used as a means of travel, there will be the following changes. The driver can be the owner or his companion. The main purpose of driving is leisure travel, not for work, transportation and other reasons. Self-driving tours are private tours and are not public tours.

With the improvement of the national economy, the number of private cars has increased rapidly, self-driving travel has gradually become a popular choice for people to travel. By analyzing the self-driving trajectory of tourists, it is possible to discover popular self-driving travel routes and provide support for travelers' travel route planning. However, the activities of self-driving tourists have high autonomy, it is difficult to collect travel trajectory data, the representativeness and coverage of the data are insufficient. Regarding these, different scholars have done research. A seasonality measurement framework is constructed from pattern and intensity, the methods such as single-index panel data cluster analysis and linear programming are used to explore the seasonal temporal and spatial characteristics of China's self-driving travel (cross-city) market *(Li X., 2021)*. A self-driving travel plan design is proposed based on distribution uniformity adaptive ant colony algorithm *(Song M J, & Wu Y H.,2021)*. Ant colony algorithm is a probabilistic simulated evolutionary algorithm, it is used to find the optimal path in the graph, the distribution uniformity of the solution is introduced in the optimization process, the information update strategy is dynamically adjusted and the path probability is selected, it can accelerate the convergence while avoiding precociousness, and get a more reasonable self-driving tour route. This provides a reasonable solution for the self-driving tour route planning. The self-driving tour in the Silk Road is taked as the research topic, the ROST CM6 software is used to analyze a total of 12,962 pieces of information of online travel notes extracted from Ctrip. Based on this, an ASEB strategy matrix is formed for the experience of self-driving tourists on the Silk Road *(Lei Li X Z,* et al.,*2021)*. ASEB is an acronym for activity, setting, experience, and benefit. A novel rough-fuzzy best-worst method is proposed to prioritize the identified requirements, simultaneously manipulating the intrapersonal and interpersonal uncertainties (*Chen Z H, et al., 2020*). The case study results of smart vehicle service system show that sixteen smart service requirements are identified in the self-driving tour with the smart vehicle, and the requirement "alerting the driver's unsafe behavior using informative diagnostic capability" emerges as the most important one in the proposed rough-fuzzy best-worst method.

Smart tourism recommendation is a relatively complex system. In the process of self-guided tours, tourists' exploration and curiosity are the strongest motivation for travel. Tourists pursue an experience that combines mind and body during travel. A recommended item has the greatest marginal relevance if and only if it has a high degree of relevance to the topic of the item, and at the same time contains the least redundant information in the selected recommendation list, so as to ensure the relevance to the topic or user needs and reduce redundant information, the distinctive content is added to ensure the novelty and high quality of recommendations. This paper mines popular autonomous driving travel routes from self-driving travel refueling information based on a semantic sparse trajectory clustering algorithm.

In this paper, a refueling dataset covering Xinjiang is used for exploring popular self-driving tour routes in Xinjiang. This data set records all the user's refueling behavior in Xinjiang, it also contains the refueling records of all self-driving tourists. A refueling trajectory can be obtained by forming a sequence of refueling stations according to the time sequence. The refueling trajectory is a sample of tourist travel routes, it can truly reflect the time and space movement trajectory of tourists, it can be used as an important data source for Xinjiang self-driving travel route mining. However,

the fueling data is used to mine popular self-driving travel routes, it faces two main challenges. First, there are a large number of people in the original data, the refueling behavior is complex and diverse, it is difficult to accurately identify the self-driving tourist group. Second, compared with the Global Positioning System (GPS) trajectory data, the frequency of refueling records is very low and the data is very sparse, this results in an uncertain path between two consecutive trajectory points in the same refueling trajectory, it is very difficult to restore a specific route from it.

There are such a group of people who dare to challenge themselves and enjoy the joy of conquering off-road. For them, off-roading is a way of life. Life is always full of challenges to human potential, and it always needs to be explored and tapped. Sometimes ordinary life can blossom to the extreme. You like challenges, adventures, and pursuit of excitement. Most importantly, you like to drive yourself to find your own preferences. The purpose of this article is also to run towards this goal, to provide its convenience.

LITERATURE REVIEW

Tourism Route Mining

In recent years, there has been a lot of research on tourism route planning and recommendation. The data in the current research is mainly GPS track data which is shared by users, geo-tagged photo data and sign-in data *(Chang L, et al.,2019)*. Based on GPS trajectory data, a series of tourism route mining and recommendation work have been carried out, and excellent results have been achieved *(Zheng Y, et al.,2011; Zheng Y, et al.,2009; Zheng V W, et al.,2010)*. Based on the user's GPS trajectory information, the user's personalized information is researched, two kinds of travel route recommendation algorithms are proposed by using collaborative filtering technology, the degree of personalization of the recommendation results is improved *(Cui G, et al.,2018)*. However, although GPS trajectories can reflect the specific travel routes of tourists, they are relatively difficult to be obtained. With the development of location services, the use of social media data has become a new research hotspot for tourism route mining and recommendation. A collaborative filtering method is proposed based on topic model, travel photos are used for travel recommendation *(Jiang S, et al.,2015)*. Based on tourist check-in data, a route reasoning framework of the collective knowledge is proposed *(Wei L Y, et al.,2012)*, popular tourist routes are mined from uncertain trajectories. At present, there are few researches on self-driving travel in using data of these shared users. The main reason is that these data are difficult to distinguish between ordinary tourists and self-driving tourists. There are very few data that can be completely determined as self-driving tourists. Based on the 924 GPS trajectories which are shared by self-driving tourists, they are combined with road network information and tourist attraction information, the temporal and spatial behavior characteristics of self-driving tourists are analyzed by methods such as seasonal intensity index, multi-dimensional buffer zone, and core area density *(Liu Y P, et al.,2019)*.

