Activity Behavior Pattern Mining and Recognition

Ling Song[1], Hongxin Liu[1], Shunming Lyu[2], Yi Liu[1(✉)], Xiaofei Niu[1], Xinfeng Liu[1], and Mosu Xu[3]

[1] Shandong Jianzhu University, Jinan 250001, China
[2] State Grid Information and Telecommunication Branch, Beijing 100031, China
[3] The University of Melbourne, Melbourne, VIC 3010, Australia

Abstract. As human activity in mobile environments is facing with an ever-increasing range of data, therefore, a deeper understanding of the human activity behavior pattern and recognition is of important research significance. However, human activity behavior that consists of a series of complex spatiotemporal processes is hard to model. In this paper, we develop a platform to do pattern mining and recognition, the main work is as follows: (1) For comparing activity behavior, similarity matrix is computed based on activity intersection, temporal connections, spatial intersection, participant intersection and activity sequence comparison. (2) For calculating activity sequence similarity, an algorithm with $O(p(m - p))$ is proposed by line segment tree, greedy algorithm and dynamic programming. (3) Activity behavior pattern and socio-demographic pattern are derived by clustering analysis and mining. (4) Pattern is recognized under the inter-dependency relationship between activity behavior pattern and socio-demographic pattern.

Keywords: Activity behavior and socio-demographic pattern mining · Activity behavior and socio-demographic pattern recognition · Clustering analysis and mining · Activity sequence similarity · Activity behavior similarity

1 Introduction

As increasingly sensors and mobile devices are becoming smarter and more powerful, a remarkable consequence of these developments has been accumulated to more and more abundant human activity dataset, providing detailed information on the way people spend time. Activity behavior pattern mining derives a representative set of activity behavior patterns, and activity behavior recognition confirms what kind of activity patterns individuals belonged to, which could provide a priori knowledge in a wide variety intelligent system such as E-recommender system, social network analysis and mining, and urban transportation.

A large amount of related work has been done by previous researchers. We reviewed the following topics from various references: the influence of socio-demographic features, trajectory, spatiotemporal features, longest common subsequence (LCS) problem, and learning model.

J. Xiong et al. (Eds.): MobiMedia 2021, LNICST 394, pp. 663–678, 2021.
https://doi.org/10.1007/978-3-030-89814-4_48

Focusing on the influence of socio-demographic features, many studies concluded that human socio-demographic features could influence activity behavior significantly. In Lu and Pas's research, there existed relationships between socio-demographics and activity behavior [1]. According to socio-demographic features, after analyzing the multiple participations of different leisure activities, Kemperman and Timmermans determined the relevance of leisure activity that individual take part in and features of the living environment [2]. Focused on the activity patterns of individuals in dual-earner family, Bernardo et al. analyzed the function relationship of household socio-demographics and time use of activity [3].

With the aim of trajectory, on the assumption that individuals' real-world activity behavior is represented by their trajectory, Ying et al. proposed that the similarity between individuals is computed in terms of their maximal trajectory and potential friends are recommended based on social networks [4]. Zhang et al. defined a community as a group of individuals with similar activity and activity-travel trajectories [5]. Cai et al. focused on mining semantic trajectory pattern [6]. Outliers often reduce the performance of pattern mining and recognition, in order to reduce the outlier effect, Xiong et al. proposed a privacy and availability data clustering scheme (PADC), which enhances the selection of the initial center points and the distance calculation through detecting outliers during the clustering process [7].

Spatial and temporal information play an important role in feature extraction and representation for representing activity behavior. Banovic et al. converted the activity behavior logs into event sequences that show individuals' activity context, which are used to represent individuals' routine activity behavior [8]. For a deeper grasping of individuals' activity dynamics, Zhang et al. extracted individuals' spatial and temporal features from activities and/or trip events [5]. Activity feature extraction from video also capture both spatial and temporal information in computer vision for activity recognition [9]. You et al. presented a clustering algorithm with a trajectory similarity computation method by considering both semantic and geographic meaning [10]. In the work of Chakri et al., knowledge discovery is used to extracted space and temporal information from semantic trajectories [11]. With a kind of deep learning, convolution neural network is presented to catch spatiotemporal information in Ke et al. 's work [12].

The LCS has been widely used in bioinformatics and text comparison. Zhang et al. used the LCS of activity sequence to represent the similarity of two individuals [5]. The general conventional algorithm is dynamic programming with time complexity O(mn). Many researchers have done a lot of work on optimized LCS on typical inputs and some kind of infrequent inputs. Based on the work of H. S. Stone, the algorithm run in O($n\log n$) on many inputs [13, 14]. Hirschberg designed two algorithms, in the general case, the running time runs in $O(pn + n \log n)$, and p is the length of the LCS. However, while p is near to m ($m < n$), the running time requires bounded by $O(p(m + 1 - p)\log n)$ [15]. When the expected length of an LCS is close to m, Nakatsu developed an algorithm in O($n(m - p)$) [16]. Liu designed a parallel algorithm for LCS to make it more efficient [17].

Lots of studies have been done on how to build the learning model for activity behavior. Zhang and Jiang classified activities into the outdoor and the indoor activities, constructing a model of attendance prediction based on the Gradient Boosting Tree [18].

