

Ying Tang · Fengfan Sheng · Hongxin Zhang · Chaojie Shi · Xujia Qin · Jing Fan

Visual analysis of traffic data based on topic modeling (ChinaVis 2017)

Received: 31 August 2017 / Revised: 26 November 2017 / Accepted: 29 January 2018 / Published online: 16 March 2018
© The Visualization Society of Japan 2018

Abstract The spatio-temporal urban movement patterns can be extracted from the massive trajectory data recorded by GPS devices. Effectively analyzing the massive and complex traffic data and then finding useful information hidden in such data constitute challenging yet meaningful research. By providing the interactive visual analysis of the underlying traffic patterns of the city, the results can guide the users in choosing ideal locations for setting up shops for business operations. We construct the topic model to analyze the GPS taxi trajectory data. The topic information is combined with the traffic volume information to choose the representative candidate areas. Then, traffic flow graphs are generated between candidate areas to show the distribution of such areas and the taxi running rules. We study the distribution and semantics of the topics from three aspects: time, space, and POIs (points of interest). Thus, we can enhance the user's understanding of area characters by semantics. In addition, inspired by the wheels of vehicles, we design a metaphor-based glyph to summarize the multi-dimensional attributes of each candidate area. Users can explore the prospective areas' multiple attributes over time through varied interactions to learn the details of the area from multiple perspectives. Finally, we design and implement a visual analysis prototype system of traffic trajectory data as well as verify the feasibility and validity of the system in the case study.

Keywords Traffic data · Topic model · Semantic analysis · Data visualization

1 Introduction

At present, big cities attract numerous migrants because of their good economic foundation, strong business atmosphere, broad space for development, and other advantages. Many of them do business as their major source of livelihood. However, before setting up shop, they need to consider various factors, such as content, cost, object, location, and so on. Among them location is the key factor for the success of the business opening. A proper location can increase the visibility of the store and provide a steady passenger flow, whereas an inappropriate one will lead to a waste of time, energy, and investment. For people who are not familiar with the city, finding a good location with convenient transportation, large passenger flow, good market environment, and other beneficial factors can be quite challenging. Traditional approaches include population statistics, people flow calculations, and travel surveys. It is time-consuming and inflexible to

Y. Tang · F. Sheng · X. Qin (✉) · J. Fan
Zhejiang University of Technology, Hangzhou, China
E-mail: qxj@zjut.edu.cn

H. Zhang (✉) · C. Shi
State Key Laboratory of CAD & CG, Zhejiang University, Hangzhou, China
E-mail: zhx@cad.zju.edu.cn
Tel.: 86-571-88206681

conduct these approaches. Furthermore, appropriate data are also difficult to collect for these approaches. The problem was considered as a data mining task, and the authors used machine learning for social network data to estimate the most promising areas that attract a large number of customers (Karamshuk et al. 2013; Weng et al. 2018). The work (Liu et al. 2017) built a visualization-driven data mining model based on taxi data to generate solutions for the locations of billboards, and designed a variety of visual presentation methods. However, they only focused on the comparison between different solutions. Directly displaying multiple values of different categories at different locations is not supported.

By adopting visual analysis methods to study the taxi GPS data, this paper attempts to help businessmen select their potential store locations without the above-mentioned restrictions. Taxi GPS data not only records individual movement and behavior history, it also provides citywide coverage. Thus, it can effectively reveal the underlying traffic patterns and traffic flow in a city (Sun et al. 2013; Zheng et al. 2014; Chen et al. 2015; Chen et al. 2017).

To design the visual analysis system dealing with taxi GPS data to tackle the problem of selecting store sites, there are two requirements to be fulfilled. First, how to mine the huge amount of GPS data to limit and extract potential regions as candidate store sites. Second, how to enable users to compare and explore these candidate sites from multiple aspects (such as traffic volume, speed or POIs, etc.) by providing intuitive and interactive visualizations of these candidate sites.

In this paper, we introduce the Latent Dirichlet Allocation (LDA) model (Blei et al. 2003) to deal with the huge search space of GPS data, which transforms the traffic data to a document library and extracts topics. The extracted topics are represented as important regions holding both spatial and temporal features. To facilitate understanding of topics, we design *the LDA view* to display spatial and temporal distributions of topics integrated with POIs which contain semantic information. The topics are combined with certain traffic attribute to determine the candidate areas to reduce the search range. The traffic flow among these key areas are visualized as a graph for users to quickly grasp the general moving patterns and identify the interested regions in *the general traffic flow view*. To support the comparison and exploration of the interested regions, we designed *the attributes view* to display multiple spatio-temporal traffic attributes. The above three views are coupled by a variety of interactive operations to help users select their desired shop sites.

The main contributions of this paper are listed as follows:

- The trajectory data analysis method based on LDA topic modelling to extract important regions from huge amount of traffic data.
- The design of visualizations of three coupled views to help users understand topic semantics implied in trajectory data as well as explore multiple traffic attributes.
- An interactive framework of LDA-based data processing integrated with visualization design to help users select shop sites and the application of a real-world dataset.