Most of the above-mentioned data comes from users' sharing, these users only account for a small part of all tourists, the analysis results have large deviations. Different from the used data in the above research, fueling trajectory data is used to mine self-driving travel routes in this article. The data set itself contains the fueling behavior of all self-driving tourists in Xinjiang of China, it plays an important role in analyzing the overall situation of self-driving travel in Xinjiang.

Trajectory Clustering

In order to find representative paths or common trends in the trajectories of different moving objects, similar trajectories are usually clustered *(Zheng Y.,2015)*. In traditional trajectory clustering methods, a certain measurement method is generally used to compare the similarity between trajectories, and then some classic clustering algorithms are used for clustering. A trajectory clustering framework is proposed, each trajectory is divided into multiple sub-trajectories, and then the density clustering

method is used to cluster these sub-trajectories *(LEE J G, et al.,2007)*. A travel behavior clustering algorithm is proposed, sampling density clustering algorithm is used to solve the noise problem in trajectory data *(Tang W, et al.,2016)*. A new distance measurement method is defined, distance-based trajectory clustering is realized *(Besse P C,et al.,2016)*.

The core of spatiotemporal trajectory clustering is to measure the similarity between trajectories. Common used trajectory similarity measurement methods include Dynamic Time Warping (DTW) *(Yi B K, Jagadish H V, Faloutsos C.,1998)* and Longest Common Subsequence (LCSS) (Vlachos M, et al.,2002) and Edit Distance on Real sequence (EDR) *(Chen L, Özsu M T, Oria V.,2005)*.

The above-mentioned measurement method mainly considers the spatial position information of the track points, and it is suitable for GPS data with high sampling frequency. However, the refueling frequency of tourists is very low, it is usually in 2~3 days, the refueling track is very sparse, and the distance between two consecutive visited refueling stations is even hundreds of kilometers. Therefore, the above method is not suitable for extremely sparse fueling trajectory data.

Distributed Representation

In the field of natural language processing, traditional methods of representing words as high-dimensional sparse vectors have been largely replaced by neural network-based language models. The neural network language model is trained by the word order and the co-occurrence of words. The concept is based on the distributed assumption, the words often appear together in sentences, they have higher statistical relevance. Word2vec was proposed by Mikolov, it is a prominent representative *(Mikolov T, et al.,2013)*. The low-dimensional vector representation of words can be learned simply and efficiently, excellent performance has been achieved on traditional natural language processing tasks including machine translation and sentiment analysis.

Recently, the concept of distributed representation has gradually expanded to other fields such as web search, e-commerce, and recommendation systems. Researchers realized that user behavior sequences could betreated as sentences, and then the representations are embedded in products or users, such as user clicks, queries or purchase sequences. Distributed representation is used in various types of online recommendations, including Taobao recommendation *(WANG J, et al.,2018)*, job search recommendation *(Kenthapadi K, LE B, & Venkataraman G.,2017)*, application recommendation *(Radosavljevic V, et al.,2016)*, housing recommendation *(Grbovic M, & Cheng H.,2018)*. Similarly, a similar method has also been proposed for social network analysis, the random walk sequence of nodes in the network is used to learn the embedding representation of network nodes *(Perozzi B, Al-Rfou R, & Skiena S..2014)*. This idea is used in this article, the gas station in the refueling trajectory is regarded as a word, and the entire trajectory is used as a sentence, and word2vec is used to learn the semantic vector representation of the gas station, it is used to cluster the refueling trajectory and restore the travel route of tourists.

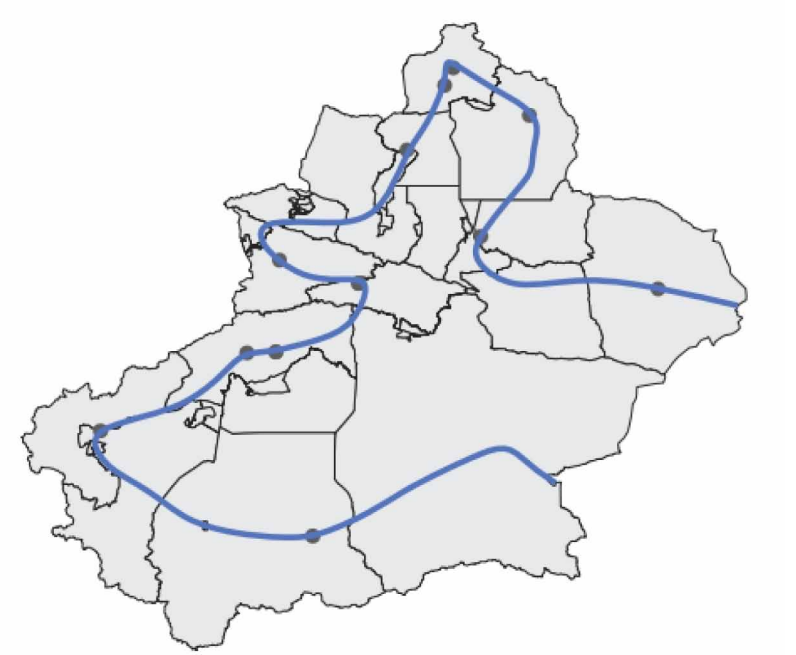
METHODOLOGY

Research Content and Technology Architecture

As is shown in Figure 1, the blue lines in the figure are the exact trajectories of a tourist on a self-driving tour, and the black dots indicate the actual location of the tourist to refuel. It can be seen that only relying on the sparse refueling trajectory point data of a single tourist is not enough to infer the actual travel route of the tourist.