Hafezi et al. presented a pattern recognition model to identify time-use daily activity patterns by Fuzzy C-Means and regression tree based on household travel of diary time use survey [19]. Zhang et al. developed activity-travel model based on similarities of people's activity trajectories and spatiotemporal connection by social network analysis and community detection algorithm [5]. Banovic et al. forecasted human routine behavior by a Markov Decision Processes framework [8]. After given individual' time and location, Benetka et al. proposed a recommender system to return a list of ranked activities according to the probability of being done [20]. Activities and durations are predicted by computing their probabilities by two kinds of Long-Short Term Memory (LSTM) [21]. LSTM is also used to predict future activities and places [22]. In more applications, occupancy patterns are profiled by learning model based on time use survey [13, 23, 24]. Data sharing among connected and autonomous vehicles without any protection will cause private activity behavior leakage, Xiong et.al. constructed secure functions and implemented a privacy-preserving convolutional neural network (P-CNN) to ensure data security [25].

Although a large amount of work has been done in the previous studies, notably, the past studies seldomly focus on both activity behavior pattern and socio-demographic pattern mining. Therefore, we should make sense of the following problems:

(1) Similarity computation: on the one hand, the method needs to be related with comprehensively activities, location, and participants for computing similarity between activity behavior; on the other hand, an effective algorithm needs to be implemented in order to reduce computational complexity for computing similarity between activity sequences.
(2) Activity granularity: too fine and too coarse activity granularity makes it difficult to discover pattern characteristics. As different applications face different requirements, pattern mining and recognition system needs to adjust activity granularity flexibly.
(3) Visualization of clusters: for the purpose of obtaining the patterns, comprehensive analysis should be done based on the clustering results by visualization of activity sequence diagrams, scale charts and Probability Density Function (PDF) distribution graph of both activity behavior features and socio-demographic features.
(4) Pattern representation and recognition: Not only the representative set of activity behavior patterns but also socio-demographic patterns should be derived, bridging the relationship between them, which is further pattern recognized by activity or socio-demographic features.

2 Overview

The description of definitions and problem statements are firstly shown with the research outline following.

2.1 Definitions and Problem Statements

We give the following description about activity behavior.

Definition 1: Let $D = \{d_1, d_2,\ldots, d_{|D|}\}$ denote the set of individuals, where $|D|$ is the number of the individuals.

Definition 2: Let $A = \{a_1, a_2,\ldots, a_{|A|}\}$ represent the set of activity categories, where $|A|$ is the number of the activity categories, such as work and sport.

Definition 3: $\forall\ a_{\mathrm{i}}$ ($a_{\mathrm{i}} \in A$) has a starting time s_i and an ending time e_i, where s_i, $e_i \in \{0, 1,\ldots, t\}$, and t can be expressed in minutes from a certain time point.

Definition 4: Let $L = \{l_1, l_2,\ldots, l_{|L|}\}$ denote the set of location categories, where $|L|$ is the number of the activity categories, such as home and workplace.

Definition 5: Let $W = \{w_1, w_2,\ldots, w_{|W|}\}$ denote the set of participant categories, where $|W|$ is the number of the participant categories, such as parent and co-workers.

Definition 6: An activity action is defined as a quintuple $b(a, [s, e], l, w)$, where $a \in A$, s, e is the time that a begins and finish, $l \in L$ is the location of a takes place, and $w \subseteq W$ are the participants in the process of a.

Definition 7: For an individual d_i ($d_i \in D$), his/her activity behavior is described as a set of activity action, $B_i = \{b_{i1}(), b_{i2}(),\ldots, b_{|iB|}()\}$, where $b_{ij}()$ is an activity action, and $|iB|$ is the number of activities d_i involves in.

Definition 8: For the set D, the set of its corresponding activity behavior is described as $O = \{B_1, B_2,\ldots, B_{|D|}\}$.

Definition 9: For the set D, the set of its corresponding socio-demographic feature is described as $M = \{m_1, m_2,\ldots, m_{|D|}\}$, such as age, gender, education, occupation, and family income etc.

Based on the above definitions, the main problems of this paper are described as the followings.

Problem 1: Pattern mining of activity behavior is characterizing the cluster memberships of the set O, deriving clusters of homogeneous activity behavior pattern, $AP = \{P_1, P_2,\ldots, P_k\}$.

Problem 2: Pattern mining of socio-demographic feature is characterizing the cluster memberships of the set O, deriving clusters of homogeneous socio-demographic pattern, $SP = \{P'_1, P'_2, \ldots, P'_k\}$.

Problem 3: Given a set of socio-demographic or activity behavior features, pattern recognition problem can be stated to identify his activity behavior pattern and socio-demographic pattern he belongs to.

2.2 Research Outline

The framework of our model involves two modules: (1) Pattern mining module: aggregated similarity matrix is computed by comparing activity behavior and activity sequence between the individuals. Clusters are obtained by clustering algorithm based on the aggregated similarity matrix, and resulted in unique clusters of homogeneous daily activity behavior patterns by mining. Meanwhile, their corresponding socio-demographic patterns of homogeneous activity patterns are derived, implying inter-dependency between two kinds of patters. (2) Pattern recognition module: in view of the inter-dependency relationship between activity behavior patterns and socio-demographic patterns, pattern is recognized through socio-demographic or activity behavior features by the classifier.

