Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

Article

Exploring Group Movement Pattern through Cellular Data: A Case Study of Tourists in Hainan

Xinning Zhu [1,*], Tianyue Sun [1] , Hao Yuan [2], Zheng Hu [2] and Jiansong Miao [1]

[1] Beijing University of Posts and Telecommunications,Beijing, China; zhuxn@bupt.edu.cn; sunty@bupt.edu.cn; miaojs@bupt.edu.cn

[2] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China; yuanhao@bupt.edu.cn; huzheng@bupt.edu.cn

* Correspondence: zhuxn@bupt.edu.cn

Abstract: Identifying group movement patterns of crowds and understanding group behaviors is valuable for urban planners, especially when the groups are special such as tourist groups. In this paper, we present a framework to discover tourist groups and investigate the tourist behaviors using mobile phone call detail records (CDRs). Unlike GPS data, CDRs are relatively poor in spatial resolution with low sampling rates, which makes it a big challenge to identify group members from thousands of tourists. Moreover, since touristic trips are not on a regular basis, no historical data of the specific group can be used to reduce the uncertainty of trajectories. To address such challenges, we propose a method called group movement pattern mining based on similarity (GMPMS) to discover tourist groups. To avoid large amounts of trajectory similarity measurements, snapshots of the trajectories are firstly generated to extract candidate groups containing co-occurring tourists. Then, considering that different groups may follow the same itineraries, additional traveling behavioral features are defined to identify the group members. Finally, with Hainan province as an example, we provide a number of interesting insights of travel behaviors of group tours as well as individual tours, which will be helpful for tourism planning and management.

Keywords: Low accuracy CDRs; Group movement pattern; Data mining; Travel behaviors

1. Introduction

With the progress of location-acquisition techniques, a large amount of spatio-temporal data can be acquired from GPS, Wi-Fi, cellular networks and Location-Based Social Networks(LBSN) in the form of trajectories. The increasing trajectory data enables us to discover knowledge that is meaningful in aiding the research of human mobility. A branch of research aimed at discovering group movement patterns in order to analyze behaviors of group members and has attracted attentions recent years [1–10]. Some group movement patterns have been proposed in previous studies like flock[3], convoy[4], swarm[5], traveling companion[6–8] and gathering[9], which are summarized in [10] and can be applied to different situations. These patterns consider the number of the snapshots corresponding to the time when objects stay together as the measurement to judge whether they constitute a group. Usually, these patterns perform well on high-quality GPS data. But for some trajectory data with low spatial resolution and sampling rate, these patterns may not be discovered correctly.

In this paper, we put efforts into discovering groups of tourists who travel together for a certain time period using anonymized Call Detail Records (CDRs) data. The travel behaviors of group members are then explored, which are of great value to the research of tourism. CDR data contains time and location information of mobile phone users corresponding to when and where the records were generated. Compared to GPS data, CDR data can be obtained at a lower costs and on a scale of millions of users easily. So CDRs are useful in trajectory data mining and human mobility analysis at a city level.

© 2018 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

However, we are still facing some challenges in the process of mining group tourists movement patterns from the huge amount of CDR data. Firstly, it can be very time-consuming if we try to discover groups from the trajectory data which consists of thousands or even millions of users. The second challenge comes from the sparsity in both spatial and temporal resolutions of CDR data. Despite the huge scale of the trajectory data acquired from CDRs, the data of every single user is of low accuracy and frequency. So the traditional similarity metrics for GPS trajectories are not applicable to CDR data. The third challenge is the complexity of tourists behaviors. Generally, tourists are willing to visit well-known scenic areas. When different groups visit the same scenic areas, their trajectories may overlap, which makes it hard to distinguish two groups with similar travel routes. Besides, group members may not travel together all the time, some members who are more active may have a bigger region of activities. In such a situation, if we discover groups only by the number of the timestamps when the objects stay together, the result will be imprecise. Therefore, other features need to be considered to help identify different tourist groups.

Considering the challenges we are facing, we propose a Group Movement Pattern Mining based on Similarity (GMPMS) to identify co-movement patterns from low accuracy trajectory data. Unlike the moving together patterns summarized in [10], GMPMS disregards the restrictions on the shape or density of a group. Instead, trajectory similarity and some other features such as accommodation similarity are used to capture the group movement patterns from sparse trajectories.

We first remove massive irrelevant trajectories and get candidate groups by employing the frequent item set mining method. In this step, only objects that stay together within a predefined distance during a certain period are retained as candidate groups. Next, we design a method to measure the similarity of tourists in each candidate group considering multiple features extracted from the trajectory data, such as trajectory similarity, accommodation similarity and other traveling features that can reflect the relationship between group members. Finally, a semi-supervised learning algorithm is used to identify real tourist groups from candidate groups.

The main contributions presented in this paper are:

- We proposed a new group movement pattern mining method based on similarity that can identify groups from the huge amount of mobile trajectory data;
- We designed an algorithm to calculate trajectory similarity of objects with low accuracy data;
- We explored different travel behaviors of group tourists and individual tourists.

The rest of the paper is organized as follows: Section 2 discusses related work, Section 3 describes the problem in this work and the overall algorithms for mining tourists group movement patterns, Section 4 shows our experimental study and discussion, and Section 5 summarizes this paper.

2. Related Works

2.1. Group Movement Mining

Discovering groups of objects that move together for a certain period is an important research objective in mobility studies. Some inferences have been proposed to explain different group patterns. These group patterns can be distinguished from three aspects: the shape of the groups, whether the groups are continuous, and whether the members are variable during the lifespan of the pattern[1].

The flock[3] pattern is a group of objects that travel together within a circle of some user-specified size for at least K consecutive timestamps. However, the group in flock is restricted in a circular region which leads to a lossy-flock problem. To solve the problem, the convoy[4,11] pattern is proposed to describe arbitrary shape of groups by using density-based clustering. Li et al. proposed swarm[5], which allows the K timestamps to not be consecutive. Because of the fact that the members of a group are not always gathered, this pattern allows the members to disappear for some time as long as they stay together for at least K timestamps. Tang et al.[8] proposed traveling companion pattern, which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

uses a data structure called traveling buddy to continuously find patterns. This pattern can be regarded as an online detection fashion of convoy.

Besides, Zheng et al.[9] proposed a gathering pattern to discover various group incidents such as celebrations, parades and so on. Wang et al.[12] proposed two kinds of loose group movement patterns namely Weakly Consistent Group Movement Pattern (WCGMP) and Weakly Consistent and Continuous Group Movement Pattern (WCCGMP), and corresponding discovery algorithms. The WCGMP and the WCCGMP are collectively referred to as the loose group movement patterns, in which the gatherings of a group could be partial and variable during the group's lifespan. Zheng et al. [13] proposed the GMOVE pattern and adopted HMMs to build a model.

Patterns mentioned above can be used to find co-travelers. But due to the low accuracy of CDRs and the complexity of travel behaviors, these patterns may be not suitable in our study. The reason is discussed in the section 3.

2.2. Trajectory Similarity

Devices that can be used to track moving objects has increased dramatically, leading to a great growth in movement data. Most movement data are captured and stored in the form of trajectories. One important class of trajectory analysis is the measurement of similarity between trajectories. Several similarity metrics have been proposed to calculate the distance between two trajectories, such as Longest Common Subsequence (LCSS), Dynamic Time Warping (DTW), Edit Distance (ED) and so on [14,15].

DTW searches all point combinations between two trajectories for the one with minimal cost even if they are not aligned in the time axis. So the similarity between trajectories of different lengths and with local time shifting can be computed. LCSS can be used as a similarity measure in which some points are able to remain unmatched in an attempt to provide an accurate similarity result. However, for trajectories with widely varying sampling rates, it may lead to many points being unmatched. ED is to count the minimum number of edits required to make two trajectories equivalent. Several variations of edit distance exist including Edit Distance with Real Penalty (ERP) and Edit Distance on Real Sequence (EDR). ERP and EDR both take into account local time shifting and allow similarities to be found between trajectories of different lengths.