The main work of this paper is to address the problem of tourist group identification. By analyzing the known refueling behaviors of tourists, the basic characteristics of tourists refueling are summarized, then tourist groups can be identified from a large number of original refueling records. For the problem of sparse trajectory points, it is inspired by word2vec, a sparse trajectory clustering algorithm is proposed based on semantic representation *(Mikolov T, et al., 2013)*. Each gas station

Figure 1. Comparison between tour route and refueling trajectory of a tourist

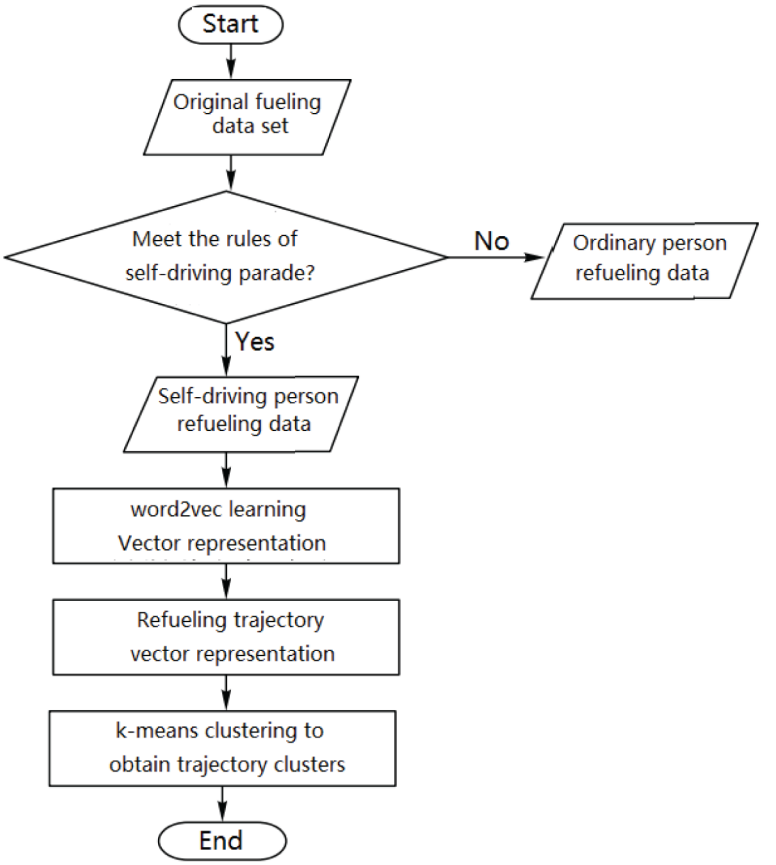


is regard as a word, and each gas trajectory is regard as a sentence, the distributed representation of the gas station is learned by word2vec, then the average of the station vectors in each trajectory is used to represent the gas trajectory. Finally, k-means algorithm is applied to complete trajectory clustering, popular self-driving travel routes are mined according to the clustering results. In Figure 2, the overall flow of the method is showed in this paper.

K-means algorithm is a clustering algorithm, it belongs to a kind of unsupervised learning algorithm, data can automatically be classified according to the distribution of data. Application scenarios include Market Segmentation, segmentation of server clusters, segmentation of social network clusters, etc. The idea of the algorithm is simple and easy to understand, and it is easier to implement. The advantages of the k-means clustering algorithm are (i) there are simple idea and fast convergence speed in the algorithm; (ii) when performing cluster analysis on large-scale data sets, the algorithm clustering is more efficient, the clustering effect is better; (iii) When the structural distribution of the data set is spherical or quasi-spherical or other convex structures, the clustered structures can efficiently be found in the k-means algorithm; (iv) The algorithm has good clustering effect on numerical data, and the clustering result has nothing to do with the input order of the data.

Different K values are initialized, and the final result of the algorithm will be different. The choice of the k value may cause the algorithm to converge to a local extremum. One solution is to do multiple random initializations and run the algorithm, and then select the optimal solution. In practical applications, if the number of K is relatively small, such as between 2 and 10, then this method will probably work. However, if the number of K is too large, such as several hundreds, the result of this method is likely to be unsatisfactory. There is a method for selecting the number of K values such as elbow method. According to the K value from small to large, the algorithm is run in turn, the size of the loss function is compared, and then a selection is made. k-means is to minimize the squared error between the sample and the mass point, the squared error is as the objective function, the sum of the squared distance error between the mass point of each cluster and the sample point

Figure 2. Overall flow of the proposed algorithm



in the cluster is called distortion. For a cluster, the lower the degree of distortion value, the tighter the members in the cluster, the higher the degree of distortion, the looser the structure in the cluster. The degree of distortion will decrease with the increase of the category, but for data with a certain degree of discrimination, the degree of distortion will be greatly improved when it reaches a certain critical point, then it will decrease slowly. This critical point can be considered as a point with good clustering performance, and its image is like an elbow, so it is named the elbow method.

Self-Driving Tourist Group Recognition

The record $r = \{u, v, s, t, area\}$ in the original fueling data set, where: u, v, s, t represent the driver, vehicle, gas station and gas timestamp respectively; *area* represents the administrative division to which the gas station belongs. The data is selected to construct the personnel refueling trajectory data set, the selected time range is from January 1, 2016 to December 31, 2018. The refueling records are reorganized according to the personnel, the refueling records of each user can form a refueling track $tra = \{s_1, s_2, ..., s_n\}$, it is a collection of refueling sites which are visited by the user in chronological order.