3 Methodology

The process of pattern mining and recognition comprises four steps: (1) similarity adjacency matrix among individuals is calculated; (2) clusters are organized by clustering algorithm based on the similarity matrix; (3) representative patterns are obtained by characterizing the cluster memberships; (4) pattern are identified by classification algorithm.

3.1 Similarity Computation

Similarity comparison between two individuals is considered from the view of activity behavior and the view of activity sequence.

(1) Activity behavior similarity matrix

Activity behavior consists of a series of activity actions, so we calculate activity action similarity firstly. For two activity actions, k and l, they are represented as $b_k(a_k, [s,e]_k, l_k, w_k)$ and $b_l(a_l, [s,e]_l, l_l, w_l)$ respectively, the similarity between them, denoted as $sim_{act}[b_k,b_l]$, is defined as:

$$sim_{act}(b_k, b_l) = sim_a(a_k, a_l) \times sim_t([s,e]_k, [s,e]_l) \times sim_l(l_k, l_l) \times \frac{1}{P} \sum\nolimits_{p=1}^{P} sim_w\left(w_{k,p}, w_{l,p}\right) \quad (1)$$

$sim_a(a_k,a_l)$ is the activity similarity function to compare a_k and a_l, $sim_l(l_k, l_l)$ is the location similarity function to compare l_k and l_l, $sim_w(w_{k,p}, w_{l,p})$ is the participants similarity function to compare $w_{k,p}$ and $w_{l,p}$, $p = 1, 2, \ldots P$, as there may be more than one participants in involved in the action. The above similarity function value is defined as 1 if the variables are the same, otherwise 0. The time similarity function $sim_t([s,e]_k, [s,e]_l)$ is defined as follows:

$$sim_t([s,e]_k, [s,e]_l) = \begin{cases} \min(e_k, e_l) - \max(s_k, s_l), & \min(e_k, e_l) > \max(s_k, s_l) \\ 0, & \min(e_k, e_l) \le \max(s_k, s_l) \end{cases} \quad (2)$$

For two individuals, $d_i, d_j \in D$, their daily activity behaviors are $B_i = \{b_{i1}(), b_{i2}(),\ldots, b_{im}()\}$ and $B_j = \{b_{j1}(), b_{j2}(),\ldots, b_{jn}()\}$, and m, n is the number of activities for d_i and d_j. The daily activity behavior similarity $Sim_{act}[B_i, B_j]$ is computed as the followings:

$$Sim_{act}(B_i, B_j) = \sum_{k=1}^{m} \sum_{l=1}^{n} sim_{act}(b_{ik}(), b_{jl}()) \tag{3}$$

In summary, activity behavior similarity matrix is defined as followings:

$$R_{act} = [Sim_{act}(B_i, B_j)]_{|D|\times|D|} \tag{4}$$

(2) Activity sequence similarity matrix

For computing activity sequence similarity, the time with adjacent time and similar activities is discretized into a time segment, and each segment is regarded as a weighted point. Thus, an individual's daily activities are described by two-dimensional representations of time and activities, and the comparison of two individual's activity sequences becomes three-dimensional data comparison. For reducing the time complexity, in view of the large degree of discretization of the time segment, we use line segment tree and greedy algorithm to reduce the dimension, and an algorithm with the complexity of $O(p(m - p))$ is proposed on the basis of previous studies of LCS problem [13–17].

For two individuals, d_i, $d_j \in D$, their activity sequences are described by S_i and S_j, notably, S_i and S_j actually refers to sequences of time segment after reducing the dimension. $LW_i[k]$ record the smallest j that S_i and S_j include a common subsequence of length k. We need to store list L to calculate LCS, so the whole space complexity is $O(n)$. Assuming that L has p values, then its space complexity is $O(p)$. When we calculate $LW_i[k + 1]$ based on $LW_i[k]$ by dynamic programming, the list L is traversed only once, the process is $O(p)$. As the process needs to be done m times, where $m_1 \in \{0\}$, $m_2 \in \{0,1\}$, $m_3 \in \{0, 1, 2\}$,.... therefore, the worst time complexity is $1 + 2 + \ldots + p + (m - p)p = O((m - p)p)$ under the case of $m_p = p$. Thus, we compute activity sequence similarity $sim_{seq}[S_i, S_j]$. Activity sequence similarity matrix is defined as followings:

$$R_{seq} = [Sim_{seq}(S_i, S_j)]_{|D|\times|D|} \tag{5}$$

R_{act}, R_{seq} are normalized by Min-Max Normalization. We aggregate the above two adjacency matrixes to a composite matrix M, as shown in Eq. (6).

$$M = \omega_1 R_{act} + \omega_2 R_{seq}, \omega_1 + \omega_2 = 1 \tag{6}$$

3.2 Pattern Mining and Identification

In this paper, spectral clustering method is used to partition a set of N individuals into k clusters, Silhouette Coefficient (SC) analysis, Calinski and Harabasz (CH) score, and Davies-Bouldin index (DB) index are used to evaluate the clustering results.Activity time series analysis, statistical analysis of activity category, location category, and activity participant category, probability density function analysis of activity category, location

category, and activity participant category are used to mine patterns, identifying the sets of representatives about activity behavior patterns. Statistical analysis of age, education, sex, metropolitan or not, race, occupation, family income, number of the family, number of children that is smaller than 18, working time, time with family, having an enterprise/farmer or not and labor force status etc. are used to mine patterns, identifying the sets of representative about socio-demographic patterns. Pattern is recognized by Random forest (RF).

