Tourist Spatial-Temporal Behavior Model in Scenic Spot Based on Social Media Data

Yanyan Yang, Xingqiang Zhang and Ying Ma

ABSTRACT

In the paper, a method based on social media data for analyzing tourist temporal-spatial behavior was presented. The web crawler method, the synonym extended text information processing technology and the fuzzy clustering method were combined in order to collect the social media traveling data. The information was transformed into tourist flow data of scenic spots, and the spatial-temporal behavior model of tourists was established, which calculated the passenger flow coefficient inhomogeneity of the scenic spot, traveler tour routes and tour time. At the same time, taking Temple of Heaven Park in Beijing as an example, the popularity of each scenic spot in the park, the tourists' tour routes and the intensity of the distribution of the tourists' stay time are collected and analyzed. The results showed that the spatial-temporal behavior analysis of tourists based on social media data mining in this paper could provide a reference for the planning and management of scenic spots.

KEYWORDS
Social Media, Data Mining, Fuzzy Clustering, Scenic Spot, Spatial-Temporal Behavior Coefficient of Inhomogeneity.

INTRODUCTION

The internet has changed people's lives, and tourists have used social media to share information such as photos, microblogs, and travel notes, which has produced a huge amount of tourism data and various forms of structure. Therefore, the development of a reasonable Internet tourism data mining method and the establishment of a spatial-temporal behavior model of the passenger flow in the scenic spot will help to improve the tourism efficiency and safety management of the scenic spot. Andrea and Others (2009) used digital footprint information such as pictures, call logs and mobile text messages to study the 2006-2008-year tourism situation in New York [1]. Van den Berg and Others (2013) studied the impact of social networks on tourism activity patterns [2]. Li Chunming and Others (2013) analyzed the spatial-temporal behavior of tourists in Gulangyu scenic spot based on network photos [3]. Chen Rong and Others (2014) proposed the ssvr-pso passenger flow prediction model based on support vector regression and particle swarm optimization.
Based on this, this paper used Python to write web crawler scripts to obtain social media text information and used synonym extension text information processing technology and fuzzy clustering analysis method to cluster the keywords of scenic spots to ensure the speed and quality of information obtained. This paper establishes a spatial-temporal behavior analysis model of tourists in scenic spots from the aspects of time and tourist routes and distribution of tourist attractions. Taking Temple of Heaven Park in Beijing as an example, it makes a case study to analyze the time and spatial behavior of tourists in scenic spots, the flow direction of tourists in scenic spots and the popular scenic spots in scenic spots, so as to provide theoretical support for Tourism Management in scenic spots.

UNSTRUCTURED SOCIAL MEDIA DATA MINING

Social media has a large amount of data, a wide variety of data types, and low-value density. This article used a Python-based web crawler program to obtain Sina Weibo and other social media and travel and other Internet data. Using Chinese word segmentation technology to split and refine the original data, according to the exported words, the passenger flow is corrected and classified (example shown in Table I and Table II). And then the fuzzy clustering analysis method is used to cluster analysis, to meet the requirements of the management accuracy of this data.

TABLE I. DATA FEATURE WORDS THAT NEED TO BE ELIMINATED OR CORRECTED.

<table>
<thead>
<tr>
<th>Data Feature Words Need to be Eliminated</th>
<th>Surrounding buildings: Temple of Heaven Hospital, The Mall, Movie, Tiantan Pharmaceutical ...... Advertisement: Tiantan East Road, nearby, lie in, Surrounding ......</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Feature Words Need to be Corrected</td>
<td>Number of people involved: A line, Together, One family, Accompany, with ......</td>
</tr>
</tbody>
</table>

TABLE II. CONVERSION OF TEXT DATA AND PASSENGER FLOW DATA.