The sections are organized as follows: Related works are covered in Sect. 2. We introduce the dataset and the system pipeline in Sect. 3. In Sect. 4, data pre-processing procedures are presented. The contents of Sect. 5 include the visualization design and visual analysis methods. Section 6 demonstrates the effectiveness of our solution with the case study. We draw conclusions in Sect. 7.

2 Related work

We mainly introduce the topic model for data mining and traffic visualization methods related to the work of this paper.

2.1 Topic model

Topic model (Blei 2012) breaks through the idea that “The higher the frequency of repetition of words between documents, the more likely it is that they are similar (Salton and Yang 1973; Salton et al. 1975a)”. The model explores the similarity between documents from the perspective of semantic relation and is commonly used in the field of machine learning, natural language processing, etc. The Vector Space Model (VSM) (Salton et al. 1975b) is a classic method for measuring document similarity. Latent Semantic Analysis model (LSA) (Deerwester et al. 1990) projects the vector of “word-document “to a lower” latent semantic” space so that it alleviates the problem of text semantics in VSM. Then, Hofmann et al. proposed

Probabilistic Latent Semantic Analysis model (pLSA) (Hofmann 1999), whose text analysis capability is better than that of LSA. Blei et al. came up with the LDA model (Blei et al. 2003) that solved the problem of linear growth of the training parameters and the “over-fitting” problem in pLSA, which is a probabilistic generative model and scalable.

2.2 Traffic data visualization

We mainly review two areas of related research: (1) aggregated visualization and (2) pattern mining and visualization of trajectory data.

2.2.1 Aggregated visualization of trajectory data

Some scholars first performed the aggregating calculation of a large number data, and then visualized the aggregated data for users to observe and explore. This method conveys the spatio-temporal information of trajectory data and reduces the visual clutter at the same time.

Zhao et al. (2008) proposed the activity ringmap to show the changes of intensities of people’s different kinds of activities over time. Andrienko (2008) presented how to use spatio-temporal aggregation for visual analysis of movement data in two views which are traffic-oriented and trajectory-oriented. Liu et al. (2011) showed the changes of amount and speed of taxis over time through similar visual design. Guo et al. (2011) used ThemeRiver to represent the changes of different types of traffic flow at the road crossing over time. Different from previous work, Landesberger et al. (2012) studied show individuals’ attributes changed over time. Based on the subdivision of the space, Ferreira et al. (2013) calculated the statistics of taxis passengers and used the shading of colors to encode the numbers of passengers taking and leaving a taxi. Sun et al. (2017) embedded spatio-temporal information into a selected route by broadening the route in a map.

As for origin–destination aggregation, locations along the trajectories can be aggregated to reduce the visual clutter and still capable to show the general moving patterns. Guo (2008, 2009) built the hierarchical partitions of the places and visualized the traffic flows between these large regions. The authors proposed to cluster spatial situations over time and the mean flows were computed to summarize the clusters (Andrienko et al. 2013; Andrienko. et al. 2012).

These techniques are capable of visualizing large amounts of movement data intuitively. However, these techniques cannot be simply applied in our work since they are not designed to support visual explorations and comparisons of different regions to find optimal site.

2.2.2 Pattern mining and visualization of trajectory data

It is of great significance to mine the hidden semantic information to assist users’ exploration in visual analytics of trajectory data. Andrienko et al. (2007) believed that it would help us understand human behaviors and the flow trajectory better by combining the movements of vehicles and human. In the study of bicycle tracks, Krüger et al. (2012) extracted keywords from Twitter data and the word cloud was generated and labelled at POIs based on keyword occurrence frequency. However, this method has a problem in scalability since the word cloud can cause visual clutter when the number of targets gets large. POIs are marked as representative icons on the map and used as the environmental context for the trajectory data (Krüger et al. 2012). The hierarchical structure is constructed for the user to interactively perform semantic reasoning of trajectory data. Andrienko et al. (2013) proposed to convert trajectory data from geographic space to abstract semantic space and the movements between different locations are transformed to movements between semantically different positions. Liao et al. (2015) linked the trajectories with the activities of suspects by connecting trajectory data with transaction data in the situation of kidnapping. These methods integrated other data of semantic meanings to trajectory data to enrich the data context to help users to analyse semantics meanings hidden in the trajectories. However, the support of user interaction is relatively simple and the analysis of knowledge in the data is limited to some extent. Chu et al. (2014) creatively transformed the geographic coordinates to street names reflecting contextual semantic information and introduced the LDA topic model for semantic analysis. Different from Chu et al. (2014), the words used in our work are grid indexes instead of the street names, which avoids the tedious work of assigning street names to GPS positions. Besides, we concatenate the grid index and the time to form the word which maintains spatio-temporal information. So the topics by our method are spatially-temporally varying which reflects the dynamics of the day. The work of SemanticTraj (Al-Dohuki et al. 2017) converted taxi

trajectories into taxi documents and employ a text search engine to mine the trajectories and propose visualizations with enhanced semantic information. Although taxi trajectories are both transformed to a corpus of documents by SemanticTraj and our work, the contents of the documents are different and the approaches to process the documents are different.

