Temporal Patterns of Shopping Behaviors Extracted from PT data

Weiying WANG *, Toshihiro OSARAGI **

Abstract: Temporal characteristics of people's daily behaviors, such as working, shopping, socializing, visiting, and so on, are complicated. In this paper, we attempt to analyze the temporal distributions of shopping behavior as the start of studying the temporal characteristics of various activities. In this research, the joint probability distributions of the start and end time of shopping behaviors are analyzed and multiple shopping patterns are extracted. The relationships between individual shopping pattern and individual factors, including socio-demographics and daily activity sequences, are further investigated and discussed.

Keywords: shopping behavior, temporal distribution, start and end time, clustering.

1. Introduction

Multi-dimensional or multivariate activity-travel patterns have been studied from various data sources in terms of different attributes. Start and end time choices of activities are important components of traveler's decision regarding trip-making. Bhat [1] examines the joint nature of travel mode and departure time choice for urban shopping trips by employing a nested structure model. Lu et al. [2] study the complex relationships among socio-demographics, activity participation and travel behavior using the structural equation model. The results show that travel behaviors can be better explained by including activity participation endogenously in the model, than through socio-demographics alone. In a previous research, we extracted sequential patterns of daily human activity from Person Trip survey data (hereafter PT data). We further notice that in most times, one type of activity happens only in a limited number of activity sequences. This indicates that activities do not happen randomly in a person's daily life, and temporal and spatial patterns may exist for each activity. In this research, we try to extract temporal patterns of shopping behavior as our first attempt to analyze the temporal characteristics of various activities.

2. Dataset and Methodology

In this research, we propose a method to extract temporal shopping patterns from Person Trip survey data.

2.1 Person Trip survey data.

Person Trip survey has been carried out every ten years by the Ministry of Land, Infrastructure and Tourism of Japan, for detecting travel behaviors in cities. In this research, the Person Trip survey data was collected in October 2008 and records each trip of people in one day in Tokyo Metropolitan area including the origin, destination, travel purpose, travel mode and time, as well as personal attributes such as age, gender, occupation, etc. The day starts from 3:00 AM and ends at 3:00 AM in the next day. Individual activity sequence is inferred from travel purposes. Detailed explanations about PT data are given in [3].

2.2 Basic ideas.

We assume that there are temporal shopping patterns. The choice of shopping pattern of a person may change with his/ her attributes and daily activity sequence. Individual attributes are age and occupation, and daily activity sequence is a sequence of activities that a person does in a day, for example, *stay at home-work-shop-go*

*　学生会員　東京工業大学環境・社会理工学院（Tokyo Institute of Technology）
〒152-8550　東京都目黒区大岡山 2-12-1　E-mail: wwy1026587697@gmail.com
**　正会員　　東京工業大学環境・社会理工学院（Tokyo Institute of Technology）

back home. For a shopping behavior, the start and end time are the temporal attributes. Shopping patterns are extracted from the 2-D space of the joint distribution of start and end time of shopping behaviors (Fig. 1-a). Following the idea introduced by Osaragi et al. [4], we cut the 2-D space into 30 min-by-30 min cells, where the center point of each cell is the time in hours or half hours (Fig. 1-b). Shopping patterns can be obtained by grouping cells. In any shopping pattern, people may prefer to start and end shopping in certain cells than others. For any group of people, their preference for cells should be the same, which means 2-D distributions of different groups in each shopping pattern should be similar (Fig. 1-c). We employ this idea and believe that if two groups prefer different cells in the same pattern, these cells can be divided into multiple patterns for a good understanding of each pattern (Fig. 1-d).