<table>
<thead>
<tr>
<th>Social media text data</th>
<th>Robby Dazhao: On May 1, ntt, three of us went to &lt;em class=&quot;red&quot;&gt;Temple of Heaven&lt;/em&gt; Park to play. It was very hot that day and there were many flowers in it. We went to pray for the New Year Hall first, then to the echo wall, and finally to the dome. The Hall of Prayer for the Year is a circular hall with three layers of glazed tiles symbolizing the &quot;blue sky&quot;. The echo wall is so magical that one person standing by the wall and another standing in the distance can hear what that person is saying. The dome symbolizes the center of heaven. In ancient times, people stood there praying for good weather.ntt2010-05-04 19:30:14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passenger Flow Data</td>
<td>ID: Da Zhao Robby ; Time:2010-05-01; Time identification: Holiday and Vacations; Attraction Name: Temple of Heaven Route number : 2 ; Passenger flow: 3 ;</td>
</tr>
</tbody>
</table>
Fuzzy Clustering Analysis Method

The steps of fuzzy cluster analysis are as follows:

Set \( y_i^j (i = 1, 2, \ldots, n) \) is a set of synonyms for the name of Attraction K, \( y_{ij} \) represents the amount of social media data for year j of the i synonym, and if the \( y_{ij} \) is arranged in matrix form by year, the original data matrix \( V \) can be obtained:

\[
V = \begin{pmatrix}
y_{11} & y_{12} & \cdots & y_{1m} \\
y_{21} & y_{22} & \cdots & y_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
y_{n1} & y_{n2} & \cdots & y_{nm}
\end{pmatrix}
\]

(1)

Let \( y_{\text{min}} \) and \( y_{\text{max}} \) be the minimum and maximum values in \( y_{ij} \). The original data of social media is standardized by range transformation, and the sample data is transformed into the number between [0,1].

\[
y_{ij}' = \begin{cases}
0, & y_{ij} = y_{\text{min}} \\
\frac{y_{ij} - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}}, & y_{\text{min}} < y_{ij} < y_{\text{max}} \\
1, & y_{ij} = y_{\text{max}}
\end{cases}
\]

(2)

Let \( r_{ij} = R(y_i, y_j) \) be the similarity coefficients of the synonym sets \( y_i \) and \( y_j \). The direct distance method is used to determine \( r_{ij} \). C is a suitable parameter to make \( 0 \leq r_{ij} \leq 1 \). \( d(y_i, y_j) \) denotes the Euclidean distance between sets \( y_i \) and \( y_j \).

\[
r_{ij} = 1 - cd(y_i, y_j)
\]

(3)

Based on the direct clustering method, the fuzzy similarity matrix \( U \) of each scenic spot is established. The clustering matrix of scenic spots and the set of similar classes are expressed by formulas (4) and (5), respectively.

\[
U = \begin{pmatrix}
r_{11} & r_{12} & \cdots & r_{1n} \\
r_{21} & r_{22} & \cdots & r_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
r_{n1} & r_{n2} & \cdots & r_{nn}
\end{pmatrix}
\]

(4)

\[
[y_i]_U = \{y_j | r_{ij} = \lambda_n\}
\]

(5)

In the formula: \([y_i]_U\) is a set of similar classes under cluster matrix \( U \) for synonyms of attraction i; \( \lambda_n \) is a similarity classification threshold, and the value of threshold \( \lambda_n \) is directly related to the size of clustering. In the process of decreasing the value of \( \lambda_n \) from 1 to 0, the category increases, and the result is divided from fine to coarse. The range of value of \( \lambda_n \) is [0,1]. Changing the value of \( \lambda_n \) can get the appropriate category. That is to say, when \( n = 1, \lambda_n = 1 \), \([y_i]_U = \{y_j | r_{ij} = 1\}\), the \( y_i \) and \( y_j \) satisfying \( r_{ij} = 1 \) will be placed in a class, forming a similar class; When \( n=2,3 \ldots \), \( \lambda_n < \lambda_{n+1} \), the element pairs with the similarity of \( \lambda_n \) are found from \( U \), and their equivalent classifications are obtained. By analogy, until \( U \) becomes a class.
SPATIO-TEMPORAL MODEL OF TOURISTS IN SCENIC SPOTS

Assuming that n categories are clustered, the passenger flow Q of scenic spot K can be expressed as follows:

\[
Q = \sum_{i=1}^{n} \theta_i \sum_{j=1}^{x} P_{ij}
\]  

(6)