In this work, we aim to develop the visual analysis system of traffic data to find proper store locations for business people. Compared with previous work, our work has the following differences. First, we employ LDA model to extract important map grids, of which the semantic meanings are analysed according to the spatio-temporal features combined with the POIs. Second, we specifically design three views to present the highlighted traffic overview and the detailed traffic attributes of different regions, as well as the LDA view to explore the semantic meanings of different LDA topics. Third, the three-views is integrated into the system to enable users to explore them collaboratively to find their desired results.

3 Overview

3.1 Data description

We study the GPS records of taxis and POI data in Hangzhou, focusing on the main urban area which is within Longitude [119.8691,120.461] and Latitude [30.012, 30.4439]. The car-held GPS devices captured the records which contain fields including car ID, speed, geographic location (latitude and longitude), car status (occupied or empty), timestamp, etc. The number of GPS records per day is over 916 million. The GPS records feature both temporal and spatial characteristics, and our LDA topic modelling method preserves both characteristics during data processing. With GPS records, we can extract the trajectory for each taxi according to the car ID.

POIs are important semantic information to be used to understand the relations between people movements and activities (Zeng et al. 2017). In our work, we also use POIs as the supplementary semantic information to derive functions of different regions. The POI data are obtained through the Place API provided by Baidu map, which consist of names, categories, and geolocation of different organizations.

3.2 The system pipeline

Our system contains two main stages, i.e., data processing and visual exploration, as shown in Fig. 1. In the data processing stage, we first perform data cleaning process to remove invalid or dirty GPS records. After the data cleansing, the LDA topic modelling is employed to find the semantically-important areas. Besides, the area attribute statistics are computed to describe the area features. Finally, the key areas are selected by taking both the semantic results and the attribute statistics into consideration. The details of the data processing process are described in Sect. 4.

After data processing, an interactive visualization system is developed to support users to explore the traffic data to compare and select candidate locations. The system is designed to achieve the following objectives:

- Display the distribution of the key areas and the traffic flows among them, thus enabling users to have the picture of global summary traffic pattern among key areas.
- Show multiple statistics and their changes over time of key areas, facilitating users to compare and evaluate these key areas from different perspectives.
- Provide the visualization of LDA topic modelling results to make users understand the semantic meanings of important trajectory data embodied in different topics with integrated POI information.

Corresponding to the above goals, we designed three coupled views, which are the general traffic flow view, the attributes view, and the LDA topic analysis view respectively. The visual design and interactions of the three coupled views are described in Sect. 5.

Our visual analysis process follows the principle of “Overview first, zoom and filter, then details-on-demand” (Shneiderman 1996). The user obtains the overview of the key areas and traffic flows in the general traffic flow view and then analyse the semantic features of each area to choose the areas of interest through the LDA view. The attributes view supports the exploration and comparison of the multiple details of different attributes in the interested areas.