Besides, a shape-based similarity measure for trajectory data is proposed by [16], the algorithm is based on vectors instead of individual data points. Liu H et al. [17] also focuses on measuring similarity between moving objects, and defines the trajectory similarity from the geographic aspect and the semantic aspect. In addition, Wang, F et al. [18] proposed that semantic trajectory should also be utilized and designed a novel semantic trajectory similarity measurement to estimate similarity among users.

However, in our work, the trajectory data extracted from CDRs is of low accuracy. Similarity measures mentioned above are not suitable because we need to deal with the influence caused by different sampling rates and data sparsity. The algorithms that go through all points to compare each pair of points to detect similarity may be not precise. In this case, a novel algorithm to measure trajectory similarity in low accuracy is needed.

2.3. Travel Behaviors

To get an insight of travel behaviors, many researchers make efforts to mine tourist behavior patterns using GPS data from mobile devices, check-ins collected from social network, geo-tagged photos uploaded by tourists, GIS information and so on. Xue M et al.[19] identified the tourists among public commuters using the public transportation data provided by Singapore's Land Transport Authority and then revealed the travelling patterns of tourists. Vu H Q et al.[20] looked insights into tourist behaviors by exploiting the socially generated and user-contributed geotagged photos that have been made publicly available on the Internet. And similarly, Yang, L et al. [21] also utilized geo-tagged photos from Flickr to extract trajectories of tourists to detect tourist mobility patterns. Sun, Y et al. [22]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

empirically investigates the travel and activity patterns of active local Foursquare users in New York City. And in addition, Phithakkitnukoon S et al.[23] analyzed tourist behaviors in Japan using massive mobile phone GPS location records to study travel behaviors.

Despite much effort put into travel behaviors mining, to have a full understanding of tourist behaviors in an area is still not an easy task. One of the most important reason is that different kinds of tourists may have different behaviors, which leads to the diversity of travel behaviors. In our work, we try to analyze travel behaviors and find out the difference between group tourists and individual tourists.

3. Materials and Methods

3.1. Problems and Framework

3.1.1. Problem Definition

In the field of group pattern mining, some common concepts are shared in many different studies. In this section, we employ some notations of trajectory mining, and then illustrate the problem we aim to solve. The notations used throughout this paper are listed in Table 1

Table 1. Description of the Notations

Notation	Description
O_i	the moving objects
T	the trajectory set
T_a, T_b	the trajectory of users
d_{th}	the distance threshold of stay points
τ	the time threshold of stay points
SP	the stay points set
sp_i	the stay points
$sp_i.t$	the timestamp of the stay point sp_i
TI	the time interval of snapshots
S	the snapshot set
s_i	the snapshots
C	the collection set
c_i	the collections
d_c	the distance threshold in collections
M	the minimum size of groups
K	the minimum snapshots for the occurrence of groups
G	the candidate group set
g_i	the candidate groups
δ_t	the time threshold of matching point
δ_d	the distance threshold of center of mass
δ_s	the distance threshold of similarity

In this paper, two types of cellular data are used actually. One is CDR which is generated when there are incoming or outgoing calls or short messages. The other is location-based data generated in the case of some passive network events, such as when hand over happens between two adjacent cells, or after half an hour of inactivity of the mobile phone. For the sake of simplicity, we refer these two types of data as CDRs. CDRs provide a compromise between spatio-temporal resolution and ubiquity. In the raw dataset, each record contains an anonymous user id, the timestamp of the record, the location id which represented by the base station id to which the user connected, the latitude and longitude of the base station, and the user's registration location which discloses where the user comes from. To discover the patterns of group movement, we need to identity the stay points which usually stand for a meaningful location and convert the raw dataset to a sequence of trajectories which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

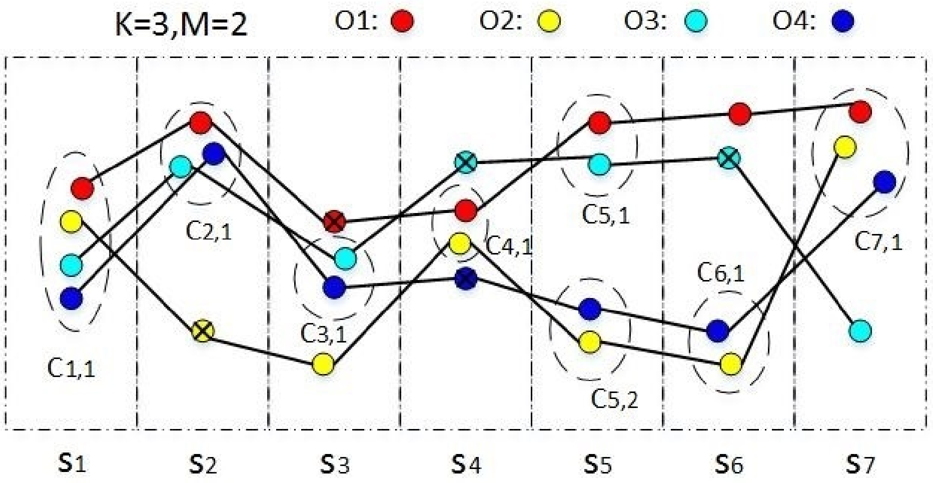
Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

separated by stay points. Group members may have the same stay points where they gather together for a certain period.

Definition 1. *(research problem) : Given a large number of CDRs with high sparsity, our task is to discover group movement patterns of tourists.*

In addition to the poor resolution in space and time, inconsistent and non-uniform sampling rates of the trajectories which is inherently the problem for CDR data have to be tackled in particular. Even if two tourists traveled together for the whole journey, the number of trajectory points of them may vary greatly and their trajectory points are not aligned with each other. In an extreme case, the trajectory points of the two tourists traveling together may be recorded at staggered times. So it is necessary to design an algorithm to deal with these issues of the CDR trajectory data.

As shown in Figure 1, trajectories of four objects are divided into seven snapshots, each of which contains trajectory points of users within a predefined period time. Considering the data missing, we use points with "x" label to indicate locations that are missed in the dataset. In each snapshots, the objects within a distance limit are gathered into clusters. The parameter M denotes the minimum number of members in a group and K denotes the minimum number of snapshots for the occurrence of groups. As shown in Figure 1, where M is 2 and K is 3, the flock and the convoy pattern require at least K consecutive timestamps. So the group $\{O2, O4\}$ and $\{O3, O4\}$ can be discovered by both flock and convoy. The swarm pattern is not restricted to the consecutiveness, so it can get more groups as $\{O1, O2\}, \{O1, O3\}, \{O2, O4\}, \{O3, O4\}$.



flock	{O2,O4},{O3,O4}
convoy	{O2,O4},{O3,O4}
swarm	{O1,O2},{O1,O3},{O2,O4},{O3,O4}

Figure 1. Group movement patterns. The results of different patterns are exhibited in the table.

However, it can be seen that $\{O1, O3, O4\}$ is also a group, but this group can't be identified because $O1$ is missed in s_3 and $O3, O4$ is missed in s_4.

Besides the sparsity of CDRs, there is another abnormal situation when identifying these moving together patterns. In the Figure 1, when K is set to be 3, $\{O1, O2\}$ is identified as a swarm pattern. However, it can be seen obviously from the figure that $O1$ and $O2$ stay together only in s_1, s_4 and s_7. The trajectory of the two objects is not similar in other four snapshots. Moreover, we can't filter out

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

this kind of groups by increasing the value of K. Because when K is set to be 4, the real group $\{O1, O3\}$ can't be identified from snapshots. In such a case, it is hard to choose the value of K when identifying groups. That is to say, a new method need to be proposed to solve this problem.

So in this paper, we proposed a method called Group Movement Pattern Mining based on Similarity (GMPMS) to solve this problem by calculating the similarity between objects in the same groups. The method measures similarity between tourists from multiple dimensions and is able to identify tourist groups from sparse CDR data.