2.3 Method of extracting shopping patterns.

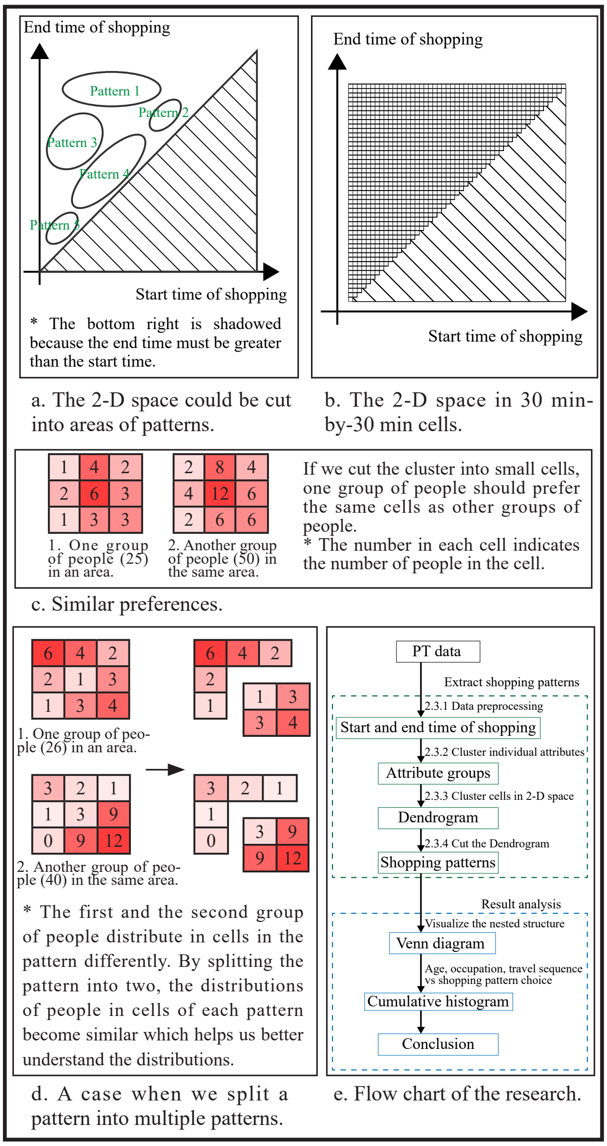
There are four steps to extract shopping patterns in the 2-D space (Fig. 1-e).

2.3.1 Data preprocessing.

In this paper, to simplify the calculation, we analyze the shopping behaviors of those who go shopping only once in the day. Also, only sequences with more than 500 people are used for analysis and people with rare sequences are excluded. Finally, we obtain a sample of 70,355 people from the dataset.

2.3.2 Cluster individual attributes.

As is mentioned in Section 2.2, the start and end time choice of different attribute groups might be similar. In the PT dataset, there are 17 age groups, 15 occupation groups and 20 activity sequences, and thus there would be 5,100 attribute combinations. Therefore, only a few people exist in each group which could be very hard to be analyzed. Given the above considerations, we cluster attributes first. Take the age groups for example, different age groups have their own 2-D distributions of points. The dissimilarity between two distributions is measured by the Jessen-Shannon divergence defined as follows:



a. The 2-D space could be cut into areas of patterns.
b. The 2-D space in 30 min-by-30 min cells.
c. Similar preferences.
d. A case when we split a pattern into multiple patterns.
e. Flow chart of the research.

Fig. 1. Basic ideas of the clustering.

$$JS(P\|Q)=\frac{1}{2}KL(P\|M)+\frac{1}{2}KL(Q\|M), \qquad (1)$$

where P, Q are arrays of probabilities that a person of age groups p, q occurs in a list of cells in the 2-D space (Fig. 2-a); $M=(P+Q)/2$ and $KL(P\|M)$ is the Kullback-Leibler divergence given by:

$$KL(P\|M)=\sum_{i=1}^{n}P(i)\log(\frac{P(i)}{M(i)}), \qquad (2)$$

where $P(i)$ is the probability that a person in age group p occurs in the i-th cell; $M(i)$ is the average probabilities that people from age group p and q are in the i-th cell, and n is the total number of cells in the 2-D space.