After investigation, it can be found that self-driving tourists usually have three characteristics: (i) They have only been to Xinjiang once and stay for no more than 30 days; (ii) They have continuous refueling behavior, the interval between two adjacent refuelings is less than 5 days; (iii) Refueling location is not fixed and scattered in various prefectures and cities, it is usually no less than 3

prefectures and cities. At the same time, in order to avoid sparse trajectories. In this paper, only trajectories with more than 8 refueling times are analyzed. Based on the above characteristics, four rules are defined to identify groups of self-driving tourists. For any trajectory $tra = \{s_1, s_2, ..., s_n\}$ in the refueling trajectory data set, if formula (1) is satisfied, the trajectory *tra* is considered to be a self-driving refueling trajectory, the corresponding person is a self-driving tourist:

$$\begin{cases} t\left(s_n\right) - t\left(s_1\right) \le 30 \\ t\left(s_j\right) - t\left(s_{j-1}\right) < 5 \\ different\left(area\right) \ge 3 \\ n > 8 \end{cases}, \ 1 < j \le n \tag{1}$$

According to the above rules, 20646 fueling trajectories are extracted from the fueling trajectory data set. In order to verify that these trajectories are actually generated by self-driving travelers, a statistical analysis of the basic characteristics is made in this paper for the corresponding trajectories. First of all, the male to female ratio of these people is as high as 21:1, it is much higher than the 3.5:1 male to female ratio in the original data set. This is in line with the fact that the majority of drivers in long-distance self-driving tours are men. Secondly, the main source provinces of tourists are also very different from the overall situation, as shown in Figure 3.

In Figure 3, the provinces from which tourists come from vary considerably. The proportion of the population in the more developed provinces has increased significantly. The number of tourists from Beijing is the largest, while the number of tourists from Gansu is very small, which reflects the important influence of residents' income and consumption level on self-driving travel. The monthly distribution of self-driving tourists is shown in Figure 4. The results show that the number of tourists is very obviously affected by seasonal changes, it is mainly concentrated in July, August, and September. At the same time, almost no tourists choose to take self-driving tours in winter. This is closely related to Xinjiang's climatic conditions, it is in line with the basic conditions of Xinjiang's tourism market.

It can be seen that the defined rules in this article effectively identify the self-driving tourist group and can support further research work.

Semantic Representation of Fueling Trajectory Clustering

Problem Definition and Analysis

Definition 1 - Self-driving route: The actual self-driving route of tourists is a continuous spatial curve, it is the exact travel path of tourists by self-driving.

Figure 3. Provincial distribution of migrants and self-driving tourists

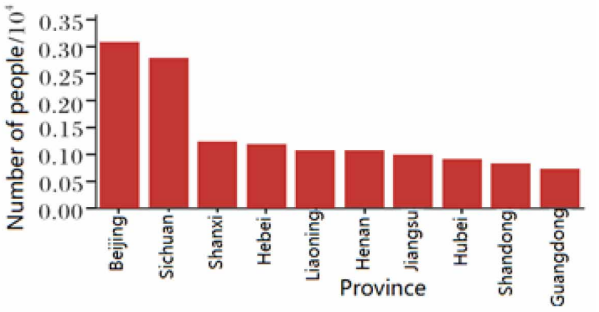
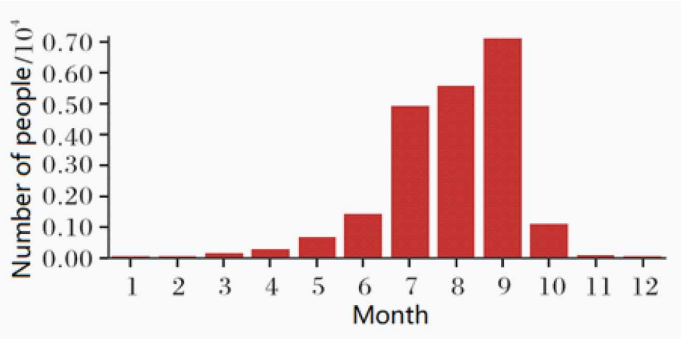


Figure 4. Histogram of self-driving tourist number variation with time



Definition 2 - Refueling trajectory *tra*: Each refueling trajectory is a sequence of stations for refueling tourists in chronological order, it is also a sampling of self-driving routes, $tra = \{s_1, s_2, ..., s_n\}$ is formally expressed. The vector representations corresponding to the refueling trajectory *tra* is ***traj***, the gas station *s* are and ***st***.

Definition 3 - Refueling trajectory clustering: A tourist refueling trajectory set T is given, the goal is to divide the trajectory set T into k disjoint clusters $C = \{C_1, C_2,..., C_k\}$, the refueling trajectory of the same trajectory cluster corresponds to the same self-driving route, $T = \{ tra_1, tra_2, ..., tra_j\}$.

There will be a large number of gas stations on the same self-driving route. Tourists have great autonomy in their refueling behavior. Each refueling trajectory can be regarded as a random sampling of all gas stations on the route. Therefore, although the tourist routes are the same, the final refueling trajectory is still very different. These trajectories all reflect part of the information of the tourist route, clustering these trajectories can get the complete tourist route information. However, the high sparseness of the fueling trajectory makes the traditional trajectory clustering algorithm based on the spatial similarity of the trajectory points unable to obtain the desired effect. Therefore, in this paper, the concept of semantic similarity in natural language processing is introduced into the fueling trajectory clustering, the similarity between gas stations is better measured.