In the formula: \(P_{ij}\) is the total amount of social media data in category i, article; \(x\) is the number of synonyms for each category, not fixed; \(\theta_i\) is the i-type synonym weight coefficient, \(i \in [1,n]\). In this paper, the passenger flow inhomogeneity coefficient \(C_k\) of the scenic spot and the tourist route non-uniformity coefficient \(C_i\) of the scenic spot are used to represent the spatial distribution characteristics of the scenic spot passenger flow, and the passenger flow time inhomogeneous coefficient \(C_t\) of the scenic spot is used to characterize the time distribution characteristics of the scenic spot.

\[
C_k = \frac{Q_k/Q_k^c}{\left(\frac{1}{m} \sum_{k=1}^{m} Q_k^c\right) / m}
\]

(7)

In the formula: \(Q_k^c\) and \(Q_k\) are the maximum passenger flow of the kth scenic spot and the average passenger flow of all m scenic spots, respectively, in the statistical period; K is the number of each scenic spot, \(k \in [1,m]\).

Because of the incompleteness of micro-blog text data, not all text data can mine the line passenger flow. Therefore, this paper uses a method similar to the traffic flow distribution, the scenic entrance is regarded as OD point, each attraction is regarded as the traffic node, takes the tour time as the road right, according to the dynamic balance principle distributes the attraction traffic to the line.

Passenger flow allocation rate of each line can be calculated by the classic Logit path selection model:

\[
P(k) = \frac{\exp\left[-\theta \cdot \frac{t(k)}{\bar{t}}\right]}{\sum_{i=1}^{n} \exp\left[-\theta \cdot \frac{t(i)}{\bar{t}}\right]}
\]

(8)

In the formula, \(P(k)\) is the passenger flow allocation rate of line k; \(t(k)\): the right of the road of the kth line (tour time); \(\bar{t}\) : Average road rights for each route (tour time); \(\theta\): Assign parameters, generally take 3.3; \(m\) : Number of lines.

The passenger flow of tourist routes in scenic spots can be expressed by the non-uniformity coefficient \(C_i\) of tourist routes in scenic spots as follows:

\[
C_i = Q_i/Q_i^c = Q_k/(\sum_{k=1}^{m} Q_k / l)
\]

(9)
In the formula, $Q_l$ and $\bar{Q}_l$ are the maximum passenger flow of route $l$ and the average passenger flow of $n$ tourist routes in scenic spots, respectively; $l$ is the symbol of each passenger flow line, $l \in [1, n]$.

Similarly, the passenger flow time distribution in the scenic spot can be expressed by the passenger flow time unevenness coefficient $C_t$:

$$C_t = \frac{Q_t}{\bar{Q}_t} = \frac{Q_t}{\left(\sum_{t=1}^{s} Q_t/s\right)}$$

(10)

In the formula, $Q_t$ and $\bar{Q}_t$ are the social media data related to scenic spots in the $t$ statistical period and the average passenger flow of scenic spots in the $s$ statistical period, respectively. $t$ is the statistical period, day/month, $t \in [1, s]$.

**SPATIO-TEMPORAL BEHAVIOR OF TOURISTS IN TEMPLE OF HEAVEN PARK**

**Distribution of Passenger Flow Time in Temple of Heaven Park**

Use the Python program to search for social media in Temple of Heaven Park in Beijing in five years. Search with the keyword “Temple of Heaven” and get 83,478 social media data. Using social media text data, we can count the passenger flow in each month of five years, as shown in Figure 1. The monthly passenger flow distribution in Temple of Heaven Park was relatively uniform, and the maximum passenger flow non-uniformity coefficient was 1.2 in August. Because of the cold weather, the minimum passenger flow appeared in January, and the coefficient of passenger flow non-uniformity was 0.74. With the warming climate, the vibrant scenery of the scenic spots, the Qingming and the May Day holiday, the March-May season has become a tourist season. At the same time, due to summer vacation, National Day and other holidays, July-October was the peak season of tourism in that period.

![Figure 1](image1.png)

Figure 1. Average monthly passenger volume and monthly non-uniformity coefficient in 15 years.