Using hierarchical clustering, age groups are clustered based on the dissimilarity between distributions of points of each age group. The same procedure is conducted for occupation groups and activity sequences.

2.3.3 Cluster cells in the 2-D space.

There are different groups of people in each cell. The dissimilarity between two cells is also measured using the Jessen-Shannon divergence defined in Equation (1), where P, Q are arrays of proportions of people of different groups in the cell p and q (Fig. 2-b). When cells are clustered using the hierarchical clustering method, a constraint is added that only adjacent cells can be clustered.

2.3.4 Cut the dendrogram.

The dendrogram obtained in Section 2.3.3 is not normal because when cells are clustered under the adjacent constraint, two similar clusters that are not adjacent to each other are not clustered until they merge with other cells that connect them. This results in the fact that, the dissimilarity between two child nodes may be larger than that between two father nodes (Fig. 3-a).

The dendrogram is also complicated with a nested structure (similar to that shown in Fig. 3-b). If the dendrogram is cut with a dissimilarity standard, the nested structure would be lost (Fig. 3-c). We attempt to cut the dendrogram into clusters that keep the nested structure (Fig. 3-b) with the following conditions:

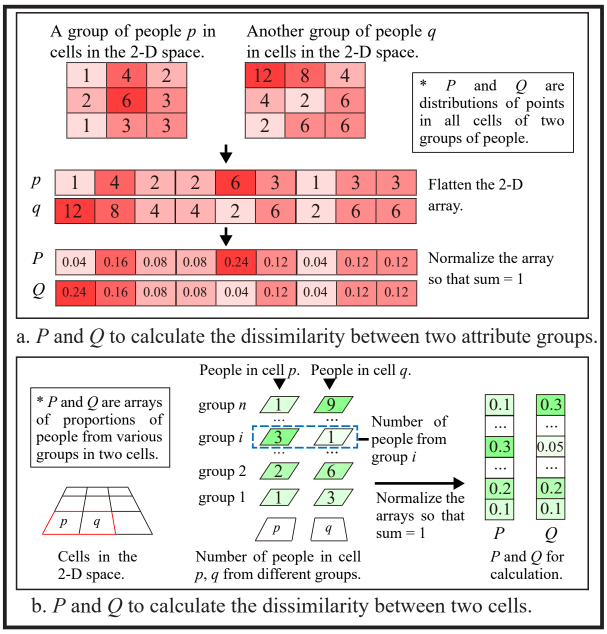
Condition 1: All clusters have at least $N_{\min}$ elements (cells).

Condition 2: The dissimilarity between the two child nodes of a cluster node should be less than $\lambda_{\max}$.

If the dendrogram is cut only from the above two principles, there might be many groups with similar elements as is shown in Fig. 3-d. We set two additional conditions to avoid that and detect the nested structure:

Condition 3: If one node meets condition (1) and (2) to be a cluster (node 1 in Fig. 3-d), and another node (node 2 in Fig. 3-d) nested below it can also be a cluster, while the other child node (node 4 in Fig. 3-d) cannot, we set that only the top node (node 1) is a cluster.

