ABSTRACT

In this paper we present the results of using visual analysis tool to analyze and visualization human movement data. The analysis is done based on data of a group of 3557 visitors’ location record. We use space-time cube to construct a 3D route in which we can not only observe the spatial movement of visitors, but also their dynamic movement in time dimension as time flow. Among datasets we figure out facility that visitor need check-in to visit (which revealed by type of data) and treat them as significant places, then we connected them to complete the construction of personal dynamic route. By checking the overlap of different dynamic route we consider visitor moving and visiting same facility together as small group, and using same method we measure the similarity of every two different dynamic route in a parameter called distance function, which is based on the interval of time if a same significant place is visited by more than one group of visitors. Once calculations of distance function among different groups completed we got a matrix that reveals the similarity of every two different dynamic route. Then we applied a clustering algorithm called Hierarchical Clustering to group data over a variety of scales by creating a binary cluster tree. In the tree clusters at one level are joined as clusters at the next level, and the height of level when two clusters joined each other presents the how exactly the similarity are between these clusters. By setting different level we can group all the movement data into clusters of different size and similarity, and by adjusting the level observers are able to observe clusters in certain size or similarity as they wish. Additionally, we can also find out potential outlier which share little similarity with all other visitors simply by paying attention to those cluster join others at an extremely high level.

1. INTRODUCTION

Movement data of human, animals, vehicles, even hurricanes and wildfire help us evacuate crowded in advance and do rescue in time. Data help us understand our environment better and enjoy our life better.

Movement data of human are as well as important, if not more important than others. Large scale of people flow has a significant influence on population. For example, the monitor of Syria refugee cross border during the war help people realize how serious the war is, and how many refugee may need help from humanitarian organization.

Movement data are once hard to get, need complex device to transfer data. But since the development of GPS technology, available low cost movement data can be easily got by using location app on personal mobile terminal. These mass of data can provide us with useful information, but the hiding knowledge can be too abstract to get if we analyze them directly. Visual analysis offer us a solution to observe those data in a more direct and understandable way.

In this project we observe a smaller scale of human movement, but the method we used can also apply to groups of larger scale and movement during a longer period of time. The data we used are from VAST challenge 2015, which record movement data of visitor in a amusement park during a weekend. The park DinoFun World is a modest-sized amusement park typically hosting thousands of people each day. Figure 1 show a geographic map of DinoWord, showing major facilities in the amusement park and roads connected them. One event last year was a weekend tribute to a famous international renowned football player. During the week a famous football player is scheduled to appear in two stage shows each on Friday, Saturday, and Sunday to talk about his life and career. During the weekend several other shows of memorabilia related to his illustrious career would also be displayed, as well as other major amusement facilities that would be visited by thousands of visitors.
By using visual analysis tools to analyze movement data of visitors we are going to cluster them into different groups, visitor in each groups are going to have more similarity about their choice of facilities visited also the order and time they choose to visit them. By clustering visitors in to smaller group with certain movement pattern we are more easily to observe, analysis, understand or even predict how they move, the pattern may be of great usage when dealing with issue like improving transfer efficiency. Also by finding out visitor have little similarity with others we may reveal outlier that may deal potential damage to amusement park.

2. RELATED WORK

In the analysis of movement data visualization is used for a better understanding of data and underlying phenomena. The most traditional tool used in visual analysis is flow line drawn on a map. The flow lines are useful to reveal certain geographic pattern of single person, exactly shoe number of different locations present in the original data, but get heavily overlap when dealing with large amount of movement data. Also it do nothing to presents the dynamic pattern of movement, which means it’s useless when time dynamic is important to decide the differences between different patterns. Another tools called flow map is useful when we need to visualize numbers of entities or volumes of materials that moved from one place to another, it does not reflect the exactly location people visited but show cumulative movements that occurred during a certain time period. Flow map do nothing with dynamic either, but can be extended to animated flow maps which use a series of flow maps showing how the flows change over time.