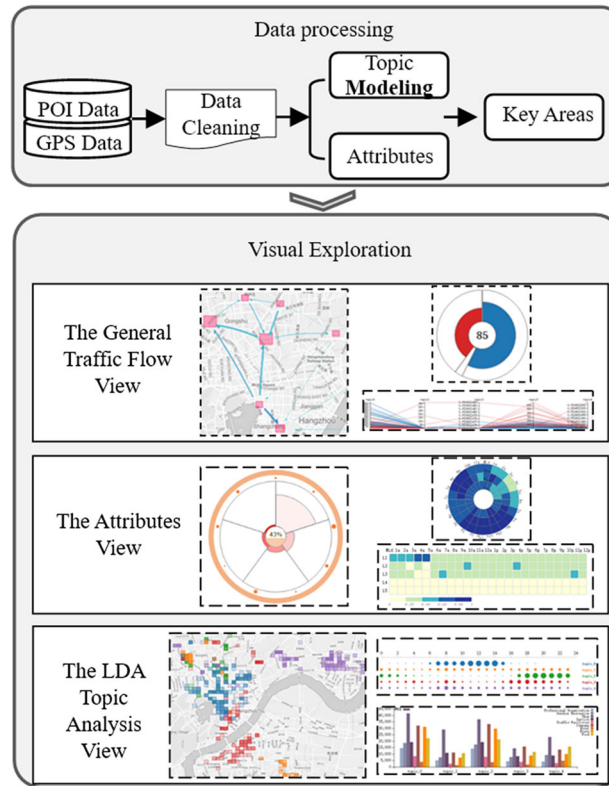


Fig. 1 The system pipeline

4 Data processing based on topic modeling

The LDA model is widely used in text analysis, classification (Cao and Li 2007), partition (Wang and Grimson 2008), pattern recognition (Lazebnik et al. 2006), etc. Hong et al. (2014) proposed to study and explore the unsteady flow field with LDA-based method. The model contains three layers: document, topic, and word. The document contains multiple topics of different proportions, and each topic is composed of a set of words with different probability distribution. The greater a word is related to the corresponding topic, the greater its conditional probability is. With LDA model we can generate a document by selecting topics first and then selecting words according to the topics. According to the inverse of the generative process, LDA model can also be used to deduce the topics based on the words in the document. In this paper, we adopt LDA to inversely deduce traffic topics from the GPS records. The main reason we employ LDA is to detect semantically-important regions from numerous records to reduce the search space for the businessmen to choose shop locations. In the following sub-sections, we first introduce how we employ LDA to find the traffic topics. Then, based on LDA results and traffic attributes on how to extract key areas, i.e., the candidate areas for users to choose store locations are described.

4.1 LDA-based topic modeling

Before LDA modelling, we perform pre-processing to all records to extract trajectories. Each trajectory is treated as a document. All trajectories comprise a corpus of documents. GPS records on the trajectories are treated as words. With such modelled “documents (trajectories)” and “words (records)”, we employ LDA to extract the traffic topics which are composed of different words (records) of different probabilities. As we have mentioned before, the GPS record is a time–space unit, so the extracted topics possess the tempo-spatial properties of trajectory data.

However, directly using the GPS records as words is not possible. If we map the location (represented by longitude and latitude) and the timestamp (represented by hour/minute/second) of a record to a word, the

number of words would be too large and each word only appears once. Such facts make the LDA modelling impossible and meaningless.

In order to make the GPS records suitable for LDA modelling, we uniformly divide the entire main urban area of Hangzhou into 100×100 grid. The grid node is denoted by Gr_i and is of the size about $500 \text{ m} \times 400 \text{ m}$. The location of the records is mapped to the index of the grid node. So the records belong to the same grid node share the same location. Such mapping greatly reduce the spatial range of the words and there are 10^4 different locations altogether. Furthermore, we cluster the GPS records to a series of time intervals $T = \{t_1, t_2, \dots, t_T\}$, using t_i instead of timestamp to represent the time of the record. This also greatly reduces the time complexity of a word. The length of the time interval can be adaptively selected, which is set to be 1 h in this paper. In our implementation, all records of the same hour share the same t_i . After the above simplification of representations of locations and time, we map the records to the “words” to be used in LDA modelling. We use $Gr_i_t_i$ to represent a *temporal-spatial node*, i.e., the “word” in LDA. Multiple records on a trajectory would belong to the same $Gr_i_t_i$, and the occurrence frequency of a node corresponds to the word frequency.

Besides words and documents as inputs, we need to set the number of topics before LDA modelling. In our implementation, we tried different topic numbers from 4 to 10 and compare their results. We find that the results show best performance in semantic modelling when the topic number was 5. The topics would become redundant for numbers larger than 5 and would be not enough to show all different traffic patterns for numbers smaller than 5. The work by Krüger et al. (2012) provides the explanation of topic number being 5 to some degree, in which human activities are divided into five categories, i.e., work, life maintenance, transport, social life, and recreation. The taxi trajectory data depict the temporal-spatial dynamics of individuals, which implies the information of human behaviour and activities. Therefore, we extract five topics from the trajectory data.

Through LDA-based topic modelling, the correlation between the trajectories and topics is obtained, as well as the correlation between the topics and nodes (words). The correlation here refers to the probability of a topic belongs to a trajectory or a word belongs to a topic computed by LDA. That is, for each trajectory we get the probability distribution of different topics, and for each topic we get the distribution of different nodes. These distributions are important for us to understand the semantical importance of the hidden topics of the trajectory data. For example, we would like to know what are the differences among different topics? Does one topic have special semantic meaning, like focusing on commuting or recreation, or concentrating in certain regions? However, users cannot understand the semantics of topics by only referring to the probabilities calculated by the LDA model. Therefore, it is important to design the visualization of the semantic topic modelling results to provide a clear and intuitive understanding of these topics. We will introduce the visualization of LDA results integrated with POIs in Sect. 5.3.

Overall, we summarize three advantages by introducing LDA into the analysis of trajectory data:

1. Obtaining the hidden topic information: The topic information can help us determine the similarity between trajectories. At the same time, the topic information also reflects the characteristics of the correlated nodes.
2. Tolerating errors: The LDA model is a probabilistic generative model that has a high fault tolerance and is capable of reducing the accuracy problem of dirty data. Thus, it greatly reduces data correction work which is needed for the traditional trajectory data.
3. Reducing data dimensions: Each trajectory consists of thousands of different records. The LDA model correlates each trajectory to several topics, using topics to describe the trajectory data, thus greatly reducing the data dimensions.

4.2 Extraction of key nodes

Choosing the appropriate shop addresses for users in a big city is equivalent to searching for the solution in an infinite space. After LDA analysis, we select the top 500 correlated words (nodes) to represent each topic. These top-ranked nodes can be used as candidate areas for user to choose.