3.1.2. Framework

In this part, we will introduce the framework of discovering the group movement patterns based on similarity. We are interested in not only discovering co-movement patterns but also analyzing travel behaviors of different groups of tourists. Figure 2 exhibits our framework diagram. The left column shows data sources, including call record details, points of interest of scenic areas and map data. The column in the middle shows the key components of our algorithm. After data preprocessing and candidate groups filtering, similarity between any two trajectories in each candidate group is computed to obtain the possible co-movement pairs. Then groups are identified based on a semi-supervised learning algorithm or threshold-based methods. Finally travel behaviors are analyzed to get some insights of the tourists. The right column shows the output of our framework.

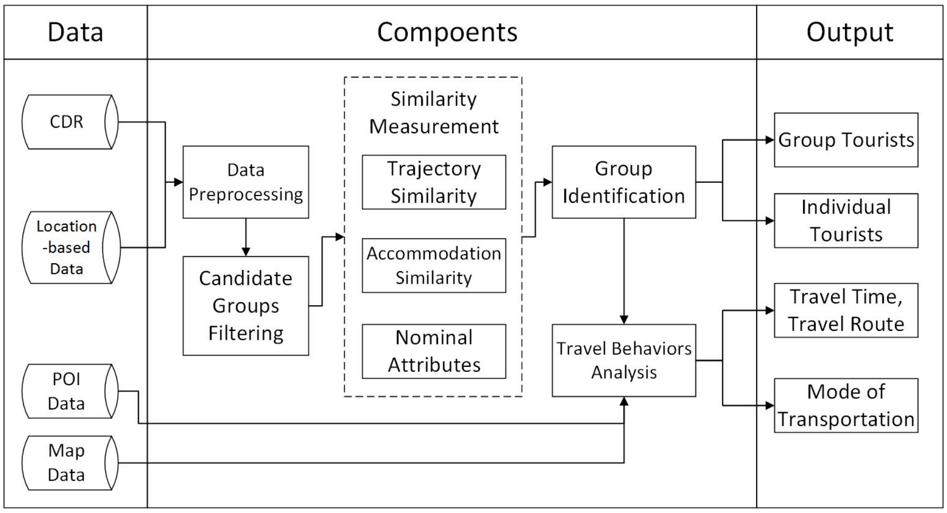


Figure 2. Modeling framework diagram

3.2. Data Preprocessing

In this paper, we focus on the travel behaviors analysis of group tourists and individual tourists in Hainan province. At the foremost, we need to extract the trajectories of the tourists. Tourists identification is non-trivial though, we simplified the problem by an assumption that the mobile phones' registration locations of tourists are provinces other than Hainan. Besides, we also discard the trajectory data of the users who have few records, which may not contribute to the analysis.

When a mobile phone is in the overlapped areas between adjacent cells, it may switch between two cells but actually the user's location hasn't changed. This phenomenon is called the Ping-pong effect which leads to abnormal trajectory in the data. To eliminate such noise, we refer to [24] for detecting and removing oscillation records. After that, we identify stay points from raw trajectories.

Definition 2. *(Stay Point) : A stay point sp represents a location characterized by a sequence of consecutive points in the raw trajectory data which is limited by both temporal and spatial constraints.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

For a given trajectory $T = \{p_1, p_2, ..., p_k\}$, a stay point sp is defined as the centroid of a sub-trajectory $T_{sub} = \{p_i, ..., p_j\}, 1 \leq i \leq j \leq k$, which satisfies the condition that distance between two points in T_{sub} is less than a threshold d_{th}, the time interval between p_i and p_j is greater than a threshold τ.

A stay point sp is generated from T_{sub} and can be denoted as $sp_{sub} = \{u, x, y, t, du\}$, where u is the id of the user, x and y are the central longitude and latitude of points in T_{sub}, t is the timestamp of the first point of T_{sub} i.e. p_i, and du is the time interval between p_i and p_j in T_{sub}, which indicates how long a user stays in this region.

After identifying stay points from raw data, the trajectory of a user converts into a sequence of pass-by points separated by some stay points, other points which do not satisfy the conditions in Definition 2 are called pass-by points.

3.3. Candidate Groups Filtering

After data preprocessing, we are able to calculate the similarity between tourists to identify tourists groups with high similarity between its members. A group of tourists traveling together must have some features in common, such as trajectory, accommodation and so on. Our algorithm aims to discover tourist groups whose members have similar behaviors in relation to each other. However, too many tourists means that the process of similarity calculation between all pairs of tourists will be overwhelmingly time- consuming as well as memory consuming. To solve the problem, we filter out massively irrelevant trajectories at first by deploying a frequent itemset mining method before the process of similarity measurement. In this step, we can get candidate groups whose members appeared together for at least K snapshots of the trajectories. So only the similarity among tourists in a candidate group rather than all tourists need to be computed, which greatly reduces the computational complexity.

Considering that the stay points of the members in one tourists group should be close to each other in space for most of the time, we can remove the irrelevant trajectories of tourists who don't satisfy these conditions and may be individual tourists by applying a frequent itemsets mining method.

At first, we divide a trajectory into a sequence of snapshots by time. Let TI be the interval of snapshots, Ts be the time span of the considered group movement patterns, snapshot set S is a sequence of snapshots $\{s_1, s_2, ..., s_i, ..., s_M\}$, $M = Ts/TI, i = 1, ..., M$. Each element s_i can be expressed as a set of stay points in trajectory, i.e $s_i = \{sp_{i,1}, sp_{i,2}, ..., sp_{i,j}, ..., sp_{i,n} | \, sp_{i,j}.t \in [t_i, t_i + TI]\}$, where t_i is the start time of the i-th snapshot, $sp_{i,j}$ is the j-th stay point whose timestamp is within the interval of the snapshot i and n is the total number of stay points in the snapshot.

For the sake of the sparsity data of CDRs and its inconsistent sampling rate of trajectories, the interval of snapshots need to be selected long enough to ensure that the trajectory points can be included in the same snapshots as long as the difference of their timestamps is no more than TI.

Definition 3. *(Collection) : A collection c is a group of stay points in a snapshot within a distance threshold d_c. The snapshot s_i can be expressed in collections as $s_i = \{c_{i,1}, c_{i,2}, ..., c_{i,j}, ..., c_{i,m}\}$, where $c_{i,j}$ is the j-th collection in s_i and m is the total number of collections in s_i.*

For example, as shown in Figure 1, s_1 to s_7 denote seven continuous snapshots, each of which contains stay points whose timestamp is within the interval of the snapshot. And the objects in a snapshot within a distance threshold belong to a collection. A snapshot may contain numbers of collections. As in s_5, there are two collections $c_{5,1} = \{O1, O3\}$ and $c_{5,2} = \{O2, O4\}$. In this work, we perform a density-based clustering method (DBSCAN) on the snapshots to get collections of objects. The objects which are close enough to each other are clustered into a collection.

By observing some group's trajectories manually, we find that even if a group of tourists are traveling companions, they won't stay together all the time. In some situations, some members in the group may leave and will not come back in the next several snapshots. The strict restrictions on

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

continuity will lead to the loss of real groups. So the problem is how to discover members of a group which appear in the same collection for at least K possibly non-consecutive snapshots. So we formulate the problem into "Market Basket Analysis".

Market Basket Analysis is a modelling technique that mines the association between different items. For example, people who buy bread may also buy butter thus bread and butter often occur together in the bills. So these kinds of problems are formulated to mine frequent itemsets from transaction records. In our study, tourists and collections can be viewed as items and transactions respectively. We aim to find tourists groups (itemsets) which frequently occurred in the collections and satisfies the threshold of support. Therefore, we adopted FP-growth, an efficient method proposed by Han et al. [25], to mine the complete set of frequent itemsets.