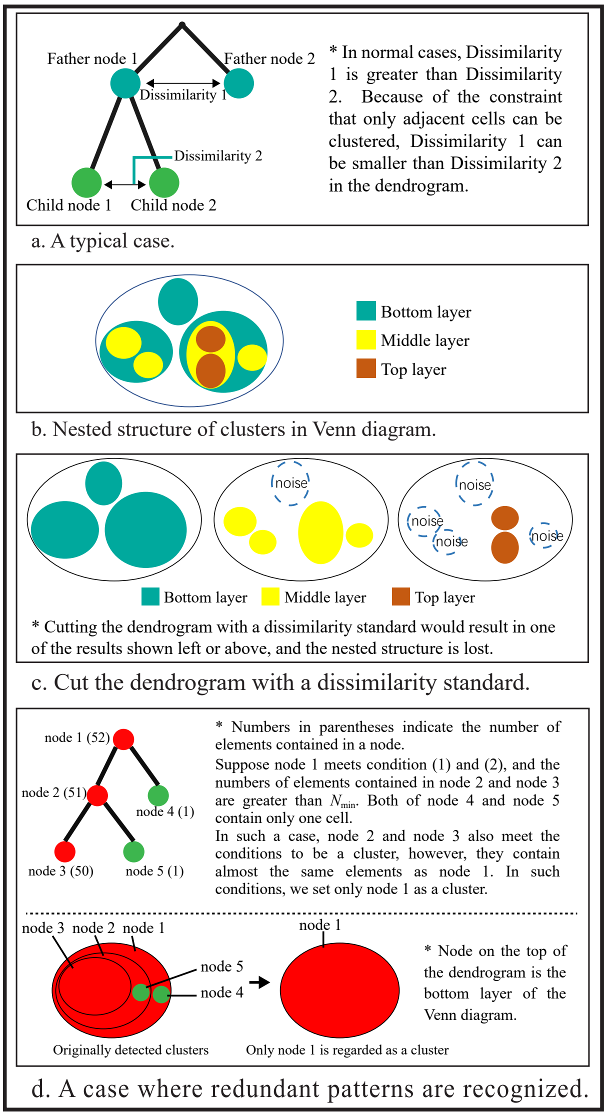
Condition 4: If a node meets condition (1) and (2) (it may not be a cluster because of condition (3)), and the number of cells under the node is greater than $N_{\max}$, and the dissimilarity between its two child nodes is bigger



a. P and Q to calculate the dissimilarity between two attribute groups.

b. P and Q to calculate the dissimilarity between two cells.

Fig. 2. P and Q in Equation (1)



a. A typical case.

b. Nested structure of clusters in Venn diagram.

c. Cut the dendrogram with a dissimilarity standard.

d. A case where redundant patterns are recognized.

Fig. 3. Cut the dendrogram.

than λ_{min}, each of its child nodes will be a cluster if both of them meet condition (1) and (2).

In total, there are 4 parameters to cut the dendrogram into clusters: λ_{max}, λ_{min}, N_{max}, and N_{min}. After getting the clusters, we merge some middle-layer clusters so that all clusters can only be nested once at most (Fig. 4). This step is performed because most middle-layer clusters have few elements by themselves and cause redundancy in the number of clusters. We should note that the increase in the number of nesting levels impedes a good understanding of the results.

3. Results

3.1 Clustered results of individual attributes.

Using the hierarchical clustering method, we cluster attributes and cut the dendrogram based on our experience. 5 age groups, 6 occupation groups and 6 sequence groups are obtained (Table 1). Theoretically, there are 180 groups of people. However, some group combinations do not contain anybody, and 82 valid group combinations are left.

3.2 Shopping patterns and the nested structure.

Different parameters used in Section 2.3.4 give different results. After multiple tests, we choose the following parameters: $\lambda_{max} = 0.5$, $\lambda_{min} = 0.1$, $N_{max} = 30$, and $N_{min} = 3$. We obtain eleven patterns with the above parameters, which is a proper size for further analysis.

Fig. 5 shows the clustered results in the 2-D space. Each color indicates a shopping pattern, while black are noise cells that do not belong to any pattern. The nested structure of shopping patterns is shown in a Venn diagram in Fig. 5, and Table 2 shows details of the start and end time of shopping patterns.

3.3 Individual attributes vs shopping pattern choice.

The relationships between individual attributes, daily activity sequences and the choice of temporal patterns of shopping behaviors are further analyzed. Fig. 6 shows the shopping patterns of each attribute group.