However, it is important for the candidate areas to have not only semantic importance, but also important attributes, e.g., the busyness of the grid node. In this paper we define *Point visit* ($Pv(Gr_i, t_i)$) as the node attribute to be taken into account for selecting key nodes:

1. *Point visit* ($Pv(Gr_i, t_i)$) refers to the number of GPS records included in each grid node Gr_i for time interval t_i . It shows the frequency of taxis visiting the node, and reflects the busyness and importance of the node's traffic.

We combine $Pv(Gr_i, t_i)$ with node semantic correlation to calculate *key nodes* so that we can retain the semantic information and consider the traffic situation at the same time. As we mentioned before, for each topic we choose 500 most correlated nodes. Since we set the topic number to be 5, so there are 2500 nodes left after topic modelling, and each node holds a correlation value denoted as $r(Gr_i, t_i)$. We can not directly add $Pv(Gr_i, t_i)$ and $r(Gr_i, t_i)$ to get the final score to determine the key nodes since both values are not in the same orders of magnitude. Before adding these two values, we perform the following normalization process.

First, the min-max method is used to normalize the values to the range of $[0, 1]$, as shown in formulas (1) and (2) where Pv_{\max} and Pv_{\min} represent the maximum and minimum values of v respectively; and r_{\max} and r_{\min} represent the maximum and minimum values of r , respectively.

$$Pv' = (Pv - Pv_{\min}) / (Pv_{\max} - Pv_{\min}), \quad (1)$$

$$r' = (r - r_{\min}) / (r_{\max} - r_{\min}), \quad (2)$$

After the transformation, we further perform equalization to make the distribution of both values similar:

$$Pv'' = (Pv' - u_{v'}) / \sigma_{v'}, \quad (3)$$

$$r'' = (r' - u_{r'}) / \sigma_{r'}. \quad (4)$$

Among them, $u_{v'}$ and $\sigma_{v'}$ respectively indicate the mean and standard deviation of v' , and $u_{r'}$ and $\sigma_{r'}$ respectively represent the mean and standard deviation of r' .

$$s = Pv'' + r'' \quad (5)$$

By adding the two measures according to formula (5), we obtain the score s to choose key nodes. We consider the nodes whose s greater than 0 as the key nodes and extract these key nodes for each time interval.

Figure 2 shows the distribution of the top 20 key nodes on the map from 9 a.m. to 10 a.m. based on different metrics. Each red rectangle represents a grid node. Figure 2a is the result based on Pv , Fig. 2b shows the result with only topic correlation value considered, and Fig. 2c is the result of the combination of the two metrics. As shown in Fig. 2c, after combining two metrics, we not only retain important nodes under either individual metric, but also find new nodes that are important for both metrics. This approach directly and quickly identify the nodes that have good attributes and semantic features, greatly reducing the huge space caused by massive data.

5 Visual design

We develop a visual analytics system to visualize and explore the heterogeneous urban data. Besides the previously-introduced three coupled views, we design the dashboard view for users to set parameters and switch to the other three different views. The users interactively explore these views cooperatively to search for the results fulfilling their requirements. In the following subsections, we introduce the visual design of these four views.

5.1 The general traffic flow view

Figure 3a-c shows the three sub-views contained in the general traffic flow view for the time interval from 10 a.m. to 11 a.m. Figure 3a is the map view, and Fig. 3b and c are the parallel coordinates and the arc chart, which present the relations between trajectories and topics from different perspectives.

5.1.1 The map view

In Fig. 2, there are some connected key nodes. These connected nodes often constitute one whole area and can be merged as a "super node". A single node marks a location that is suitable for small businesses, whereas the "super node" is more appropriate for large-scale shopping malls. We employ the DBSCAN algorithm to merge these connected nodes. The E neighborhood in DBSCAN is the "nine-box" region. The