In some situation dynamic is more important than spatial pattern because there are little differences of spatial significant places people visited. For example groups of people can all visit same significant places but in different order, or one person could visit one places multiple times while others never go back after one visit. These all leads to same geographic pattern but dynamic patterns show much differences. Traditional visual analysis tools pay little attention to dynamic pattern for the data they analyze seldom overlap with each other so spatial pattern is more important, in area which dynamic counts other visual analysis tools like space-time cube is wildly used to reveal the dynamic pattern of certain person. Space-time cube uses a 3D line graph to present how the location changes as time flow. While using two dimensions to represent geographical space like other visual tools, a third dimension is imported to represents time, movement behavior of single person is revealed as a three-dimensional line connecting places of significant places. The line also presents other information like speed of movement and acceleration[1].

In our project we mainly use space-time cube to presents the dynamic route of every single group of visitors. We also use the three-dimensional line to measure the similarity of different group, then use the distance function to cluster all visitors.

3. DATASETS

The dataset we used are offered from VAST challenge 2015, in which a group of thousands visitor visit an amusement park named DinoFun Amusement park during a weekend (from 6/6/2014 to 6/8/2014). The running time of the amusement park is from 8:am to midnight, information of personal location are recorded by a map app that applied in every visitor’s personal devices. An additional service are offered and work as others if any visitor doesn’t hold a personal device which could run the location app. The app also work as e-ticket so visiting any facility that requires a ticket in the amusement park leave a record in the datasets too, which show as check-in type data. If not in any facility the location data are updated every 10 seconds to reveal your movement from one facility to other, staying around with your friend or simply waiting in a line. The datasets are stored in csv format as figure 2.
1. However, ‘check-in’ are useful for much of how people think can be learned through them. However, we abandon ‘movement’ datasets for reasons as follows:

   1. There is usually only one closest approach people prefer to choose when moving from one facility to another. For situations there is more than one choice which leads to different geographic patterns, choose one approach means spend more or less time than another.

   2. People hang around together may have some differences of ‘movement’ datasets, but they should still be consider as having same pattern. For example a group of 10 people are visiting the amusement park together, they divide into 2 smaller group, one go to roller coaster directly to check the line while another buy some drink first before they join each other to check-in together. There is some spatial pattern difference exists but they are not important as long as they have same motivation and check-in datasets.

   3. While facilities need check-in have solid coordinate, places people like to stay but record as ‘movement’ datasets are not solid. For example facility like roller coaster can only be recorded as (check-in, 66, 47) when people visited, but a small park people going to rest can be recorded as (movement 44, 56), (movement 46, 55) or any coordinate between them. These datasets are regarded as different places, people visit them will also be mistakenly clustered as different group.

   4. Clustering with over 8,000,000 datasets is expensive, we are going to compare every two different routes in following steps so the comparison algorithm is even more expensive. Using pure ‘check-in’ date means dealing with datasets around 100,000 but with enough accuracy.

Considering all inconveniences we are only going to use ‘check-in’ data-set as significant point when we construct 3D dynamic route. All ‘check-in’ datasets are revealed as scatter in 3D space-time cube, and for every single person their points are connected with line. The height of point presents the time stamp they are visited, details like visit same places in succession or return for one places with overlapped spatial pattern can all be observed using 3D dynamic route. An example of dynamic route are shown as Figure 3.

There are 4 parts consist a complete dataset. Time like ‘2014-6-07 08:00:10’ presents the exactly timestamp when this datasets was record. Timestamp are continuous even if you are not moving, as long as you are not in any facility. Number like ‘1102394’ presents ID number of a certain visitor. Every ID number is bound to an e-ticket so every visitor can only have one ID number and one ID number only presents one visitor. Type of datasets are either ‘check-in’ or ‘movement’, while type ‘check-in’ are recorded every time the visitor visit any facility that need a check-in and no other data will be recorded as long as you stayed in the facility, ‘movement’ datasets are recorded once you are not in any facility, even you are not moving. Places visitor visit and stay but do not need a check-in like small shop visitor can have a snack or a park everyone can have a nap are also record as ‘movement’ when visited. The last part of data is the exactly location of a certain time stamp in format as coordinate. Facilities which need e-tickets have solid location so coordinate of ‘check-in’ data type places are restricted to 40 different arrays. Facilities have multiple entrances (like 30) also get different ‘check-in’ data for each entrance. We treat them as different places for we are going to consider visitor groups chose different entrances as different clusters.