With the minimum number of snapshots K and the minimum size M set, the FP-growth algorithm aims to find out groups containing at least M members that traveled together for at least K possibly non-consecutive snapshots. A candidate group g_i is denoted as $g_i = \{size, object1|object2|...|objectN, frequency\}$, where $frequency$ is the number of times that g_i occurred in collections. However, the itemsets obtained from FP-growth are not closed frequent itemsets, resulting an increase in computation. To solve this problem, we define a filtering rule as: For tourists group g_i and g_j, if $g_i \subseteq g_j$ and g_i's support is less than g_j's support, then g_i is removed from the result set. After this step, candidate groups in the result set are guaranteed to be closed.

3.4. Similarity Measurement

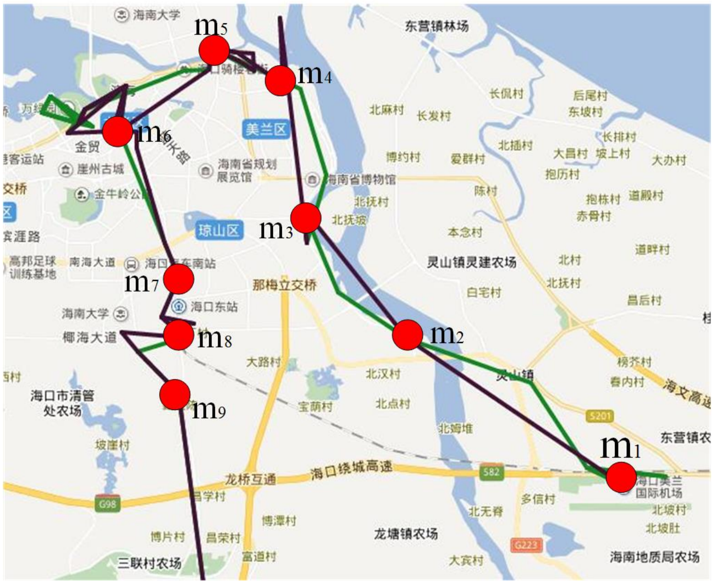


Figure 3. Two trajectories moving together with different sampling rate

After filtering out the irrelevant trajectories, we get a number of closed candidate groups. To discover real tourist groups, we propose a similarity measurement taking into account four features. That is, the trajectory of tourists, the accommodation of tourists, the attribution of tourists and the number of days tourists stayed in Hainan. The similarity of tourist a and b is defined as a vector:

$$Sim\,(a,b) = (Tsim\,(a,b)\,,Asim\,(a,b)\,,Nsim\,(a,b)) \tag{1}$$

where $Tsim(a,b)$ is the trajectory similarity of a and b, $Asim(a,b)$ is the accommodation similarity and $Nsim(a,b)$ is the similarity of the other two features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

3.4.1. Trajectory Similarity

In this part, we perform trajectory similarity measurement for each pair of tourists in the same candidate group.

Because tourists in one group may not stay together all the way and trajectories from CDR are in sparsity, their travel routes and places visited may be different in a local area. Figure 3 illustrates the trajectories of the two tourists belonging to the same group. Although they traveled together in Haikou city, but their trajectories are not similar to each other in some areas. We can see that there are 9 points they stayed together in their trajectories. The trajectory of tourist a in green line has more sampling points than tourist b from m_1 to m_4, which causes the difference between trajectories in these areas. Another problem we can see from Figure 3 is the different travel routes of tourists between m_5 and m_6. In such a case, some existing trajectory similarity algorithms such as LCSS, DTW, ED are not suitable. So we design a trajectory similarity measurement method to deal with CDR trajectories, which is shown in Algorithm 1.

```
Algorithm 1: Trajectory Similarity
Input: $trajectory\ T_a, trajectory\ T_b, \delta_t, \delta_d$
Output: $Tsim(a, b)$
function $TraSimilarity(T_a, T_b)$ begin
    $v \leftarrow 0, w \leftarrow 0;$
    $start_i \leftarrow 0, start_j \leftarrow 0, end_i \leftarrow 0, end_j \leftarrow 0;$
    for each $sp_{a,i} \in T_a, sp_{b,j} \in T_b$ do
        if $|sp_{a,i}.t - sp_{b,j}.t < \delta_t|$ and $dis(sp_{a,i}, sp_{b,j}) < d_{th}$ then
            if $i \notin M_a$ and $j \notin M_b$ then
                add $i$ to $M_a$, add $j$ to $M_b$;
                $end_i \leftarrow i, end_j \leftarrow j;$
                $subT_a \leftarrow sp_{start_i}$ to $sp_{end_i};$
                $subT_b \leftarrow sp_{start_j}$ to $sp_{end_j};$
                if $Measure(subT_a, subT_b) = true$ then
                    $w \leftarrow w + 1;$
                end
                $v \leftarrow v + 1;$
            end
        end
        $start_i \leftarrow end_i,\ start_j \leftarrow end_j;$
    end
    return $w/v;$
end
function $Measure(subT_a, subT_b)$ begin
    $Flag \leftarrow false;$
    $C_a \leftarrow\ Centroid(subT_a);$
    $C_b \leftarrow\ Centroid(subT_b);$
    if $Distance(C_a, C_b) < \delta_d$ then
        $Flag \leftarrow true;$
    end
    return $Flag;$
end
```

The core concept of the algorithm is to divide an entire trajectory into sub-trajectories by the matching points on two trajectories, then measure the similarity of two trajectories based on the distance between centroids of two sub-trajectories.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 74; doi:10.3390/ijgi8020074

Definition 4. *(Matching Point):Given trajectory T_a and T_b, sp_i and sp_j are the stay point of T_a and T_b respectively. δ_t is a time threshold. sp_i and sp_j are called matching points if:*
(1) $|sp_i.t - sp_j.t| < \delta_t$
(2) $dis(sp_i, sp_j) < d_{th}$
where $dis()$ is the distance of two points.

The algorithm consists of two functions. The function TraSimilarity (Line 1-20) is to find the matching points and then divide trajectories into sub-trajectories by the matching points. M_a and M_b are the set of the matching points in T_a and T_b. The function Measure (Line 21-29) is to judge if the two sub-trajectories are similar by calculating the distance between two centroids of sub-trajectories. First, we try to find the matching points of the two trajectories (Line 5). To avoid the situation that one point can be matched with two different points in another trajectory, we choose the first matching point in another trajectory as the start of the sub-trajectory (Line 6-7). To measure similarity of sub-trajectories with different sampling rate, we merge stay points on the sub-trajectory into a centroid and then use the distance between two centroids to estimate the similarity of sub-trajectories (Line 25-27). In this way, the issue caused by different sampling rate can be addressed somehow. When the distance between centroids of sub-trajectories is within δ_d, two tourists are considered as travel companion in the sub-trajectory. Finally, the trajectory similarity of tourist a and b is denoted as:

$$Tsim\,(a,b) = w/v \tag{2}$$

where v is the number of sub-trajectories in the entire trajectories and w is the number of sub-trajectories on which tourist a and b are considered to be traveling companion.

3.4.2. Accommodation Similarity

In the process of measuring similarity of tourists, we consider not only the trajectories of tourists, but also the places they stayed at night. Group movement patterns for tourists have distinct characteristics compared with other kinds of group movement patterns. For example, in the peak season, thousands of tourists crowd to famous scenic areas in Hainan at the same period, which leads to overlapped trajectories of different tourist groups in the daytime. So it's hard to distinguish different tourists groups only by trajectory data. We try to find places tourists stayed at night to measure their similarities in accommodations. Generally speaking, a group of tourists will stay in the same place (maybe a hotel or a residence) at night which can be an important feature to measure the similarity of tourists in a group.

The first step is to identity lodgings tourists stayed at each night. We define 21:00 to 9:00 as *Hometime*. It is obvious that tourists will spend most time in lodgings at Hometime. In the algorithm, we try to find the stay points with the longest duration at Hometime, and identified them as tourists' lodgings. Supposing the length of stay for tourist a and b in Hainan is z nights, his/her lodgings in Hainan are denoted as a sequence of lodgings with date attached, i.e. $H_a = \{h_{a1}, h_{a2}, ..., h_{az}\}$. So we define the accommodation similarity of tourist a and b as

$$Asim\,(a,b) = samelod/z \tag{3}$$

where $samelod$ denotes the number of same lodgings during z nights.