Figure 2 (a), (b), (c) are the daily passenger flow distribution of Temple of Heaven Park in 2012. The daily passenger flow of Temple of Heaven Park was influenced by holidays, weekends and special activities in scenic spots, such as the Spring Festival Temple Fair and Lantern Festival. The largest daily passenger flow occurred on May 2 and October 2, followed by January 1 and February 6, and $C_t$ was 2.69, 2.87, 2.31 and 2.01, respectively, which indicated that the holidays of May 1, National Day and New Year's Day and the Lantern Festival Temple Fair were the main time for peak passenger flow in Temple of Heaven Park. The
weekend passenger flow was obviously higher than the working day passenger flow, and the monthly passenger flow peak change was very obvious. The peak passenger flow dates of Temple of Heaven Park in 2012 were August (15 days), May (14 days) and July (11 days), which indicated that holidays and suitable climatic conditions were the main factors for tourists to choose Temple of Heaven scenic spot. The biggest change of passenger flow was in October, the National Day holiday has 7 days, and the tourist flow in the scenic spot was a blowout, after which the whole tourist flow in the scenic spot not high.

Figure 2. Daily unevenness coefficient for each month of 2012.

Spatial Distribution of Passenger Flow in Temple of Heaven Park

On social media, visitors generally use different words to describe the attraction, so determining the synonym describing the attraction can guarantee the integrity of the data obtained. Take the scenic spot "Praying Hall" as an example. The synonymous keywords of "Qinian Temple" in scenic spots are "Qigutan Temple, Daqidian Hall, Taihengdian Hall, twelve golden pillars, Dahengdian Hall, Zaojing Hall, Longjing Pillar" and so on. Using the social media data of "Qinian Hall" and its synonyms as the original matrix, the similarity matrix is obtained as follows.

\[
\begin{pmatrix}
1 & 0.9 & 1 \\
0.6 & 0.7 & 1 \\
0.6 & 0.3 & 0.8 & 1 \\
0.4 & 0.1 & 0.8 & 0.8 & 1 \\
0.7 & 0.7 & 0.5 & 0.4 & 0.6 & 1 \\
0.2 & 0.5 & 0.7 & 0.6 & 0.5 & 0.3 & 1 \\
0.2 & 0.4 & 0.6 & 0.7 & 0.1 & 0.7 & 0.7 & 1
\end{pmatrix}
\]

According to the similarity matrix, the synonyms of "Praying Hall" are clustered, and the clustering results are obtained from different similarity classification thresholds $\lambda_n$. Finally, it can be grouped into three categories: group 1 includes Longjing pillar, algae well and Dahengdian hall, group 2 includes...
Taihengdian hall, Daqidian hall, twelve "golden pillars", group 3 includes Qinian hall and Qigutan hall.

Assign weights to various synonyms to characterize their similar relationship to keywords [7]. The weight of the degree of correlation between words can be calculated by the following formula:

$$\theta_k = \frac{F_{xy}}{\sqrt{F_x \cdot F_y}}$$ (11)

In the formula: $F_{xy}$ denotes the frequency of two words appearing together in the data set; $F_x$ and $F_y$ denote the frequency of each word appearing independently. According to the weight coefficient, the search results of various synonyms are screened, that is, the number of data searched by various synonyms is multiplied by the weight $\theta$. The results are shown in Table III.

<table>
<thead>
<tr>
<th>$\theta_k$</th>
<th>Meaning</th>
<th>Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Theta_1=0.38$</td>
<td>Synonyms and keywords have very different meanings</td>
<td>Long Jing Zhu; Caisson ceiling; Daxiang Temple</td>
</tr>
<tr>
<td>$\Theta_2=0.67$</td>
<td>Synonyms have weaker meanings than keyword meanings</td>
<td>Taixiang Temple; Great prayer hall; Twelve &quot;golden columns&quot;</td>
</tr>
<tr>
<td>$\Theta_3=1$</td>
<td>Synonyms and keywords have the same meaning</td>
<td>Hall of Prayer; Prayer Valley altar</td>
</tr>
</tbody>
</table>

Similarly, the amount of synonym data of other attractions can be obtained. After calculation, the most popular tourist attraction is the Hall of Prayer, which has a passenger flow unevenness coefficient of 4.47, followed by the Qiuqiu and Echo Walls, which were 3.27 and 2.41 respectively. As the landmark building of the Temple of Heaven and even Beijing, the Hall of Prayer has attracted a large number of domestic and foreign tourists. When you come to Beijing, you will go to the Temple of Heaven, and you will go to the Hall of Prayer. At the same time, the scenic spots on the north-south axis of the scenic spot are densely populated, and the newly opened Zhaigong and Kagura Departments and seasonal visitors have more passengers. Wanshouting has the lowest passenger flow intensity and its passenger flow unevenness coefficient is 0.09.