In the field of natural language processing, words with similar semantics have similar contexts. Similarly, in the refueling trajectory, the context of the gas station is usually the gas station on the same route, these stations have higher semantic similarity. Word2vec is used to learn the semantic information of the site, the vector representation of the site is obtained, then the average of all site vectors is taken as the vector of the refueling trajectory, and finally, the classic k-means algorithm is used to achieve trajectory clustering. The overall framework of the algorithm in this paper is described as follows.

Algorithm 1: Refueling trajectory clustering based on semantic representation.

Input: Refueling trajectory data set T, Clustering number of clusters k, Skip-gram model window size m, Embedding vector dimension d. The main purpose of the skip gram model in Word2Vec is to avoid general shallow introductions and abstract ideas, but to explore Word2Vec in more detail.
Output: Cluster division $C = \{C_1, C_2,..., C_k\}$.
Initialize the weight matrix $\boldsymbol{W}$ and $\boldsymbol{W'}$
for $tra \in T$ do

```
   SkipGram(W, W', m, d, tra)  //word2vec model training
end for
//Each row in the weight matrix W corresponds to a vector
representation of a site
for tra∈T do
```

$$traj = \frac{1}{n}\sum_{k=1}^{n} st_k$$

```
//The average value of the site vector in the trajectory is
represented as the vector of the trajectory
end for
D = {traj_1, traj_2,..., traj_f}
C = Kmeans(D,k) //k-means algorithm gets the trajectory clustering
result
return C
```

Site Vector Representation

Word2vec is used for site representation. The fueling trajectory data set T is given, the goal is to learn the d-dimensional vector representation of each station. In fact, word2vec includes two models, namely Continuous Bag-of-Words model (CBOW) and skip-gram model. The goal of CBOW is to predict the probability of the central word based on the context, while the skip-gram model is the opposite. Generally speaking, the effect of skip-gram model will be better. In this article, skip-gram model is selected to learn the vector representation of the site.

Although skip-gram is an unsupervised method, it still defines an auxiliary prediction task internally. As shown in Figure 5, the center word s_i is selected first, the context is selected within m word distances before and after it, the training word pair (s_i, s_{i-m}), ..., (s_i, s_{i-1}), (s_i, s_{i+1}),..., (s_i, s_{i+m}) are formed. During training, a sliding window containing the current central word and its context is used to move in the corpus, all training samples are obtained. The purpose is to use the central word to predict the probability of the context. The goal of constructing this supervised learning is not to solve the supervised learning problem itself, but this problem is used to learn a good word embedding model.

The structure of the skip-gram model is shown in Figure 6. It is a simple neural network model with only three layers of input layer, hidden layer and output layer.

Both the input layer and the output layer are represented by an *N*-dimensional *one-hot* encoding vector, where *N* represents the total number of gas stations, and the hidden layer is represented by a vector of dimension *d*. The weight matrices $\boldsymbol{W}_{N\times d}$ and $\boldsymbol{W'}_{d\times N}$ are located between the input layer and the hidden layer and between the hidden layer and the output layer, respectively. The hidden layer does not use any activation function, and the output layer uses *softmax* as the activation function. The specific training process of the model is as follows:

Figure 5. Skip-gram model-based station embedding representation

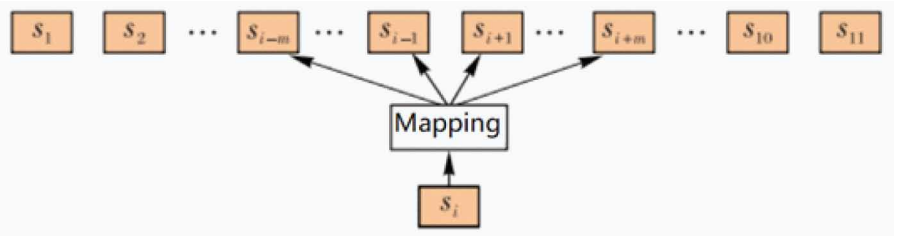
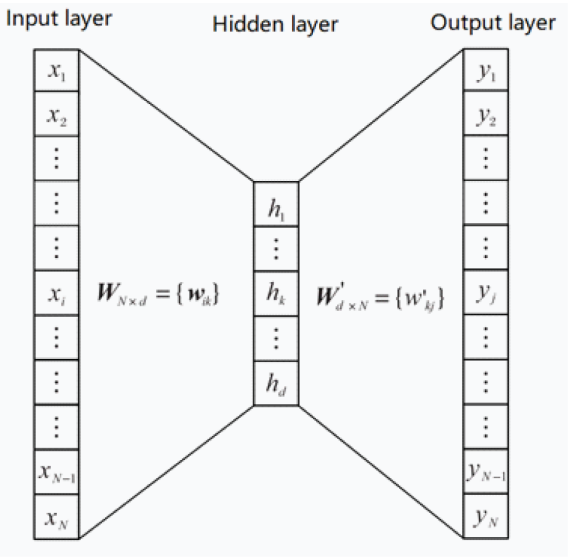


Figure 6. Skip-gram model structure



1. Initialize the weight matrices ***W*** and ***W'*** randomly.
2. Predict the vector ŷ of the target word, which is calculated as formula (2):

$$\hat{y}_j = P\left(s_j \mid s_i\right) = \frac{\exp\left(u_j\right)}{\sum_{n=1}^{N}\exp\left(u_n\right)},\ h = W^T x, u = W^{'T} h \tag{2}$$