3.4.3. The Similarity of other features

Besides similarity measurements mentioned above, there are also other features related to travel behaviors that can be used to measure similarity. In this part, we extract other two features from tourists which can help us identify the relationship between tourists in the groups, then combine the two features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

The first feature is the mobile phones' registration locations of tourists, which can help to discover groups from a certain province. In CDRs, each record has a field to indicate the mobile phones' registration locations, for example, "301" represents Guangdong Province, "302" represents Shandong Province and so on. Utilizing this we can easily discover tourists from the same province. Two tourists traveling together within the same tourist group are more likely to be from the same province.

The second feature is the number of days tourists spent in Hainan which is also an important feature to distinguish different tourist groups. Generally, the members' arrival and departure time in a group is usually consistent and the days they spent in Hainan will also be the same. We calculate the maximum continuous days in Hainan for each tourist as his/her second feature.

For tourist a and b, when the feature of tourist a and b has the same value, this feature is considered to be matched. We measure the similarity of the two features of tourists a and b by this equation:

$$Nsim\left(a,b\right) = matfeas/allfeas \tag{4}$$

where $allfeas$ is the total number of features. Here the value of $allfeas$ is 2. The $matfeas$ is the number of matches of a and b.

3.5. Identify Group Tourists

After obtaining the similarity of tourist a and b which is denoted as $Sim(a,b) = (Tsim, Asim, Nsim)$, we need to judge whether a and b is a pair of traveling companions or not by the similarity vector.

We use two different methods to determine which pairs of tourists are the traveling companions. The first method is to set a threshold to filter out the tourists who have low similarity with others in candidate groups. We define $totalsim = w_1 * Tsim + w_2 * Asim + w_3 * Nsim$, where w_1, w_2, w_3 are the weight of three features. $w_1 + w_2 + w_3 = 1$ and $w_1, w_2, w_3 \in [0,1]$. If $totalsim > \delta_s$, the two tourists are identified as traveling companions, otherwise not. Because of the complexity of travel behaviors, it is not so easy to choose the proper value of w_1, w_2, w_3, δ_s. We set two groups of value in our work to analyze.

In the second method, we apply S4VMs, a semi-supervised learning algorithm proposed by [26], after labeling some pairs of tourists manually, to identify the traveling companions. The algorithm uses unlabeled data to improve the performance of classification results when labeled data are limited.

Traveling companions that are in the same candidate group are identified as a real group. For example, in a candidate group $\{a, b, c, d\}$, we find two pairs of traveling companions, namely a, b and b, c, then we consider the real group as $\{a, b, c\}$.

4. Experiment and discussion

4.1. Experiment setup

Data set: In this paper, we use anonymized Call Detail Records (CDRs) and location-based data provided by one of the largest telecom operators in China. The data set contains more than 10 millions anonymized mobile phone records in Hainan province in December 2015. To deal with the large-scale data, we use Apache Spark, a cluster computing framework for big data processing.

Parameter setting: In the data preprocessing, the time threshold d and distance threshold t for stay points identifying is set to be 200 meters and 10 minutes respectively. In the process of frequent groups mining, the interval of snapshots TI is set to be 30 minutes, which is long enough to ensure that the trajectory points of the tourists in the same group can be included in the same snapshots. The minimum size of candidate group M is 2 and the minimum snapshots for occurrence K is 4 which can avoid many accidental meeting events in scenic spots. In the process of trajectory similarity measurement, the time threshold of matching points δ_t is set to be 5 minutes and the distance threshold between centroids of sub-trajectories δ_d is set to be 1 kilometer. In the process of identifying group

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 74; doi:10.3390/ijgi8020074

tourists, we choose two groups of weights, w_1, w_2, w_3 ,namely weight1, {0.33,0.33,0.33} and weight2, {0.5,0.25,0.25}, to examine the effects of different features when calculating the similarity of the tourist pairs. The distance threshold of similarity δ_s is set to be 0.5.

Results: By performing frequent itemsets mining on the collections generated from CDRs, we got 50471 frequent itemsets and 43894 of them are closed itemsets. For each candidate group, we calculate the similarity of each pair of the tourists in the group.

Table 2. Number of groups of two methods

group size	weight 1	weight 2	S4VMs	overlap
2	1624	2320	3088	1255
3	173	331	255	151
4	61	79	88	59
5	34	35	49	33
6	16	19	21	16
7	5	6	6	6
8	2	2	2	2

In order to identify real tourists groups from candidate groups, we picked 200 of them as the training set and added labels manually. Then we used the threshold-based method and S4VMs algorithm to get real groups from candidate groups. The results of different methods are shown in Table 2. It can be seen that the size of groups in the result is not large, ranging from 2 to 8. Considering that the CDR data set comes from the telecom operator which market share is less then 20%, the actual group size may vary from about 10 to 40. The threshold-based method can identify more groups when the weight of trajectory similarity is set to be 0.5, compared to 0.33. It indicates that trajectory similarity is an important feature in group identifying. The S4VMs can identify more tourist groups than the threshold-based method and more groups in small size can be discovered by S4VMs. The column overlap shows the number of groups which are identified by both three methods. For larger group size, more groups are overlapped for different methods. It can also be noted that, with the aims of getting better understanding of the group movement behaviors, the proposed group identification method is relatively tight. In addition, only when more than one member in a group whose phone number belongs to this telecom operator, a group can be identified. So the number of discovered groups are smaller than expected, especially for large group size.

Since it is difficult to obtain the ground truth of the identified groups, we validated the results of the proposed methods in an indirect way. We first picked out the groups which traveled from Haikou city to Sanya city from the identified groups. And then the transportation mode of each group member is detected. If the members in the same group have the identical transportation modes during their journey from Haikou to Sanya, they are very likely to be traveling companions.

There are four major transportation routes between the two cities, namely the eastern expressway, the middle expressway, the western expressway and the railway. By calculating the similarity between tourist trip routes and the four major transportation routes, we can identify the transportation mode of each tourist.

In Table 3, the percentage of groups which members share the same transportation modes is shown. It can be seen that, for the group size larger than 2, the groups identified by S4VM can achieve 100% matching ratio, which means all the group members have the same transportation mode.

4.2. Travel behaviors analysis

With the results of tourist groups identified by our algorithm, we analyze the travel behaviors of group tourists and individual tourists, such as travel routes, travel time and so on.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

Table 3. Matching ratio of transportation mode

Group size	weight={0.33,0.33,0.33}	weight={0.5,0.25,0.25}	S4VMs
2	0.81	0.77	0.91
3	0.89	0.87	1.00
4	0.96	0.93	1.00
5	0.96	0.89	1.00
6	1.00	0.94	1.00
7	1.00	1.00	1.00
8	1.00	1.00	1.00

4.2.1. Time to visit the scenic spot

The time that tourists arrived at scenic areas is shown in Figure 4. The differences between group tourists and individual tourists can be seen clearly. More group tourists visited scenic areas before 9 o'clock than individual tourists. And in the afternoon, more individual tourists visited scenic areas than group tourists. This is reasonable because travel agencies often start their tours in the early morning, but individual tourists have a more flexible schedule so they are more likely to visit scenic areas in the afternoon.