From Fig. 6-a, people aging 20-24 are more likely to shop at night and less likely to shop in the afternoon than

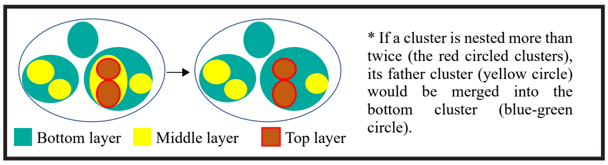


Fig. 4. Merge middle-layer clusters.

Table. 1. Clustered results of individual attributes and activity sequences.

ID	Age range
A1	5-14
A2	15-19
A3	20-24
A4	25-59
A5	60-

a. Age groups.

ID	Occupations
O1	Skilled workers, retailers, service-related, transport-related, office workers, specialized workers, managers, other workers
O2	University students
O3	Kids before middle school
O4	Security guards
O5	High school students
O6	Farmers, household wives/ husbands, and the unemployed

b. Occupation groups.

ID	Activity sequences
S1	Home-Shop-Work-Home
S2	Home-Work-Shop-Work-Home
S3	Home-Shop-Home-Entertainment-Home, Home-Shop-Home-Personal activity-Home, Home-Pick up-Home-Shop-Home-Pick up-Home
S4	Home-Shop-Home, Home-Hospital-Shop-Home, Home-Entertainment-Shop-Home, Home-Personal activity-Shop-Home, Home-Shop-Personal activity-Home.
S5	Home-Other works-Shop-Home, Home-Leisure-Shop-Home, Home-Hospital-Home-Shop-Home, Home-Personal activity-Home-Shop-home, Home-Entertainment-Home-Shop-Home
S6	Home-Work-Shop-Home, Home-Work-Home-Shop-Home, Home-Educational-Shop-Home, Home-Educational-Home-Shop-Home

c. Activity sequence groups.

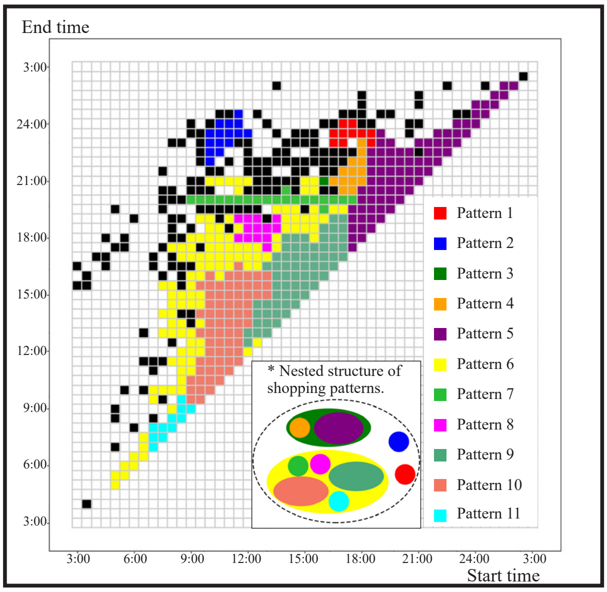


Fig. 5. The 2-D distribution of clustered shopping patterns.

the younger and older generations. As age increases, the preference to shop in the morning is strengthened. In Fig. 6-b, kids (O3) are more likely to shop in the afternoon, while office workers (O1), university students (O2) and high school students (O5) are more likely to shop at night. A large proportion of household wives/ husbands, the unemployed and farmers (O6) shop in the morning. In Fig. 6-c, people who shop before work usually have shopping behaviors in the morning (S1), and who shop after work or school usually do it at night (S6). Most of those who go back home after personal or leisure activities and go shopping again from home have shopping behaviors in the afternoon (S5), while those who do not go back home have more chance to go shopping in the morning (S4). People with different individual attributes and activity sequences have distinct preferences in the time choice of shopping behavior, and daily activity sequence plays a more important role than other attributes.