4 Experiments and Analysis

The Multinational Time Use Study gathers more than a million diary days from over 70 national surveys [27]. The American Time Use Survey (ATUS) gives details of how, where, and with whom Americans take their time [28]. The experiment process consists of the followings: (1) Determining parameters of clustering; (2) Activity behavior pattern mining; (3) Socio-demographic pattern mining; (4) Pattern identification.

4.1 Data Preprocessing

We use ATUS of 2018 as our experiment dataset to explore individuals' daily life in 24 h (from 4:00am to 3:59am). Features about what, where, when and with whom are extracted. After preprocessing, we obtain a database of 13,133 individual samples. Notably, granularity of category in ATUS is too fine, so it's difficult to discover characteristics of pattern for some kind of application.

4.2 Determining Parameters of Clustering

In our experiments, granularity of categories can be merged flexibly. We merge some categories and select appropriate weights ω_1, ω_2 (Eq. (6)) and k (number of clusters) by SC, CH Score and DB Index. The parameters are set as followings:

(1) From the overall trend, with ω_1 increasing and ω_2 decreasing, the SC value has the trend of decreasing, CH Score has the trend of decreasing and DB Index has the trend of increasing, which means that ω_2 plays a more important role than ω_1. So, we set $\omega_1 = 0.1$, $\omega_2 = 0.9$, clustering result shows the best performance.
(2) It shows better clustering effect when k is smaller, but we hope to get more refined clusters that mine the diversity of patterns. Therefore, we need to look for a tradeoff. No matter what values ω_1, ω_2 are taken, $k = 7$ performs best under evaluating comprehensively with SC, CH Score and DB index, so we set $k = 7$.

4.3 Activity Behavior Pattern Mining

For discovery of activity behavior pattern, we operate more detailed data analysis and mining from the view of activity time series, activity category, location category and participant category based on the cluster results.

(1) Activity time series analysis

Figure 1 gives the activity time series of 7 clusters. One category corresponds to one color. X-axis is time (4:00am to 3:59am) and Y-axis is the number of individuals. We can observe that different clusters show obvious characteristics. Figure 2 shows probability density function (PDF) of number of individuals by time.

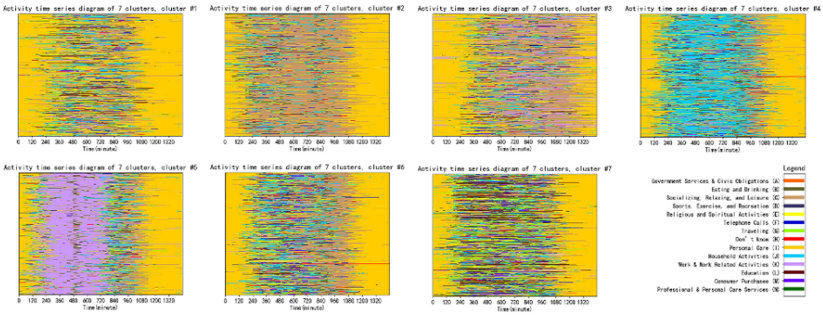

Fig. 1. Activity time series diagram with 14 categories (after activity merge).

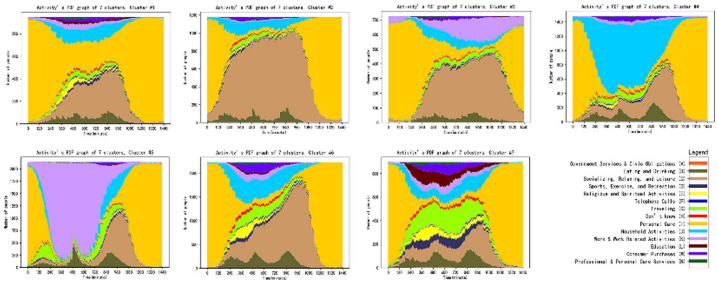

Fig. 2. Activity's PDF graph of clusters.

The typical activity characteristics of clusters as followings:

Cluster #1 has the largest proportion (59%) in I (Personal Care) with average 853 min. **Cluster #2** has the largest proportion (47%) in C (Socializing, Relaxing, and Leisure) with 682 average minutes. However, it takes up average 38min to work. **Cluster #3** has the third proportion (12%) in K (work & work related activities) with average 169 min. And the working hours are mainly distributed at 00:00 am, as Fig. 2 show. **Cluster #4** has the largest proportion (38%) is J (household activities) with average 473 min. **Cluster #5** has the largest proportion (37%) in K with average 534 min. And the working hours are mainly distributed at day. **Cluster #6** is the largest cluster with 23% of the individuals. It spends more time in B, D, E, G, L, M, N (including eating and drinking, sports, exercise, recreation, religious and spiritual activities, education, volunteer, travel, purchases of goods and services) than the other clusters, except #7. **Cluster #7**, compared with other clusters, the average time spend in B, D, E, G, L, M, N is the most. And it spends less time in C than #6.