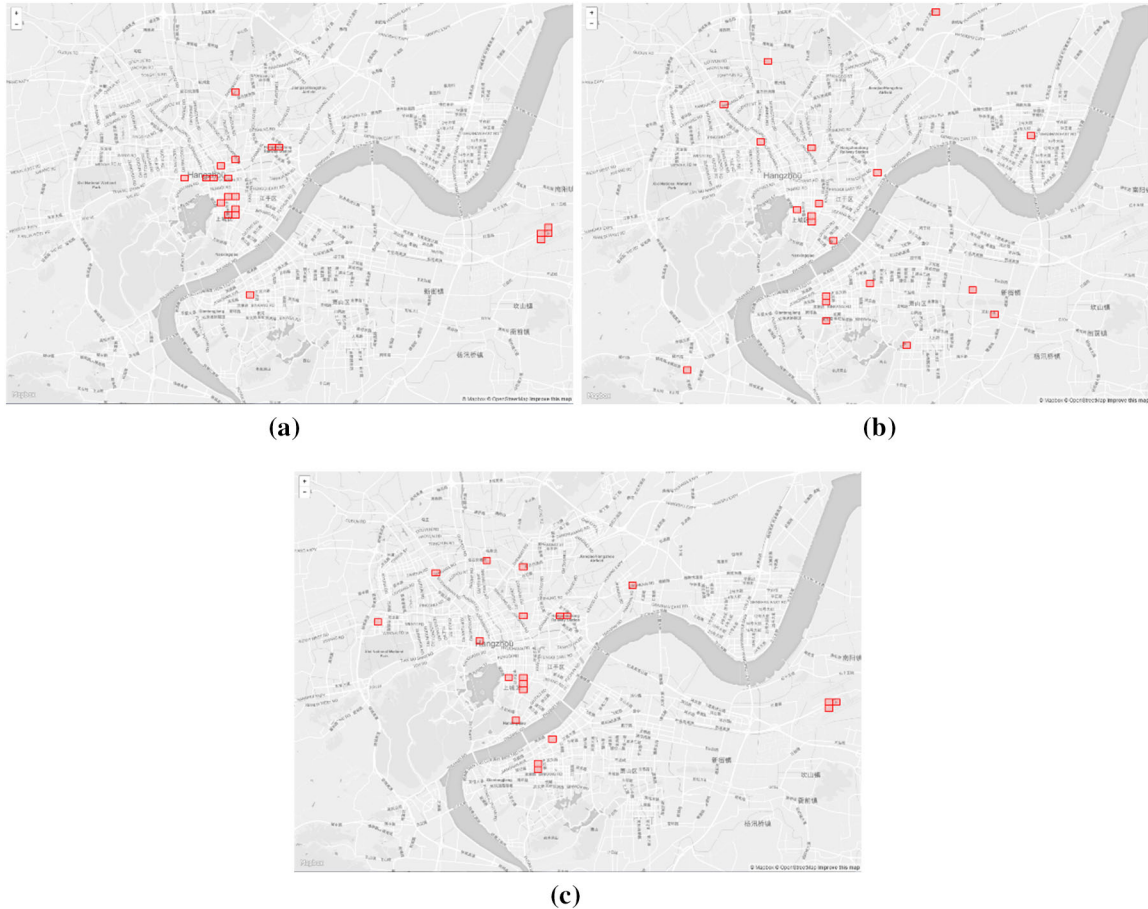


Fig. 2 Key nodes selected on the map based on different metrics. **a** Key nodes under the attribute of point visit; **b** key nodes under the topic-node correlation values; **c** key nodes selected by combining the above two metrics

core point is judged by whether the score s is greater than 0. Figure 4 shows the result of merging by DBSCAN based on Fig. 2c, in which the previous connected nodes are represented as a larger red rectangle. Such merged super nodes are treated in the same way as the single node and the “nodes” in the later parts of this paper can be referred to either a single node or a super node.

In order to generate the general traffic flow among the key nodes, we partition the map into a Voronoi graph with the key nodes as the center points of each region. We compute the traffic flows between adjacent nodes. The set of trajectories starting from A to B or from B to A constitute the in–out flows between A and B. For the merged super node $A = \{a_1, a_2, \dots, a_i, \dots, a_n\}$, we calculate the trajectories between each single node a_i in A and the area B. Finally, the total number of trajectories is calculated by summing up all non-repetitive trajectories.

Mapping the key nodes to rectangles on the map helps users understand the spatial distribution of the key nodes that have convenient traffic and high passenger volume. A traffic flow graph is constructed for displaying the summary transportation conditions among key nodes. The thicker the flow lines are, the greater the flows are. The arrow indicates the larger flow direction (in or out) of two connected areas. The difference between flow amounts along in/out directions is encoded by different intensities of the blue color, with the darker blue color indicating a larger difference. This traffic flow view helps users understand the general flow patterns and passenger flow rules quickly.

From Fig. 3a, we find that the traffic flows from left to right and bottom to top. The area A outlined by the black circle is a crowded area since the thick and dark flows from B and C to A indicate a large number of taxis going from B and C to A and fewer from A to B and C. The color of flow between A and D is light, indicating that the number of trajectories in both directions is similar, whereas that of B is the opposite, indicating that many trajectories flow out from area B.