3. The back propagation algorithm and stochastic gradient descent are used to update the weight matrices ***W*** and ***W'*** to minimize the loss function. The loss function on the entire training sample set is equation (3):

$$L = -\sum_{tra\in T}\sum_{s_i\in tra}\left(\sum_{-m\le j\le m, i\ne 0}\log(s_{i+j}\mid s_i)\right) \tag{3}$$

4. Each row of the weight matrix ***W*** is taken as the vector representation of the site.

Trajectory Clustering

After the vector representation of the site is obtained, the site vector can be used to obtain the vector representation of the trajectory. In natural language processing, a simple and effective method of sentence vector representation is to average all word vectors in the sentence *(Perozzi B, Al-Rfou R, & Skiena S.,2014)*. Similarly, for the trajectory *tra*, its vector representation ***traj*** is the average value of all gas stations s_1, s_2, ..., s_n vectors it contains, the formula is expressed as equation (4):

$$traj = \frac{1}{n}\sum_{k=1}^{n} st_k \tag{4}$$

The vector representation of each trajectory can be obtained, and then the trajectory clustering can be performed by using the classic clustering algorithm. In this paper, the *k*-means algorithm is used for trajectory clustering. The sample set *D* and the number of clusters *k* are given, the *k*-means algorithm is used to divide the sample set into k different clusters C_1, C_2,..., C_k, the minimized square error is equation (5):

$$E = \sum_{i=1}^{n} \sum_{traj \in C_i} \left\| traj - u_i \right\|_2^2 \tag{5}$$

wherein:

$$u_i = \frac{1}{\left| C_i \right|} \sum_{traj \in C_i} traj$$

is the mean vector of the cluster C_i. The specific process of the *k*-means algorithm is as follows.

Step 1: Randomly select k samples from the sample set D as the mean vector of the initial clusters.
Step 2: Calculate the distance between each sample and the mean vector of each cluster, and select the cluster with the closest distance as the cluster label of the sample.
Step 3: According to the newly divided clusters, recalculate the cluster mean vector.
Step 4: If the cluster mean vector does not change, the algorithm terminates and returns to the cluster; otherwise, repeat steps 2~3.

RESULTS AND DISCUSSION

Experimental Configuration

In order to mine Xinjiang self-driving travel routes, a refueling data set covering the entire region of Xinjiang was used as the original data set. First, the original refueling data set was preprocessed, the refueling trajectories of 20646 self-driving tourists were excavated as an experimental data set, involving 1856 gas stations in Xinjiang. The experimental machine system is Ubuntu 18.04, the CPU model is Intel Core i7-3770 CPU @ 3.4 GHz, the memory is 12 GB, and the Python version 3.6.

The main parameters of this paper include the window size m, the site vector dimension d and the number of clusters k. Wherein, m and d are set as default values, which are 5 and 100 respectively. The self-driving tour routes in Xinjiang are mainly divided into the Northern Xinjiang Line and the North-South Xinjiang Great Circle Line, so the value of k is set to 2.

Site Distributed Representation Results Analysis

In order to verify whether the algorithm has learned an effective site embedding representation, in this paper, cosine similarity is used to measure the similarity between site vectors and observe how similar sites are related. Kanas gas station in Altay area is selected as the target, which is located near the famous scenic spot Kanas, it is suitable for analyzing tourists' refueling behavior. Table 1 shows the 10 stations most similar to Kanas gas station. No. 1, No. 2, No. 4, No. 6 and No. 8 stations are also located near the Kanas Scenic Area, this indicats that the vectors of the stations have effectively learned the spatial location characteristics of the stations. In addition, it was also found that 10 stations are located near famous tourist attractions, the No. 3, No. 5 and No. 10 gas stations are located near the Baisha Lake Scenic Area, the No. 7 and No. 9 gas stations are located near the Nalati Scenic

Table 1. Ten most similar stations to Kanas gas station

Gas station No.	Gas station name	Cosine similarity
1	Altay Burqin Gas Station in Eastern Suburbs	0. 936 8
2	Altay Beitun Northwest Road Gas Station	0. 847 6
3	Altay Haba River City West Gas Station	0. 819 1
4	Altay Burqin Chonghuer Gas Station	0. 802 3
5	Altay Farm Tenth Division Tianshan Road Gas Station	0. 791 6
6	Altai Alahake Gas Station	0. 788 3
7	Yili Nilke Jorma Gas Station	0. 787 8
8	Altay Burqin Xingfu Road Gas Station	0. 787 1
9	Yili Xinyuan Tianhe Gas Station	0. 785 0
10	Altai Habahe Qibal Gas Station	0. 781 3

Area. This result shows that the site vector effectively contains the semantic information of the site, it includes the information of the scenic spots which are visited by tourists during their travels, it is conducive to further route mining.

Cluster Visualization Result Analysis

Since the data itself is completely unlabeled, the visualization of the clustering results is chosen to manually verify the validity of the results. As shown in Figure 7, the dots in the figure represent the top 100 stations in each cluster, the triangles represent 12 5A-level scenic spots in Xinjiang, the lines represent popular tourist routes.