(2) Activity location categories analysis
Cluster #1 has the largest proportion (58%) in J (including other Unspecified place, Blank, Other place, Don't know, Refused) with average 876 min. It spends average 463 min at A (Home or yard). **Cluster #2** has the largest proportion (52%) in A with average 766 min. **Cluster #3** has the third proportion (11%) in B(Workplace) with average 154 min, and the working places are mainly distributed at 00:00 am. **Cluster #4** has the second proportion (47%) in A with average 686 min. **Cluster #5** has the larger proportion (33%) in B with average 479 min. **Cluster #6** spends more time in D, E, F (including Restaurant or Bar, Place of worship, Store or Grocery) than the other clusters, except #7. **Cluster #7** takes least time at home, compared with other clusters, the average time in D, E, F, G, H, I (including Restaurant or Bar, Place of worship, Store or Grocery, School, Library, various mode of transportation, Gym, Outdoors) is the most.

(3) Activity participant categories analysis
Cluster #1 spends the least time with family (including Spouse, Unmarried partners, Own household children, Grandchildren, Parents, Brothers/sisters, Other related persons, Foster children, Own nonhousehold children < 18, Parents that not living in household, Other nonhousehold family members < 18, Other nonhousehold family members18 and older) and less time with other nonrelatives (including Housemates/roommates, Roomers/boarders, Other nonrelatives, Friends, Neighbors/acquaintances). **Cluster #2** takes the most time alone and least time with Co-workers (including Co-workers, people whom I supervise, customers, boss or manager). **Cluster #3** takes more time (average 127 min) with Co-workers. **Cluster #4** spends the most time (average 534 min) with family. **Cluster #5** spends the most time (average 212 min) with Co-workers. **Cluster #6** takes more time (average 505 min) with family and other nonrelatives. **Cluster #7** takes the least time alone, and takes the most time (average 573 min) with family and other nonrelatives.

4.4 Socio-demographic Pattern Mining

For discovery of socio-demographic pattern, we operate more detailed data analysis and mining from the view of age, sex, race, labor force status, occupation, family income, number of the family and underage children, time for work and family, having an enterprise/farmer or not, and Education level.

(1) Age and Sex
Cluster #3 is the youngest group, 59% of the people is between 20 and 59, and there are more male than female (53.7% vs. 46.3%). **Cluster #2** and **#6** are two old-aged groups. **Cluster #2** is the oldest group, 78% of the individuals are older than 50, and there are more male than female (55.3% vs. 44.7%). In **cluster #6**, the proportion from 40–79 years old takes up 66%, and there are more female than male (60.0% vs. 40.0%).

Cluster #1, **#4**, **#5** and **# 7** are the middle-aged groups. In **cluster #1**, the proportion from 30 to 69 years old takes up 62%, and there are more female than male (61.5% vs. 38.5%). **Cluster #4** is younger than **cluster #1**, the proportion from 30 to 69 years old takes up 76%, and there are more female than male (67.9% vs. 32.1%). In **Cluster #5,** the proportion from 30 to 69 years old takes up 85%, and there are more male than

female (53.2% vs. 46.8%). In cluster #7, the proportion from 30 to 69 years old takes up 65%, and there are more female than male (54.5% vs. 45.5%).

(2) Race and Labor force status

Compared with the other clusters, **cluster #1**, **#2**, **and #3** have higher proportion of the blacks. In **cluster #1**, the black account for the largest proportion of all clusters (the white and the black, 67% vs. 24%). The proportion that is not labor force is higher, which takes up 48%. In **cluster #2**, the proportion of the white and the black is 79% vs. 18%. The proportion that is employed at work is the lowest, which takes up only 28%, on the contrary, the proportion that is not labor force is the highest, which takes up 68%. In **cluster #3**, the proportion of the white and the black is 71% vs. 22%. The proportion that employed at work and not in labor force is 54% and 38%, respectively.

Cluster #4, #5, #6, and #7 have higher proportion of the whites. In **cluster #4**, the proportion of the white is the largest, which takes up 86%, and the proportion of the black is the smallest, which takes up only 8%. The proportion that employed at work and not in labor force is 48% and 43%, respectively. In **cluster #5**, the proportion of the white takes up 82%, and the proportion of the black takes up 10%. The proportion that employed at work is 99%. In **cluster #6**, the proportion of the white takes up 83%, and the proportion of the black takes up 12%. The proportion that employed at work and not in labor force is 47% and 46%, respectively. In **cluster #7**, the proportion of the white and the black is 78% vs. 13%. The proportion that employed at work and not in labor force is 58% and 35%, respectively.

(3) Occupation and family income

Family income distribution and division is based on Pew classification [29]. **Cluster #1, #2, and #3** belong to the low family income groups. People who earn less than $40000 take up around one half, people who earn $40000-$99999 take up 32%–39%, and people who earn more than $100000 take up only 15%. People in these clusters are minority in the occupation of Science/Technology/Management. For **cluster #1 and #2**, there is no obvious difference in occupation distribution, however, for **cluster #3**, people that work in service and related occupation take up 22%.