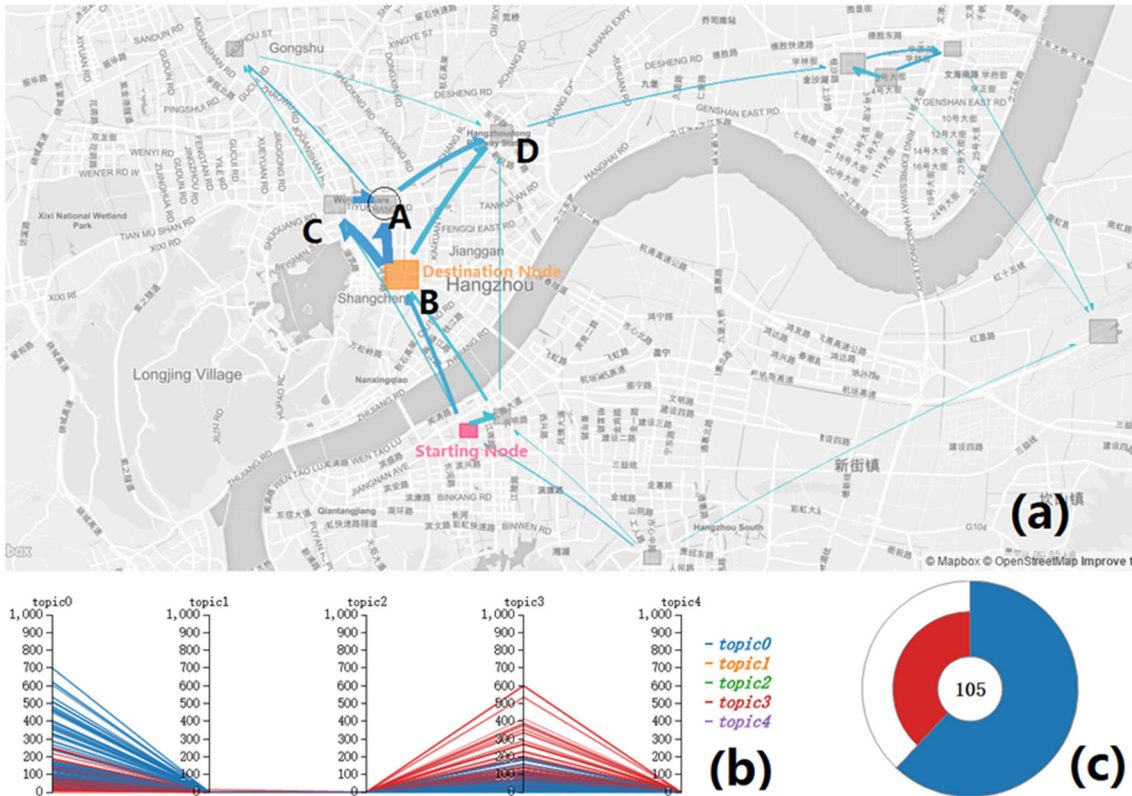


Fig. 3 The General traffic flow view. **a** The map view shows the general traffic flow; **b** the parallel coordinates shows the correlation to five topics for selected traffic trajectories; **c** the arc chart shows the topics distribution of selected traffic trajectories



Fig. 4 The nodes-merging result by DBSCAN

In the traffic flow view, users can examine topic features of specific flows between two nodes by clicking the starting and the destination nodes, which are highlighted in colors of rose red and oranges respectively in Fig. 3a. Other nodes are suppressed to be grey in color. The lower partial view of the arc chart and the parallel coordinate show the topic features related to the trajectories (Fig. 3b, c).

5.1.2 The arc chart

The arc chart visualizes the general topic distributions of the selected traffic flow. The number in the inner circle reveals the total trajectory number between two nodes. The trajectories are counted in terms of their most related topics and the statistical results are visualized by distributions of sections. The diameter of each sector is proportional to the sum of the topic correlation of each trajectory. From Fig. 3c we can see there are 105 trajectories in the selected flow, and most are related to topic 0 and topic 3.

5.1.3 The parallel coordinates

The parallel coordinate shows how trajectories correlate to the five topics (Fig. 3b). Each trajectory is visualized by its correlation factors to five topics and colored by the most related topic. Through the parallel coordinates, users can observe the changes of distributions of the trajectories on each topic, allowing them to gain an intuitive understanding of the trajectory types in terms of the topic distribution. As shown in Fig. 5, the users can select certain trajectories by restricting the value ranges at any coordinate axis. In Fig. 5a and b, the users respectively selected trajectories most related to topic 0 and 3, we find that the trajectories most related topic 0 always has a lower correlation with topic 3 and vice versa, indicating that topic 0 or topic 3 are exclusively represented by these trajectories. Besides, the corresponding trajectories related to the selected trajectories are also shown on the map which is helpful for users to discern the geographical difference between these two types of trajectories. In Fig. 5a, the trajectories associated with topic 0 are biased towards through the Fuxing Bridge to reach the target area and hovering around the target area. In Fig. 5b, besides the Fuxing Bridge, the trajectories most related to topic 3 also choose the Xixing Bridge to reach the target area, and hover around the start area.

5.2 The attributes view

The users want to inspect and compare traffic attributes like traffic speed or volume of different nodes in order to make their selection of store locations. Besides *Point visit* ($Pv(Gr_i, t_i)$) defined in Sect. 4.2, the following four attributes are calculated for each node:

2. *Traffic volume* ($Tv(Gr_i, t_i)$) refers to the number of trajectories passing each Gr_i during interval t_i . It represents the taxi flow in the node and reflects the convenience of the node.

3. *Taxi state* ($Tst(Gr_i, t_i)$) indicates the taxi status as empty or occupied. The proportion of occupied status of each Gr_i during interval t_i is calculated, which represents the passenger occupancy of taxis and indicates the number of potential customers in the node.