4. CONSTRUCTION OF DYNAMIC ROUTE

We use datasets of Friday as an example to settle visual analysis. There are total over 8,000,000 datasets recorded during Friday, among them 100,000 are ‘check-in’ datasets and others are ‘movement’ datasets. To complete the construction of dynamic route we should decide of which data we are going to use and of which we are not. Our goal of visual analysis is to observe the movement pattern of different cluster of people so we can understand the meaning under these pattern, thus datasets of ‘check-in’ are useful for much of how people think can be learned through them. However we abandon ‘movement’ datasets for reasons as follows:

Figure 2, Example of datasets in csv format
In the 3D space-time cube, X axes and Y axes present the geographic location of every significant palace, while the Z axes presents the time stamp when the place is visited. The time of beginner check-in points are all set to 0, as everyone needs to check-in at entrances first when they enter the amusement park the first check-in point for all people can only be places of entrances, this set bring much convenience for observation. We can transfer every point into an array $P_i,k(x,y,t)$, $P_i,k$ presents one visit to significant point $k$ for visitor with id $i$, the coordinate presents the location $(x,y)$ and time stamp $t$ when the place is visited.

Using time-space cube to visualize movement data give us a much more directly observation of movement pattern. We could clearly see how people move from significant place to another as time flow in one single static graph. Presents dynamic pattern in one single graph is important for we are going to compare route using all points it contains in following steps.

5. **CALCULATION OF DISTANT FUNCTION**

For now we have finished the construction of dynamic route for every single visitor, but to cluster all different routes into groups we still need to find relations between them. Common clustering algorithm are used to dealing with scatter graph which uses their coordinate as input of clustering. But for dynamic route they have more than one point in each route and it’s necessary to take every point into consideration when we compare and cluster among them. We use a parameter called ‘distant function’ to measure the similarity between different routes, this method is inspired by the research of similarity search among trajectory database[2] [3].

The parameter we are going to use must take all point in one route into calculation. Distance function is measured by the interval of significant places with same location. We present our algorithm to calculate the distant function between any two different dynamic routes $P$ and $Q$ in figure 4. To measure distant function there have be at least one pair of same location significant point for two different routes, or the distant function is set to 0. If there are pair of same location significant point exist, for every significant point in the dynamic route with fewer total point, the relative interval is measured by the time distance between the closet point that have same location and total distance of dynamic route with longer total time distance, then the distance function is calculated with the amount of all relative interval and the number of point in the dynamic route with fewer points. We give several example of distant function calculation in figure 5.

**Algorithm: distant function**

**Input:** dynamic route $P (P_1, P_2, \ldots, P_m)$, $Q (Q_1, Q_2, \ldots, Q_n)$

**Output:** distant function $D$

1. $s = 0$
2. FOR $j = 1$ TO $m$
3. IF EXIST $x_{pj} = x_{qk}, y_{pj} = y_{qk}$ THEN
4. $k = |t_{pj} - t_{qk}|$
5. $s = s + (\text{MAX}(t_{pm}, t_{qn}) - k) / \text{MAX}(t_{pm}, t_{qn})$
6. END FOR
7. $D = s / \text{MIN}(n, m)$

**Figure 4. The algorithm to calculate distant function**
6. Hierarchical Clustering

Now we have the distance function of every two different dynamic routes, the next step is to cluster all different routes. The clustering algorithm that we are going to use is hierarchical clustering. Hierarchical clustering clusters data over a variety of scales, it doesn’t need certain information of location or coordinate, what we need to input is simply the distances of every two different objects, and we happened to have it in the format of distance function matrix. By inputting the matrix of every two different dynamic route, we can get another matrix Z called linkage of former matrix like figure 7 and a binary cluster tree like figure 8 (a).