Cluster #4 and #6 belong to the middle family income groups. People who earn less than $40000 take up around 34%–35%, people who earn $40000–$99999 take up37%–39%, and people who earn more than $100000 take up around 26%–29%. People engaged in Science/Technology/Management is the majority, which takes up 23%–26%; people engaged in service and related occupation take up 12%–15%; people engaged in manual occupation takes up 10%–13%, and people with unknown occupation takes up 47%–50%.

Cluster #5 and #7 belong to the high family income groups. People who earn less than $40000 take up around 24%–26%, people who earn $40000–$99999 take up 42%–43%, and people who earn more than 100000 take up 34%. Almost all people in **#5** have work. People engaged in Science/Technology/Management is the majority, which take up 45%, while people engaged in service and related occupation take up 27%, and people engaged in manual occupation take up 23%. People in **#7** engaged in Science/Technology/Management is the majority, which takes up 29%, people engaged

in service and related occupation takes up 18%, people engaged in manual occupation takes up 11%, and people with unknown occupation takes up 38%.

(4) Number of the family and underage children

Cluster # 2 has the least family population. Single accounts for 42%, more than three people only account for 22.9%. At least one underage children account for only 18.5%.

Cluster #1, #3, and #6 are medium family. More than three people account for 35.8%–41.4%, and more than one underage child account for 29.4%–37.2%.

Cluster #4, #5, and #7 are large family. More than three people account for more than 50% and they have more underage children. **#4** has the largest number of family and underage children.

(5) Time for work and family

Cluster #2 and 3 has the longer working hours. **Cluster #2** has the longest working time,40.3% people work more than 700 min. Family time presents "Two ends big, middle small" state, that is, 50.8% people don't spend any time with family, and 24.9% people spend more than 600 min with family. In **# 3,** 41.3% work more than 500 min, and 80.5% take up less than 399 min with their family.

Cluster#1, #4, #5, #6, and #7 have the normal working hours ($\leq$500 min). In **#1,** 75.1% work less than 500 min, 10.2% people don't work, and 41.5% work less than 300 min. 67.3% people in **#4** work less than 500 min, 8.7% people don't work, and 39.1% work less than 300 min. They spend most of the time with the family and 51.7% spend more than 400 min with their family. 74.3% people in **#5** work less than 500 min, 1.3% people don't work, and 53.9% work less than 300 min. Although they are busy on work, they still spend more time with the family, in detail, 60.9% spend 1–399 min and 4.3% spend more than 400 min. 66.2% people in **#6** work less than 500 min, 8.3% people don't work, and 37.2% work less than 300 min. They spend more time with the family, with 34.7% spending more than 400 min with their family. 77.3% people in **#7** work less than 500 min, 11.7% people don't work, and 44.6% work less than 300 min. They spend more time with the family, about 36.8% spending more than 400 min with their family.

(6) Having an enterprise/farmer or not

Only less than 9% people in **cluster #1, #2, and #3** have an enterprise/farmer, however. More than 14% people in **cluster #4, #5, #6 and #7** have an enterprise/farmer.

(7) Education level

Cluster #1, #2, and #3 are three groups with the higher proportion of low education level and lower proportion of the high education level. People who are less than bachelor' degree take up about 75% and greater than or equal to bachelor' degree take up only 25%.

Cluster #4, #5, #6 and #7 are four groups with the higher proportion of the high education level and lower proportion of the low education level. People who are less than bachelor' degree take up only 51.0–57.3% and greater than or equal to bachelor' degree take up 41.4–49.0%.

4.5 Characteristics of Activity Behavior and Socio-demographic Pattern

From the above analysis, activity behavior pattern and socio-demographics pattern are derived, as Table 1 shown.

Table 1. Characteristics of activity behavior pattern and socio-demographic pattern

Pattern	Characteristics of activity behavior	Characteristics of socio-demographic
#1 10% (945)	The largest activity proportion in personal care, the least time to work, the largest location proportion is unspecified place, the least time with family and less time with other nonrelative	Middle-aged group, more female than male, with the higher proportion of the blacks, low family income, medium family population, the normal working hours, low education level
#2 (1179) 12%	The largest activity proportion in socializing, relaxing, and leisure, less time to work. the largest location proportion in home/yard, the most time alone and least time with co-workers	The oldest group, more male than female, with the higher proportion of the blacks, not labor force is the highest, low family income, the least family population, the employed people have the longest working hours, low education level
#3 (723) 8%	The working hours are mainly distributed after 00:00 am, the main location is the workplace, more time with co-workers	Youngest group, more male than female, with the higher proportion of the blacks, low family income, higher proportion in service and related occupation, medium family population, longer working time, less time with the family, low education level
#4 (1484) 15%	The largest proportion is household activities, caring for members, more time in home/yard, the most time with family	Middle-aged group, more female than male, with the higher proportion of the whites, middle family income, higher proportion in occupation of science, technology and management, large family population, the normal working hours, the most time with the family, with the higher proportion of having an enterprise/farmer, high education level
#5 (2136) 22%	The largest proportion is work related activities, the larger location proportion is workplace, the larger proportion with co-workers	Middle-aged group, more male than female, with the higher proportion of the whites, high family income, higher proportion in occupation of science, technology and management, large family population, the normal working hours, less time with the family, with the higher proportion of having an enterprise/farmer, high education level

(*continued*)

Table 1. (*continued*)