It can be seen that the tourism resources of northern Xinjiang are more abundant and concentrated. Almost all tourists will go to northern Xinjiang for sightseeing. Therefore, the two routes involve sightseeing in northern Xinjiang. Their itineraries in northern Xinjiang are almost the same, including 10 5A-level scenic spots in northern Xinjiang. Compared with northern Xinjiang, tourist attractions in southern Xinjiang are concentrated in the Kashgar area, it is far away from other attractions. Therefore, some tourists choose to return directly after visiting northern Xinjiang (route A), while the rest continue to visit Kashgar, they travel to Qinghai Province via Ruoqiang County in Bazhou (Route B). Figure 8 is the Xinjiang self-driving tour route which is recommended by Mafengwo

Figure 7. Visualization of clustering results

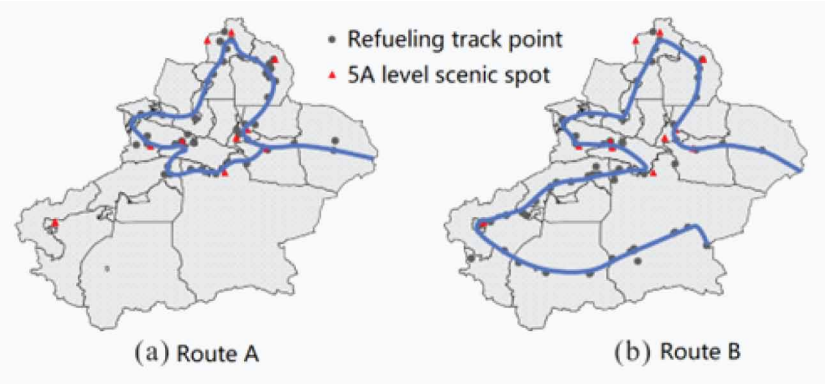
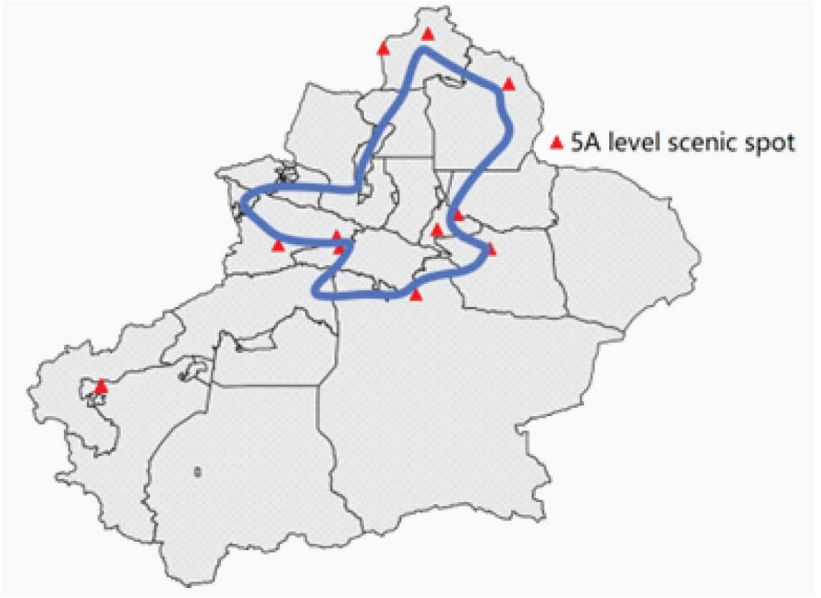


Figure 8. A self-driving tour route recommended by Mafengwo website



Tourism Network. It can be seen that this route has a high degree of overlap with Route A in Figure 7. The difference is that the Mafengwo recommended route starts and ends in Urumqi, it is because of tourism websites usually use Urumqi as a gathering place for tourists.

CONCLUSION AND OUTLOOK

Self-driving tour is a type of self-guided tour, it is a new type of tour that is different from the traditional group tour. Self-driving travel provides tourists with a flexible space in terms of object selection, participation procedures, and freedom of experience. Self-driving tour itself has inherent characteristics such as freedom and individuality, flexibility and comfort, selectivity and seasonality, there is its own characteristics and charm compared with the traditional way of participating in a group.

The user's refueling data in Xinjiang is used in this paper, the definitions of self-driving route, refueling trajectory and refueling trajectory clustering are given in the refueling trajectory clustering, the station representation is researched, the refueling trajectory clustering algorithm is designed based on semantic representation. According to the behavior characteristics of self-driving tourists, the group of self-driving tourists is excavated from the All Xinjiang Refueling Data Set, the basic characteristics of tourists is analyzed. The result showed the reliability of the data on the group of tourists, two popular self-driving travel routes in Xinjiang have been successfully excavated. In view of the fact that the existing trajectory clustering algorithm cannot solve the problem of too sparse refueling trajectory, a semantic representation-based refueling trajectory clustering algorithm is proposed. The final visualization results show that this method can restore the travel route of tourists well. However, the semantic information of the trajectory points is only considered in the method of this paper, the spatial information of the trajectory points is not taken into account. In subsequent research work, spatial and semantic information are combined to learn better stations and trajectory. In addition, the excavated popular routes in this paper are long-distance self-driving tours. The follow-up work will further explore short-distance self-driving tours based on information such as festivals and seasons.

Behind the content is the information content, behind the information content is the Semantic Web, and behind the Semantic Web is the use value. In the era when everyone is a self-media

platform, it is not too difficult to produce content, but the threshold for producing content with use value is self-evident. In this paper, we try to study autonomous driving travel route mining from the application point of view.

With the popularity of automobiles, self-driving tours have become more and more common. The fun of self-driving travel is to do whatever we want, and the biggest advantage of self-driving travel is that we can do whatever we want. But if we want to have a perfect and happy self-driving travel experience, there are many details that we need to pay attention to. Here is just a study of a self-driving tour route extracted from the refueling information of the gas station. During the self-driving tour, there are many factors such as weather, geography and human beings. If enough data is accumulated, deep learning can be used to mine self-driving tour routes in the future.

FUNDING AGENCY

This work was supported by the Scientific Research Project (No: 20B337) of Hunan Provincial Education Department, China.