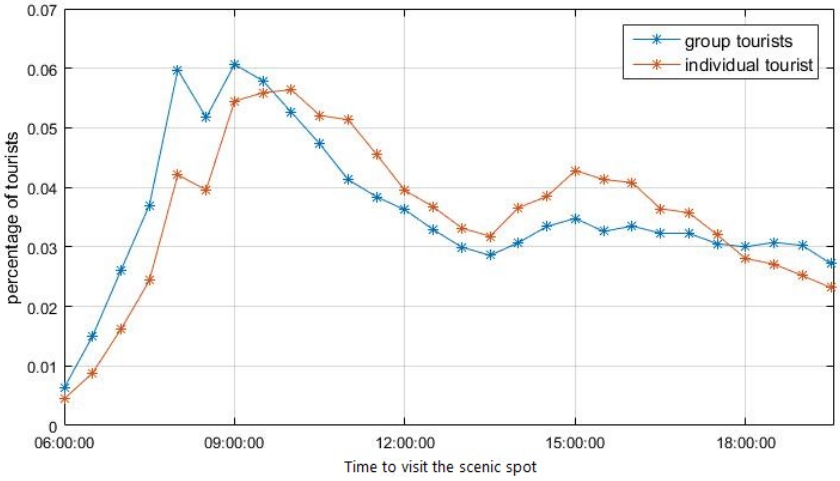


Figure 4. Time to visit the scenic spot for group tourists and individual tourists

Figure 5 exhibits the average time spent in different scenic ares by group tourists and individual tourists. The scenic areas and the corresponding IDs are exhibited in Table 4. It can be seen that group tourists spent less time than individual tourists in most scenic areas, because the tourist groups often have a tight schedule so that they have much shorter visiting time at each location.. But in Capital Outlets, a famous shopping mall, the average time spent by group tourists are longer than individual tourists.

4.2.2. Trip distance

The trip distance is another importance aspect of tourist behaviors. The trip distance of tourists is defined as the cumulative distance of their travel routes in Hainan province. We calculate trip distance of group tourists and individual tourists from extracted trajectory data. The results are shown in Figure 6. The group tourists represented in blue have a larger average trip distance than the individual tourists shown in red curve. This can be explained that group tourists are likely to visit more places. Individual tourists prefer to stay at one area to relax.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

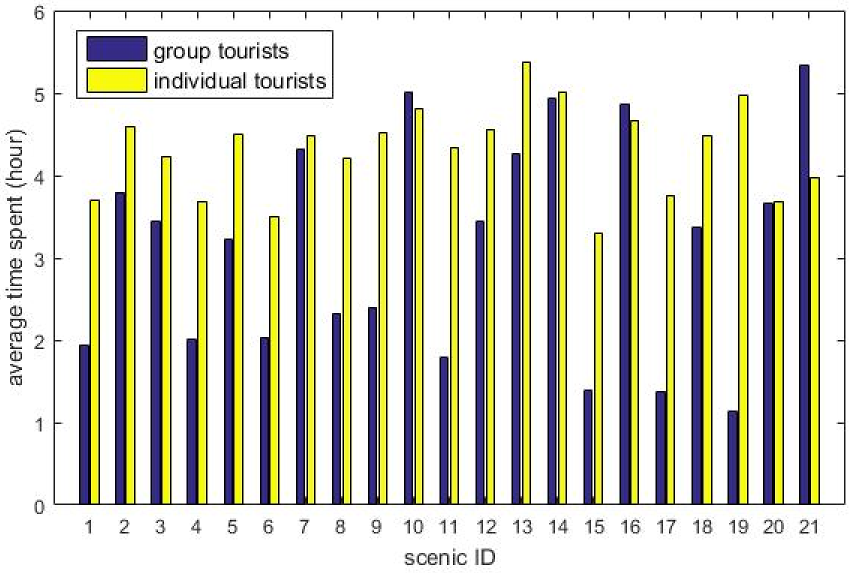


Figure 5. Average time spent in different scenic areas.

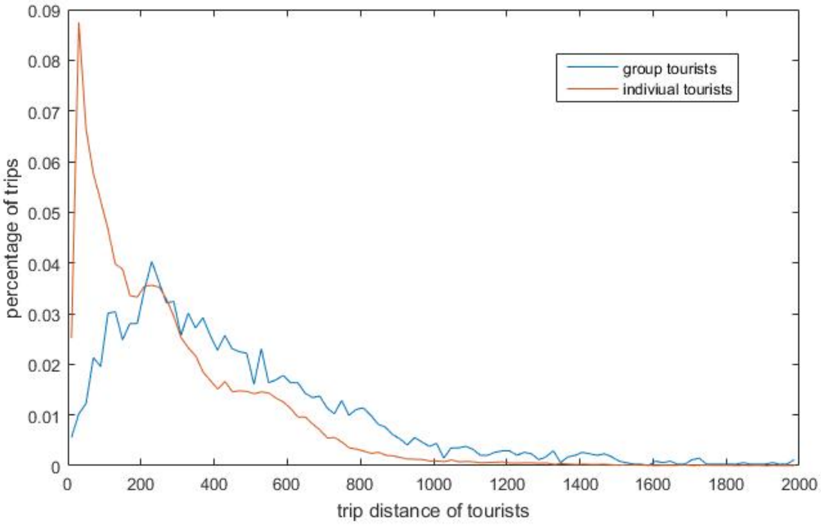


Figure 6. The PDF plot of trip distance. X-axis is the trip distance (kilometers) of tourists in their tour.

4.2.3. Origin and destination

Besides the mobility of tourists, we are also interested in spatial distribution of tourists. We calculate the total group number of tourists from different provinces and the result is shown in Figure 7. We found that Beijing, Sichuan and Heilongjiang are the top 3 provinces with the largest group of tourists.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 74; doi:10.3390/ijgi8020074

Table 4. Scenic area identifications and corresponding names.

Id	Scenic area	Id	Scenic area
1	Nanshan Cultural Zone	12	Permanent Site of Boao
2	Daxiaodongtian	13	YaLong Bay
3	Yanuoda Rain Forest	14	Dadonghai
4	Fenjiezhou Island	15	Nanwan Houdao Island
5	Volcanic Cluster Geopark	16	Mission Hills Haikou
6	Binlanggu	17	Hainan Wenbi Mountain
7	Holiday Beachside Resort	18	Sanya Xidao
8	Tianyahaijiao	19	Dongshan Ridge
9	Tropical Garden of Fauna	20	Sanya Duty Free Shop
10	Wuzhizhou Island	21	Capital Outlets
11	Xinglong Botanical Garden		

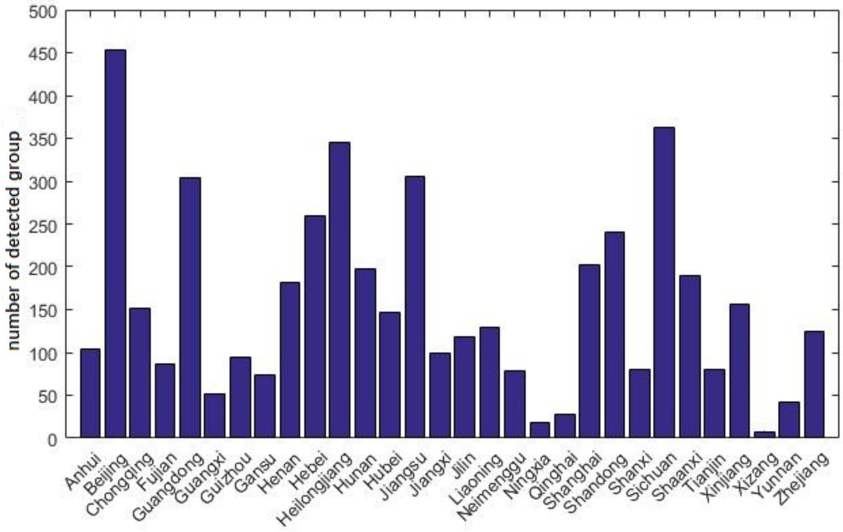


Figure 7. Geographical distribution of group tourists, X-axis is provinces, Y-axis is the number of detected group from each province.