Pattern	Characteristics of activity behavior	Characteristics of socio-demographic
#6 (2229) 23%	More time in eating, sports, exercise, recreation, spiritual activities, education, travel, purchases of goods and various services, socializing, and leisure, more time in the place of restaurant, worship and store, more time with family and other nonrelatives	old-aged group, more female than male, with the higher proportion of the whites, middle family income, higher proportion in manual occupation, medium family population, the normal working hours, spend more time with the family, with the higher proportion of having an enterprise/farmer, high education level
#7 (897) 9%	The most time in eating, sports, exercise, recreation, spiritual activities, education, travel, purchases of goods and various services, socializing, and leisure, the least time at home, the most time in the place of restaurant, worship, store, library, school, varies mode of transportation and gym, least time alone and more time in family	Middle-aged group, more male than female, with the higher proportion of the whites, high family income, higher proportion in occupation of science, technology and management, large family population, the normal working hours, more time with the family, with the higher proportion of having an enterprise/farmer, high education level

From Table 1, both representative set of activity behavior pattern and socio-demographic pattern are derived, bridging the relationship between them, which is further pattern recognized by activity or socio-demographic features.

4.6 Pattern Identification

In this section, the pattern is recognized through socio-demographic or activity features by RF and parameters are set. For different application, we can select any kind of features as input for pattern recognition.

we use accuracy as evaluation criteria. Experiments are performed by 10 cross validation, and recognition accuracy is only 0.478. The reason lies in that some individuals in the clusters have no obvious characteristics, resulting the lower accuracy. For each cluster, we select 80% samples which squared Euclidean distance are closer to the centroid. With these samples, we do pattern recognition experiment as above, and recognition accuracy achieve 0.856.

Figure 3 and Fig. 4 are two test cases for pattern identification. The left of the figures is the input interface of pattern recognition. We input number of family, labor force status, number of underage children, relaxing time, family time, having an enterprise/farmer or not, metropolitan or not, race, age, sex, and family income. The right of the figures is the output interface of pattern recognition. The test case 1 is a typical working people and test case 2 is a typical housewife. We can see that largest probability that test case 1 is classified to the pattern #5 is 0.7270 and the largest probability that test case 2 is classified to the pattern #4 is 0.5994, which are consistent with analysis results in the Table 1.

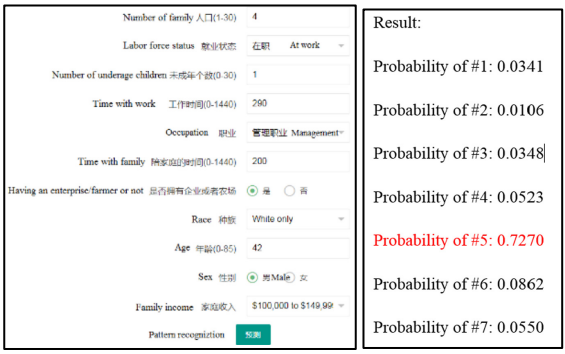

Fig. 3. The interface of pattern recognition (Test case 1).

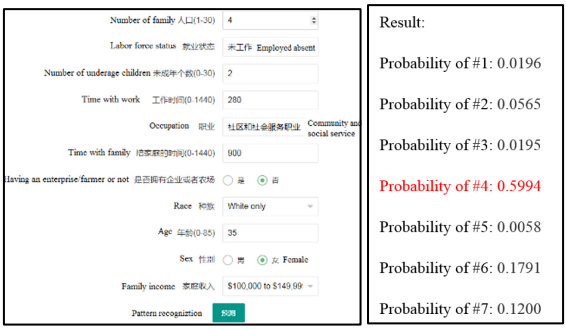

Fig. 4. The interface of pattern recognition (Test case 2).

5 Conclusion and Future Work

We address the human behavior pattern mining and recognition problems. This study contributes pattern mining and recognition by providing the linkage between activity behavior pattern and socio-demographic pattern based on the overall activity-based time use survey. In detail, we mine activity behavior patterns by deriving clusters of homogeneous daily activity behavior and activity sequence, where each pattern provides vital characteristics of activity behavior and socio-demographic characteristics. Furthermore, exploring more accurate activity behavior patterns play an important basic role in pattern recognition. Our proposed method can be applied to recommender system, activity schedule, social network analysis and mining, and urban planning.

Similarity computation between individuals is of primary importance in order to mine and recognize activity behavior pattern. The innovation is similarity computation between activity sequences. The time with adjacent time and similar activities is discretized into a time segment, as the degree of discretization is relatively large, the LCS algorithm with the complexity of $O(p(m-p))$ is proposed in this paper, which improves the efficiency of the algorithm. But it isn't suitable for the case when the degree of discretization is relatively low. In the future work, we will explore the semantic similarity computation for improving the understanding of semantic information. Furthermore,

activity sequence pattern will be mined by network analysis and community detection algorithm.

Acknowledgments. This work is supported by the Major Science and Technology Innovation Project of Shandong Province under grant No. 2019JZZY010435, and National Natural Science Foundation of China under grant No. 51975332.