REFERENCES

Besse, P. C., Guillouet, B., Loubes, J.-M., & Royer, F. (2016). Review and perspective for distance-based clustering of vehicle trajectories. *IEEE Transactions on Intelligent Transportation Systems*, *17*(11), 3306–3317. doi:10.1109/TITS.2016.2547641

Chang, L., Sun, W. P., & Zhang, W. T. (2019). Review of tourism route planning. *Zhineng Xitong Xuebao*, *14*(1), 82–92.

Chen, L., & Özsu, M. T., ORIA V. (2005). Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*. New York: ACM. doi:10.1145/1066157.1066213

Chen Z H, Ming X G,Zhou T T & Chang Y. (2020). A hybrid framework integrating rough-fuzzy best-worst method to identify and evaluate user activity-oriented service requirement for smart product service system. *Journal of Cleaner Production*, *253*(20), 119954.1-119954.19. 10.1016/j.jclepro.2020.119954

Cui, G., Luo, J., & Wang, X. (2018). Personalized travel route recommendation using collaborative filtering based on GPS trajectories. *International Journal of Digital Earth*, *11*(3), 284–307. doi:10.1080/17538947.2017.1326535

Grbovic, M., & Cheng, H. (2018). Real-time personalization using embeddings for search ranking at Airbnb. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM. doi:10.1145/3219819.3219885

Jiang, S., Qian, X., Shen, J., Fu, Y., & Mei, T. (2015). Author topic model-based collaborative filtering for personalized POI recommendations. *IEEE Transactions on Multimedia*, *17*(6), 907–918. doi:10.1109/TMM.2015.2417506

Kenthapadi, K., Le, B., & Venkataraman, G. (2017). Personalized job recommendation system at LinkedIn: practical challenges and lessons learned. In *Proceedings of the 11th ACM Conference on Recommender Systems*. New York: ACM. doi:10.1145/3109859.3109921

Lee, J. G., Han, J., & Whang, K Y. (2007). Trajectory clustering: A partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. New York: ACM. doi:10.1145/1247480.1247546

Lei Lim X. Z. (2021). Research on the experience of domestic Silk Road self-driving tour based on network text. *Journal of Chengde Vocational College*, *26*(4), 13-19. doi:.1674-2079.2021.04.00410.3969/j.issn

Li, X. (2021). Study on Measurement of Seasonality and its Regional Differences of Self-driving Tours Market in China. *Luyou Xuekan*, *36*(8), 140–154. doi:10.19765/j.cnki.1002-5006.2021.08.016

Liu, Y. P., Bao, J. G., & Huang, Y. H. (2019). Study on spatio-temporal behaviors of self-driving tourists based on GPS data: A case study of Tibet. *World Regional Studies*, *28*(1), 149–160.

Mikolov, T., Sutskever, I., & Chen, K. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 2013 Conference on Neural Information Processing Systems*. MIT Press.

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM. doi:10.1145/2623330.2623732

Radosavljevic, V., Grbovic, M., & Djuric, N. (2016). Smartphone app categorization for interest targeting in advertising marketplace. In *Proceedings of the 25th International Conference Companion on World Wide Web*. New York: ACM. doi:10.1145/2872518.2889411

Song, M. J., & Wu, Y. H. (2021). A Distributed Uniformity Adaptive Ant Colony Algorithm for Driving and Traveling Route Planning. *New Generation of Information Technology*, *4*(4), 36–41. doi:10.3969/j.issn.2096-6091.2021.04.006

Tang, W., Pi, D., & He, Y. (2016). A density-based clustering algorithm with sampling for travel behavior analysis. In *Proceedings of the 2016 International Conference on Intelligent Data Engineering and Automated Learning*. Cham: Springer. doi:10.1007/978-3-319-46257-8_25

Vlachos, M., Kollios, G., & Gunopoulos, D. (2002). Discovering similar multidimensional trajectories. In *Proceedings of the 18th International Conference on Data Engineering*. IEEE. doi:10.1109/ICDE.2002.994784

Wang, J., Huang, P., & Zhao, H. (2018). Billion-scale commodity embedding for e-commerce recommendation in Alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM. doi:10.1145/3219819.3219869

Wei, L. Y., Zheng, Y., & Peng, W C. (2012). Constructing popular routes from uncertain trajectories. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM. doi:10.1145/2339530.2339562

Wieting, J., Bansal, M., & Gimpel, K. (2015). *Towards universal paraphrastic sentence embeddings.* arXiv preprint arXiv:1511.08198.

Yi, B. K., & Jagadish, H. V., & Faloutsos, C. (1998). Efficient retrieval of similar time sequences under time warping. In *Proceedings of the 14th International Conference on Data Engineering*. IEEE.

Zheng, V. W., Zheng, Y., & Xie, X. (2010). Collaborative location and activity recommendations with GPS history data. In *Proceedings of the 19th International Conference on World Wide Web*. New York: ACM. doi:10.1145/1772690.1772795

Zheng, Y. (2015). Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology*, *6*(3), 1–41. doi:10.1145/2743025

Zheng, Y., Zhang, L., Ma, Z., Xie, X., & Ma, W.-Y. (2011). Recommending friends and locations based on individual location history. *ACM Transactions on the Web*, *5*(1), 5. doi:10.1145/1921591.1921596

Zheng, Y., Zhang, L., & Xie, X. (2009). Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th International Conference on World Wide Web*. New York: ACM. doi:10.1145/1526709.1526816

Can Yang (b. 1983) received the M.S. degree in Electronics and Communication Engineering from Central South University, China in 2004. Now, she is a researcher at Hunan Institute of Transportation Engineering, China. Her research interests include information excavation and image processing.