In addition to that, we are interested in the Origin-Destination matrix, which describes the tourists flow and the different distributions in each scenic areas. In Figure 8, the origin indicates the top 10 provinces which have the greatest number of group tourists, and the destination is the scenic areas they visited. We can observe the top origins and destinations, as well as the trip distributions. Tourists from Sichuan and Beijing preferred to visit YaLong Bay and Dadonghai than other scenic areas. But few tourists visit Tropical Garden of Fauna and Dongshan Ridge.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

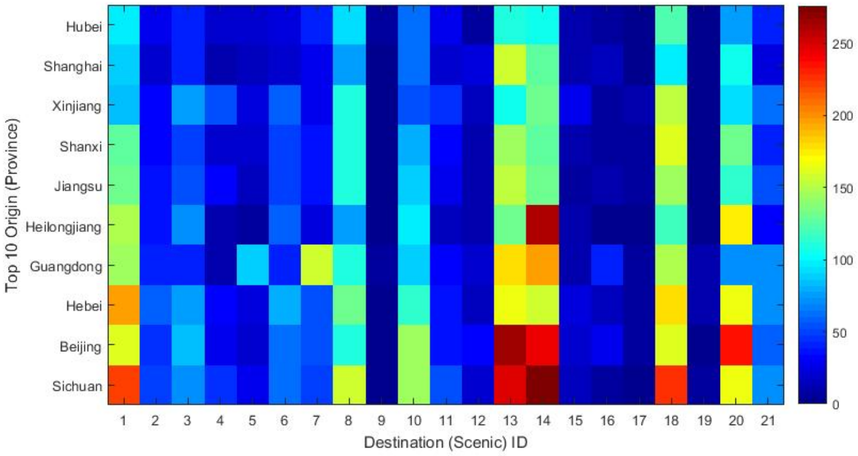


Figure 8. OD matrix of group tourists. X-axis is the scenic ID of scenic areas and Y-axis is top 10 provinces with the most group tourists. (Top 1 is Sichuan, Top 10 is Hubei).

4.2.4. Popular tourist routes

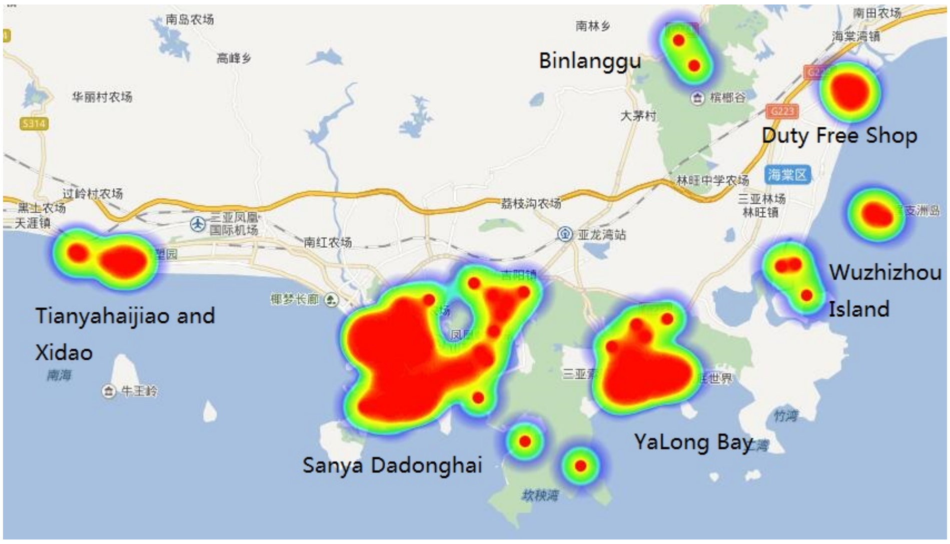


Figure 9. The distribution of tourists in Sanya.

To analyze the different travel preferences of group tourists and individual tourists, we try to mine the most popular scenic areas and popluar tourist routes in Sanya. In Figure 9, it can be seen that Dadonghai and YaLong Bay are the most popular scenic areas in Sanya.

The popular patterns of tourists are shown in Table 5. The most popular route of group tourists is Yalong Bay⟶Dadonghai⟶Duty Free Shop. The top 2 route is Tianyahaijiao⟶Sanya Xidao⟶Dadonghai. The top 3 route is Tianyahaijiao⟶Sanya Xidao⟶Nanshan Cultural Zone. Different from group tourists, the most popular route of individual tourists is Tianyahaijiao⟶Sanya Xidao⟶Dadonghai. The top 2 route is Tianyehaijiao⟶Nanshan Cultural Zone⟶YaLong Bay. The results indicate that group tourists are more likely to go shopping than individual tourists. And individual tourists are likely to choose to visit natural scenic areas.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2019, 8, 74; doi:10.3390/ijgi8020074

Table 5. Top 10 popular routes with three scenic areas

Id	Group tourists	Individual tourists
1	13-14-20	8-18-14
2	8-18-14	8-1-13
3	8-18-1	8-18-13
4	1-13-14	20-13-14
5	1-20-14	21-8-18
6	10-20-14	10-8-18
7	10-1-14	1-13-14
8	6-8-18	18-13-14
9	21-8-18	1-8-14
10	3-8-18	11-8-13

5. Conclusions

In this work, we tried to identify tourists groups and analyze behaviors of tourist groups using CDRs and location-based information. Though the mobile data is poor in spatial resolution with low sampling rate, it has a much larger scale and costs less to obtain. Considering the the complexity of tourist groups, we propose a method called GMPMS to identify tourist groups using spares CDRs. The algorithm discovers groups not only by trajectories but also by the accommodations tourists stayed at night and other features that can reflect the behaviors of tourists. In the experiments, we found more than 3500 groups from a big amount of anonymized mobile records. Then we analyzed the difference between group tourists and individual tourists from several aspects such as trip distance, average travel time and popular tourist routes. The results show that the two kinds of tourists have different travel behaviors and preferences which may benefits tourism in understanding behaviors of tourists in an insightful way.

Author Contributions: Xinning Zhu, Tianyue Sun, and Hao Yuan designed the framework and Tianyue Sun contributed to the group identification method. Hao Yuan and Tianyue Sun contributed to data processing. The group tourist exploration and the analysis of travel behaviors were conducted by Hao Yuan and Tianyue Sun. The paper was reviewed by Xinning Zhu. Finally, the project administration and resources were provided by Zheng Hu and Jiansong Miao.

Funding: This research received no external funding.

Acknowledgments: This work was supported by the Major Science and Technology Plan Project of Hainan Province, ZDKJ201808.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tsai, H.P.; Yang, D.N.; Chen, M.S. Mining group movement patterns for tracking moving objects efficiently. *Ieee T Knowl Data En 2011* **2011**, *23*, 266-281.
2. Zhou, Y.; Zhang, Y.; Ge, Y.; Xue, Z.; Fu, Y.; Guo, D.; Shao, J.; Zhu, T.; Wang, X.; Li, J. An efficient data processing framework for mining the massive trajectory of moving objects. *Computers, Environment and Urban Systems* **2017**, *61*, 129-140.
3. Gudmundsson, J.; Van Kreveld, M. In Computing longest duration flocks in trajectory data, 14th Annual ACM International Symposium on Advances in Geographic Information Systems, ACM-GIS'06, November 6, 2006 - November 11, 2006, Arlington, VA, United states, 2006; Association for Computing Machinery: Arlington, VA, United states, pp 35-42.
4. Jeung, H.; Yiu, M.L.; Zhou, X.; Jensen, C.S.; Shen, H.T. Discovery of convoys in trajectory databases. *Proceedings of the VLDB Endowment* **2008**, *1*, 1068-1080.
5. Li, Z.; Ding, B.; Han, J.; Kays, R. Swarm: Mining relaxed temporal moving object clusters. *Proceedings of the VLDB Endowment* **2010**, *3*, 723-734.
6. Naserian, E.; Wang, X.; Xu, X.; Dong, Y. A framework of loose travelling companion discovery from human trajectories. *IEEE Transactions on Mobile Computing* **2018**, *17*, 2497-2511.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 74; doi:10.3390/ijgi8020074