\[
\begin{array}{ccc}
236 & 250 & 3.2600 \\
59 & 107 & 3.2900 \\
1 & 989 & 3.3100 \\
\end{array}
\]

Figure 7. Example of linkage matrix

In the output of linkage matrix, each row presents the level of two different clusters join each other. The first two columns show the groups that joined with each other. The third column presents the height of level that when two groups joins happen. In the example of figure 7, the linkage begins with group of 236 and 250, which have the largest distance function and join each other at a level of 3.26. The linkage function continues by grouping objects 59 and 107, which join each other at a level higher than 236 and 250, means they share less similarity with each other.

The third row indicates that the group 1 and group 989 join each other. But if we take an observation of binary cluster tree we get at figure 9 (a) or figure 10 (a), there is not a group of 989. In fact group 989 is the newly formed binary cluster created by the grouping of objects 236 and 250. When the two groups join each other and come into exist with a new cluster, it must assign the cluster a new group number, starting with the number m + 1, where m is the number of objects in the original data set.

Figure 5. Two examples of distant function calculation. Point with color other than blue and orange are pair of same location point. (a) shows a limited distant function and (b) shows little distant function.

The larger the distance function is, more similarity the two dynamic routes have. People hang around as same group have dynamic route totally overlap, and their distance function is largest as 1. Those ID number with 1 as their distance function are treated as a small group and we have 985 groups overall. Groups in (a) have 3 pairs of overlapped significant points, each group have totally same time stamp when visited so for these three pairs they got a distance function of 3, but there are total 10 points in blue graph and 11 points in orange graph, orange graph has a smaller amount of points yet still have 6 points that share no same location with orange one, so they have a limited distance function of 0.3. Groups in (b) have only 2 pairs of significant points with same location, what’s more they are not overlapped but having a small time interval, the interval makes the distance function of these 2 single point less than 1, while other points share no same location significant places make them have a even smaller distance function.

We get the distance function of two different dynamic routes, using same method we could get a matrix of every two different dynamic route. The distance function matrix can be seen as figure 6

\[
\begin{array}{ccc}
1 \\
D21 & 1 & ^{^}\\\n... & ... & 1 \\
Dml & ... & Dmj \end{array}
\]

Figure 6. Matrix of Distant Function
The binary cluster tree we got is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. Figure 8 (a) shows an example of binary cluster tree with only 5 groups. The binary cluster tree we got by inputting distant function matrix with 985 groups and running hierarchical clustering is shown in figure 8 (b).

The height in the binary tree represents the distance between different groups. In figure 8 (b) most of groups join each other at height around 3.5 to 3.7, which means majority of groups share a limited similarity with each other at the same level. The higher level two clusters join together, the larger differences they have. So the groups join with other on the right side in figure 8 (b) means they have little similarity with the majority.

To settle a level of which height is decided by ourselves, that means by setting different level we can group all the movement data into clusters of different size and similarity, and we can adjust the level if we want to observers clusters in certain size or similarity as we hope.

Figure 8 and figure 9 show a example of setting different level to get cluster of different size and similarity. In figure 8 we set the level at a little higher than 3.25 but lower than 3.3, and then what we get is a group of 2: visitor of group 236 and 250. We can give a observe of this two different dynamic routes. 236 and 250 have two dynamic routes with limited similarity, though never go together in the former part of their visit, they route form a specific pattern when they are going to leave the park. During the 7h to 7.5h time period of their visit, these two different groups of visitor visit 3 different significant points (47,11),(16,66) and (43,56) in certain order, which consist a certain movement pattern and clustering them in same group. In figure 9, we set the level a little higher at above 3.3 but lower than 3.35, what we get is a larger group of 5, including group 1, 59 and 107 visit these 3 places in their earlier part of their visit, at time stamp around 3.5h to 4h, while group 1 visit these 3 places in the middle of their visit, around time 4.5h to 5h. So although the similarity of geographic pattern make them join together at a higher level of 3.3, the differences in dynamic make them separate at level 3.25, which show the importance of dynamic in our project.
Overall the larger the level we set, the bigger the size of cluster we will get, but in each cluster we get fewer similarities. On the contrary the smaller the level we set, the higher accuracy we will get for each cluster, which means dynamic routes in same cluster are going to have certain movement pattern more easily to find out, and the accuracy comes with exchange of size.