References

1. Ying, J.C., Lu, H.C., Lee, W.C., et al.: Mining user similarity from semantic trajectories. In: ACM SIGSPATIAL International Workshop on Location Based Social Networks. ACM (2010)
2. Zhang, W., Thill, J.C.: Detecting and visualizing cohesive activity-travel patterns: a network analysis approach. Comput. Environ. Urban Syst. **66**, 117–129 (2017)
3. Guochen, C., Kyungmi, L., Ickjai, L.: Mining semantic trajectory patterns from geo-tagged data. J. Comput. Sci. Technol. **33**(4), 849–862 (2018)
4. Banovic, N., Buzali, T., Chevalier, F., et al.: Modeling and understanding human routine behavior. In: CHI Conference on Human Factors in Computing Systems. ACM (2016)
5. Trong, N.P., Nguyen, H., Kazunori, K., Le Hoai, B.: A comprehensive survey on human activity prediction. In: Gervasi, O., et al. (eds.) ICCSA 2017. LNCS, vol. 10404, pp. 411–425. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-62392-4_30
6. Lu, X., Pas, E.I.: Socio-demographics, activity participation and travel behavior. Transp. Res. Part A (Policy Pract.) **33**(1), 0–18(1999)
7. Xiong, J., et al.: Enhancing privacy and availability for data clustering in intelligent electrical service of IoT. IEEE Internet Things J. **6**(2), 1530–1540 (2019)
8. Kemperman, A., Timmermans, H.: Influence of socio-demographics and residential environment on leisure activity participation. Leis. Sci. **30**(4), 306–324 (2008)
9. Bernardo, C., Paleti, R., Hoklas, M., et al.: An empirical investigation into the time-use and activity patterns of dual-earner couples with and without young children. Transp. Res. Part A Policy Pract. **76**, 71–91 (2015)
10. You, W., Chenghu, Z., Tao, P.: Semantic-geographic trajectory pattern mining based on a new similarity measurement. ISPRS Int. J. Geo-Inform. **6**(7), 212 (2017)
11. Chakri, S., Raghay, S., Hadaj, S.E.: Semantic trajectory knowledge discovery: a promising way to extract meaningful patterns from spatiotemporal data. Int. J. Softw. Eng. Knowl. Eng. (2017)
12. Ke, Q., Bennamoun, M., An, S., Boussaid, F., Sohel, F.: Human interaction prediction using deep temporal features. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 403–414. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_28
13. Hunt, J.W., Szymanski, T.G.: A fast algorithm for computing longest common subsequences. Commun. ACM **20**(5), 350–353 (1977)
14. Aho, A.V., Hirschberg, D.S., Ullman, J.D.: Bounds on the complexity of the longest common subsequence problem. In: Symposium on Switching & Automata Theory. IEEE Computer Society (1974)
15. Hirschberg, D.S.: Algorithms for the longest common subsequence problem. J. ACM (JACM) (1977)
16. Nakatsu, N., Kambayashi, Y., Yajima, S.: A longest common subsequence algorithm suitable for similar text strings. Acta Informatica **18**(2), 171–179 (1982)

17. Liu, W., Chen, L.: Parallel longest common subsequence algorithm based on pruning technology. J. Comput. Appl. **26**(6), 1422–1424(2006)
18. Hafezi, M.H., Liu, L., Millward, H.: A time-use activity-pattern recognition model for activity-based travel demand modeling. Transportation **46**(4), 1369–1394 (2017). https://doi.org/10.1007/s11116-017-9840-9
19. Benetka, J.R., Krumm, J. Bennett, P.N.: Understanding context for tasks and activities, pp. 133–142 (2019). https://doi.org/10.1145/3295750.3298929
20. Krishna, K., Jain, D., Mehta, S.V., et al: An LSTM based system for prediction of human activities with durations. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 1, no. 4, pp.1–31 (2018)
21. Moon, G., Hamm, J.: A large-scale study in predictability of daily activities and places. In: The 8th EAI International Conference on Mobile Computing, Applications and Services. ICST, Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering (2016)
22. Flett, G., Kelly, N.: An occupant-differentiated, higher-order Markov chain method for prediction of domestic occupancy. Energy Build. **125**, 219–230 (2016)
23. Diao, L., Sun, Y., Chen, Z., et al.: Modeling energy consumption in residential buildings: a bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation. Energy Build. **147**, 47–66 (2017)
24. Barthelmes, V.M., Li, R., Andersen, R.K., et al.: Profiling occupant behaviour in danish dwellings using time use survey data. Energy Build., S0378778817342044-(2018)
25. Xiong, J., Bi, R., Zhao, M., Guo, J., Yang, Q.: Edge-assisted privacy-preserving raw data sharing framework for connected autonomous vehicles. IEEE Wirel. Commun. **27**(3), 24–30 (2020)
26. UCL Homepage. https://www.timeuse.org/mtus
27. BLS Homepage. https://www.bls.gov/tus/
28. U.S.Census Bureau: Highest Median Household Income on Record. http://www.census.gov/library/stories/2018/09/highest-median-household-income-on-record.html. Acessed 19 Stept (2019)
29. Zhang, J., Jiang, W., Zhang, J., Wu, J., Wang, G.: Exploring weather data to predict activity attendance in event-based social network: from the organizer's view. ACM Trans. Web **15**, 2 (2021). Article 10, 25 pages. https://doi.org/10.1145/34401341