The passenger flow distribution rates of 10 routes for sightseeing in the scenic spot are as follows.

Route 1. Eastern Gate - Zaizhu Pavilion - Shenchef - Long Corridor - Qinian Hall - Echo Wall - Yuqiu - South Gate: 36.67%.
Route 2. Eastern Gate - Qixingshi-Long Corridor-Qinian Hall-Echo Wall-Yuqiu-South Gate: 29.75%.
Route 3. East Gate - Long Corridor - Qinian Hall - Baiyuan - Kowloon Bai - Echo Wall - Yuqiu - South Gate: 18.61%.

119
Route 4. East Gate - Long Corridor - Qinian Hall - Longevity Pavilion - Baiyuan - Cypress Forest - Echo Wall - Yuqiu - South Gate: 14.97%.
Route 5. East Gate - Long Corridor - Qinian Hall - Echo Wall - Yuqiu - Cypress Forest - Zhaigong - Shenyue Department – West Gate: 100%.
Route 6. South Gate - Iwaqiu - Echowall - Jiulongbai - Zhaigong - Baiyuan - Qinian Palace - Emperor Qiandian - North Gate: 32.26%.
Route 7. South Gate - Yuqiu - Echowall - Bailin - Baiyuan - Qinian Hall - Emperor Qiandian - North Gate: 23.94%.
Route 8. South Gate - Yiqiu - Echowall - Jiulongbai - Danweiqiao - Qinian Hall - Emperor Qiandian - North Gate: 43.8%.
Route 9. South Gate - Yiqiu - Echowall - Danweiqiao - Qinian Hall - Baiyuan - Zhaigong - West Gate: 100%.

The non-uniformity coefficient of passenger flow line is shown in Figure 3.

![Figure 3. passenger flow lines coefficient inhomogeneity of the scenic spot.](image)

The uneven distribution of spatial-temporal passenger flow routes in Temple of Heaven scenic spot varies greatly. The convenience of import and export traffic in the scenic spot, the hot spots of the scenic spots and the sightseeing time are the most important factors affecting tourists’ choice of scenic spots. The route East 2 has the largest number of tourists, and its passenger flow line unevenness coefficient reaches 3.44, which is because the route contains all the popular attractions of the Temple of Heaven Park, spending less time, and convenient transportation between the East Gate and the South Gate. Followed by the number of visitors to route 1, the passenger flow line unevenness coefficient reached 3.09. The least passenger flow is route 9, and the passenger line non-uniformity coefficient is only 0.06. It is 57 times different from line 2. Tourist routes with non-uniformity coefficient of passenger flow not exceeding 1 include 6, 7, 8 and 9.

**CONCLUSION**

1) This paper constructed a tourism data mining solution of social media. Python was used to write web crawler scripts to obtain social media text information; based on Chinese word segmentation technology, Python was used to merge synonyms, and fuzzy clustering method was used to cluster keywords of scenic spots, which improved the utilization rate and processing accuracy of social
media text information, and ensured the speed and quality of information acquisition.

2) Establishing the spatial-temporal behavior characteristics analysis model of tourists in tourist scenic spots, characterizing the temporal behavior characteristics of tourists by the uneven time coefficient of passenger flow, characterizing the spatial behavior characteristics of tourists by the uneven coefficient of passenger flow and the uneven coefficient of the line. The model can provide a reference for the management of tourists in scenic spots.

3) The spatial and temporal characteristics of passenger flow in Temple of Heaven Park in Beijing are analyzed. The distribution of passenger flow time in Temple of Heaven Park is affected by factors such as climate, holidays and scenic spots. The May 1st and National Holiday are the peak days of passenger flow, April, August, and September are peak months of passenger flow. the attractions of the Hall of Prayer, Qiuqiu and Echo Wall are the tourist hotspots of the scenic spot; the passenger route is mainly influenced by the convenience of the import and export of the scenic spot, the hot spot of the route and the time of the tour.

ACKNOWLEDGEMENTS

Corresponding author: Yanyan Yang.

REFERENCES