4. Conclusions

In this paper, we propose a method to extract temporal patterns of shopping behaviors from PT data. Individual attributes and daily activity sequences are clustered first, which decreases the number of groups of people. Cells on the 2-D temporal plane are grouped with the hierarchical clustering method, where the dissimilarity of two cells is measured by the Jessen-Shannon divergence between the proportions of groups of people in each cell. Method to cut the dendrogram from hierarchical clustering is stated, and four parameters are adopted to control the size and dissimilarity of elements in each cluster. We further merge middle-layer clusters and compare the shopping pattern choice of people by attributes. The distinguishing feature of the proposed method is that clustered results can maintain the nested structure. Another appealing characteristic is that we can cluster the points (people) in the 2-D space and find temporal patterns that are easy to interpret and understand with properly selected parameters.

Table. 2. Details of each shopping pattern.

Pattern	Description
Pattern 1	Start in the afternoon and long-time shopping.
Pattern 2	Start in the morning and long-time shopping.
Pattern 3	Start in the afternoon/ night and short-/medium-time shopping.
Pattern 4	Start in the afternoon and medium-time shopping.
Pattern 5	Start in the afternoon or at night and short-/ medium-time shopping.
Pattern 6	Start in the morning/ afternoon, and short-, medium- or long-time shopping.
Pattern 7	Start from working time and end shopping until shops close.
Pattern 8	Start at noon and shop in the whole afternoon.
Pattern 9	Start in the afternoon. Short-/ medium-time shopping.
Pattern 10	Start in the morning. Short-/ medium-time shopping.
Pattern 11	Start in the early morning and short-time shopping.

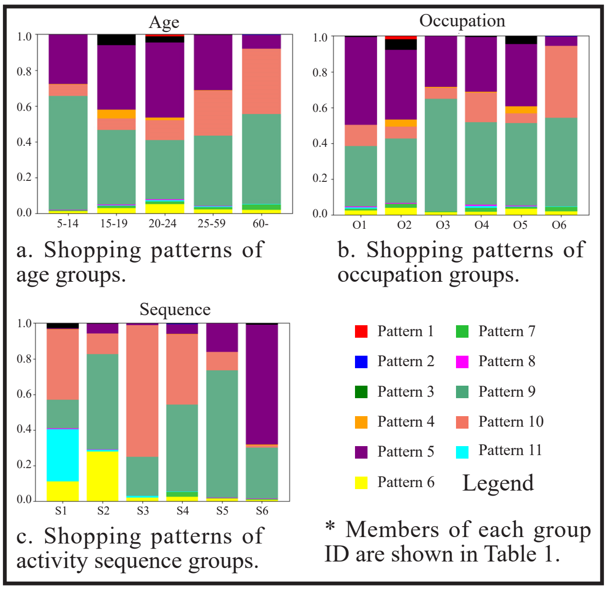


Fig. 6. Shopping patterns in each attribute group.

5. Acknowledgements

This paper is part of the research outcomes funded by KAKENHI (Grant Number 17H00843). The authors would like to thank Dr. Maki Tagashira for valuable comments on the paper.

References

[1] Bhat, Chandra R. “Analysis of Travel Mode and Departure Time Choice for Urban Shopping Trips.” Transportation Research Part B: Methodological 32, No. 6 (August 1998): 361–71.

[2] Lu, Xuedong, and Eric I. Pas. “Socio-Demographics, Activity Participation and Travel Behavior.” Transportation Research Part A: Policy and Practice 33, No. 1 (January 1999): 1–18.

[3] Osaragi, Toshihiro, and Ryo Kudo. “Enhancing the Use of Population Statistics Derived from Mobile Phone Users by Considering Building-Use Dependent Purpose of Stay | SpringerLink.” Accessed April 1, 2020.

[4] Osaragi, Toshihiro. “Classification and Space Cluster for Visualizing GeoInformation:” International Journal of Data Warehousing and Mining 15, No. 1 (January 2019): 19–38.