7. Liu, S.; Wang, S. Trajectory community discovery and recommendation by multi-source diffusion modeling. *IEEE Transactions on Knowledge and Data Engineering* **2017**, *29*, 898-911.
8. Tang, L.-A.; Zheng, Y.; Yuan, J.; Han, J.; Leung, A.; Hung, C.-C.; Peng, W.-C. In On discovery of traveling companions from streaming trajectories, IEEE 28th International Conference on Data Engineering, ICDE 2012, April 1, 2012 - April 5, 2012, Arlington, VA, United states, 2012; IEEE Computer Society: Arlington, VA, United states, pp 186-197.
9. Zheng, K.; Zheng, Y.; Yuan, N.J.; Shang, S.; Zhou, X. Online discovery of gathering patterns over trajectories. *Ieee T Knowl Data En* **2014**, *26*, 1974-1988.
10. Zheng, Y. Trajectory data mining: An overview. *Acm T Intel Syst Tec* **2015**, *6*.
11. Tang, L.-A.; Zheng, Y.; Yuan, J.; Han, J.; Leung, A.; Peng, W.-C.; Porta, T.L. A framework of traveling companion discovery on trajectory data streams. *Acm T Intel Syst Tec* **2013**, *5*.
12. Wang, Y.; Luo, Z.; Xiong, Y.; Prosser, D.J.; Newman, S.H.; Takekawa, J.Y.; Yan, B. In Discovering loose group movement patterns from animal trajectories, 11th IEEE International Conference on eScience, eScience 2015, August 31, 2015 - September 4, 2015, Munich, Germany, 2015; Institute of Electrical and Electronics Engineers Inc.: Munich, Germany, pp 196-206.
13. Zhang, C.; Han, J.; Shou, L.; Lu, J.; Porta, T.L. Splitter: Mining finegrained sequential patterns in semantic trajectories. *Proceedings of the VLDB Endowment* **2014**, *7*, 769-780.
14. Toohey, K.; Duckham, M. Trajectory similarity measures. *Sigspatial Special* **2015**, *7*, 43-50.
15. Magdy, N.; Sakr, M.A.; Mostafa, T.; El-Bahnasy, K. In Review on trajectory similarity measures, 7th IEEE International Conference on Intelligent Computing and Information Systems, ICICIS 2015, December 12, 2015 - December 14, 2015, Cairo, Egypt, 2015; Institute of Electrical and Electronics Engineers Inc.: Cairo, Egypt, pp 613-619.
16. Nakamura, T.; Taki, K.; Nomiya, H.; Seki, K.; Uehara, K. A shape-based similarity measure for time series data with ensemble learning. *Pattern Analysis and Applications* **2013**, *16*, 535-548.
17. Liu, H.; Schneider, M. In Similarity measurement of moving object trajectories, 3rd ACM SIGSPATIAL International Workshop on GeoStreaming, IWGS 2012, November 6, 2012 - November 6, 2012, Redondo Beach, CA, United states, 2012; Association for Computing Machinery: Redondo Beach, CA, United states, pp 19-22.
18. Wang, F.; Zhu, X.; Miao, J. Semantic trajectories-based social relationships discovery using wifi monitors. *Personal and Ubiquitous Computing* **2017**, *21*, 85-96.
19. Xue, M.; Wu, H.; Chen, W.; Ng, W.S.; Goh, G.H. In Identifying tourists from public transport commuters, 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2014, August 24, 2014 - August 27, 2014, New York, NY, United states, 2014; Association for Computing Machinery: New York, NY, United states, pp 1779-1788.
20. Vu, H.Q.; Li, G.; Law, R.; Ye, B.H. Exploring the travel behaviors of inbound tourists to hong kong using geotagged photos. *Tourism Management* **2015**, *46*, 222-232.
21. Yang, L.; Wu, L.; Liu, Y.; Kang, C. Quantifying tourist behavior patterns by travel motifs and geo-tagged photos from flickr. *Pervasive and Mobile Computing* **2015**, *18*, 18-39.
22. Sun, Y.; Li, M. Investigation of travel and activity patterns using location-based social network data: A case study of active mobile social media users. *ISPRS International Journal of Geo-Information* **2015**, *4*, 1512.
23. Phithakkitnukoon, S.; Horanont, T.; Witayangkurn, A.; Siri, R.; Sekimoto, Y.; Shibasaki, R. Understanding tourist behavior using large-scale mobile sensing approach: A case study of mobile phone users in japan. *Pervasive and Mobile Computing* **2015**, *18*, 18-39.
24. Wu, W.; Wang, Y.; Gomes, J.B.; Anh, D.T.; Antonatos, S.; Xue, M.; Yang, P.; Yap, G.E.; Li, X.; Krishnaswamy, S., et al. In Oscillation resolution for mobile phone cellular tower data to enable mobility modelling, 15th IEEE International Conference on Mobile Data Management, IEEE MDM 2014, July 15, 2014 - July 18, 2014, Brisbane, QLD, Australia, 2014; Institute of Electrical and Electronics Engineers Inc.: Brisbane, QLD, Australia, pp 317-324.
25. Han, J.; Pei, J.; Yin, Y. In Mining frequent patterns without candidate generation, 2000 ACM SIGMOD - International Conference on Management of Data, May 16 - May 18 2000, Dallas, TX, United states, 2000; Dallas, TX, United states, pp 1-12.
26. Li, Y.-F.; Zhou, Z.-H. Towards making unlabeled data never hurt. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2015**, *37*, 175-188.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2018 doi:10.20944/preprints201812.0056.v1

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 74; doi:10.3390/ijgi8020074

27. Zheng, K.; Zheng, Y.; Yuan, N.J.; Shang, S. In On discovery of gathering patterns from trajectories, 29th International Conference on Data Engineering, ICDE 2013, April 8, 2013 - April 11, 2013, Brisbane, QLD, Australia, 2013; IEEE Computer Society: Brisbane, QLD, Australia, pp 242-253.
28. Yan-Wei, Y.U.; Jian-Peng, Q.I.; Yun-Hui, L.U.; Zhao, J.D.; Zhang, Y.G. Distributed swarm pattern mining algorithm in big spatio-temporal trajectory data. *Computer Engineering Science* **2016**.
29. Zhang, C.; Zhang, K.; Yuan, Q.; Zhang, L.; Hanratty, T.; Han, J. In Gmove: Group-level mobility modeling using geo-tagged social media, 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, August 13, 2016 - August 17, 2016, San Francisco, CA, United states, 2016; Association for Computing Machinery: San Francisco, CA, United states, pp 1305-1314.
30. Chen, X.C.; Faghmous, J.H.; Khandelwal, A.; Kumar, V. In Clustering dynamic spatio-temporal patterns in the presence of noise and missing data, 24th International Joint Conference on Artificial Intelligence, IJCAI 2015, July 25, 2015 - July 31, 2015, Buenos Aires, Argentina, 2015; International Joint Conferences on Artificial Intelligence: Buenos Aires, Argentina, pp 2575-2581.
31. Banerjee, P.; Ranu, S.; Raghavan, S. In Inferring uncertain trajectories from partial observations, 14th IEEE International Conference on Data Mining, ICDM 2014, December 14, 2014 - December 17, 2014, Shenzhen, China, 2015; Institute of Electrical and Electronics Engineers Inc.: Shenzhen, China, pp 30-39.
32. Lan, R.; Yu, Y.; Cao, L.; Song, P.; Wang, Y. In Discovering evolving moving object groups from massive-scale trajectory streams, 18th IEEE International Conference on Mobile Data Management, MDM 2017, May 29, 2017 - June 1, 2017, Daejeon, Korea, Republic of, 2017; Institute of Electrical and Electronics Engineers Inc.: Daejeon, Korea, Republic of, pp 256-265.
33. Wu, G.; Ding, Y.; Li, Y.; Bao, J.; Zheng, Y.; Luo, J. In Mining spatio-temporal reachable regions over massive trajectory data, 33rd IEEE International Conference on Data Engineering, ICDE 2017, April 19, 2017 - April 22, 2017, San Diego, CA, United states, 2017; IEEE Computer Society: San Diego, CA, United states, pp 1283-1294.