The flexibility of hierarchical clustering gives us much freedom to settle the level as we wish. If we want to observe clusters with pattern that share high similarity and can be easily understood, it’s better for us to set a relatively low level; if we want to cluster all the groups into a smaller amount of cluster, we’d better set the level a little higher. If we want to find balance between easily understood pattern and relatively smaller amount of clusters, we need to get a further understand of the cluster tree we got. In the binary cluster tree we got in figure 8 (b), we can find two large gap around level 3.6. The adjacent groups display a significant differences of height joined with each other, and they join each other at height much higher than the level they join with another adjacent groups. These gaps present big differences in similarity and clear movement patterns. If we set level to group objects on each side of gap into different clusters we can get the balance between clear patterns and relatively smaller amount of clusters.

7. **Outlier**

Outlier means groups that share little similarity with majority groups. They join other clusters at a extremely high level like in figure 11 (a) and figure 11 (b).

Outliers join with other at a extremely high level so they are always treated as single cluster when we run the clustering algorithm. But the cluster of outlier are always meaningless for a normal height of level set them as cluster of only one, and a cluster of no others means you are not able to find certain pattern for there are only one pattern in the group. If you are trying cluster these outlier with others the only way is to set a extremely high level, and that means you are clusters all objects into one single group, that’s also not possible to find a pattern there for a group of all is nothing different with a group of none. The best way to deal with them is to set them aside when doing the cluster work by setting an appropriate level.

The cause of outlier varies. One possible cause is that the outliers are group of large amount of people. A group of 40 people moving together is rare for DinoFun amusement park, and so is the movement pattern of this super large group. A group of 40 is certainly not moving around like a group of 4, and the differences may cause unique pattern that not similar with any other smaller group of people. Another cause is much dangerous, those outlier are potential damage maker. This may often happens with small group like group of 2 or 1, for the majority of people are visiting amusement park in smaller group, and it’s really uncommon to have no similarity with majority of others. Park manager better pay more attention to those small outlier groups.
8. CONCLUSION AND FUTURE WORK

In this project our goal is to cluster visitors into different groups with similar movement pattern, also we would like to find those outliers share little similarity with majority of visitors. We construct the dynamic route of every single person by connected significant places of ‘check-in’ data in a 3D time-space cube, then based on the algorithm of searching similar pattern in projector data base we use a parameter called ‘distant function’ to measure the similarity between every two different dynamic route. Got the distant function matrix we used a clustering algorithm called hierarchical clustering to cluster all groups into a binary cluster tree in which groups join each other at different level of height. By setting the height at appropriate level we are going to cluster visitors into groups with sizes and similarity as we wish, and those small group join with other at extremely high level are outliers we should pay attention to.

One direction of future work is to reduce the heavily overlap of space-time cube when doing visual analysis. We use connected line graph represents movement of single person form one place to another so it’s unavailable to get heavily overlap graph when we trying to compare multiple route in a single graph, these may leads to difficulties when we trying to find the common pattern. This problem could also be settled down if we find a reliable algorithm to find the common pattern automatically, how to describe the pattern and use parameter to measure them are details needs further discussion.

Another direction of future work is using exist movement datasets to predict move of human groups. For those who share common pattern with majority the prediction is useful in area like traffic management, and for those outliers the prediction may prevent potential terror assault.

9. REFERENCES