4. *Traffic speed* ($Tsp(Gr_i, t_i)$) indicates the distribution of taxi speeds in the grid node Gr_i during interval t_i . Similar to the work by Al-Dohuki et al. (2017), we divide the speeds into five levels L1–L5 from low to fast. The distribution of traffic speed at different levels is calculated.

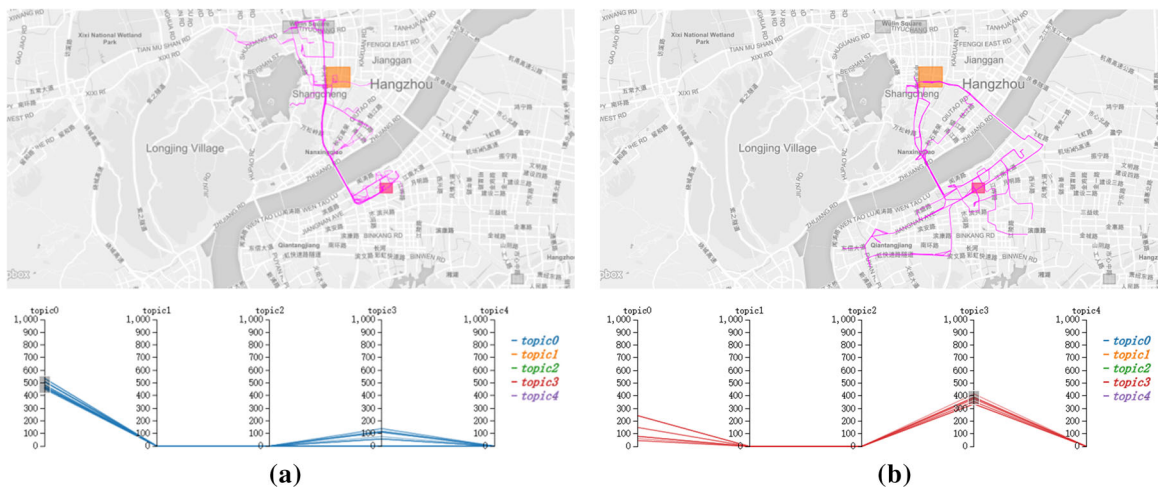


Fig. 5 Selecting different trajectories through parallel coordinates. **a** High correlation with topic 0. **b** High correlation with topic 3

5. *Environment* calculates the number of different POIs in each category within grid node Gr_i . It reflects the market characteristics and circumstances in the node.

Although seems similar at first sight, there is subtle yet important difference between *Point visit* Pv and *Traffic volume* Tv . These two indicators are not absolutely proportionally-related. For example, a node of large Pv may not be that large in Tv , since many taxis may just hang around in that area without going out. Larger value of Tv indicates higher taxi flow in and out at both time so it can be more properly used as the indicator for the smooth and convenience of the traffic in the node. Larger value of Pv indicates that taxi drivers pay many times of visits there, so it is more likely to use as the indicator of the traffic importance of the node.

We design the attribute glyph to show multiple attributes of key nodes simultaneously so that users can compare different nodes in terms of all attributes conveniently. Besides the attribute glyph, we also design two heatmaps to show how different attributes change over time within a node.

5.2.1 The attribute glyph

A familiar metaphor in people's daily lives greatly enhances comprehension and reduce the cognitive burden. Inspired by the vehicle wheel, we design a radial-based visual metaphor to represent multiple attributes of a key node as shown in Fig. 6.

The radius of the entire glyph is proportional to the point visit $Pv(Gr_i, t_i)$. The proportion of taxi status $Tst(Gr_i, t_i)$ being occupied is identified by the fill of the inner circle with the number shown as 43% in Fig. 6. Five uniformly-divided sectors outside the inner circle respectively represent the five speed levels $Tsp(Gr_i, t_i)$ in the clockwise direction, with the darker red colors indicating higher speed levels. The radius of each sector is proportional to the number of records having speed levels of this sector. In Fig. 6, the section with the lightest red corresponding to speed level 1 has the largest radius, and the radii of sections for speed levels 2–5 are shown in descending order. This indicates most of the taxis move slowly and the traffic in this area may be congested. The POI details are shown as nine dots outside the sectors, with each dot representing a POI category. In the clockwise direction, the dots represent professional organization, outdoor recreation, shopping, service, traffic facilities, company, hotel, and estate and food. The names are listed beside the dots. We use the size of the dot to indicate the number of different POIs in each category. The outer circle represents the traffic volume $Tv(Gr_i, t_i)$. The thicker the outer circle is, the larger the traffic volume $Tv(Gr_i, t_i)$ is.

5.2.2 The matrix heatmap

Except for POI environment, the other attributes change over time. Different users care about different time periods based on their own needs, e.g., the breakfast shop and the bar have different business hours. We design two heatmaps to represent the attribute changes over time.



Fig. 6 The attribute glyph which shows the five attributes of a key node

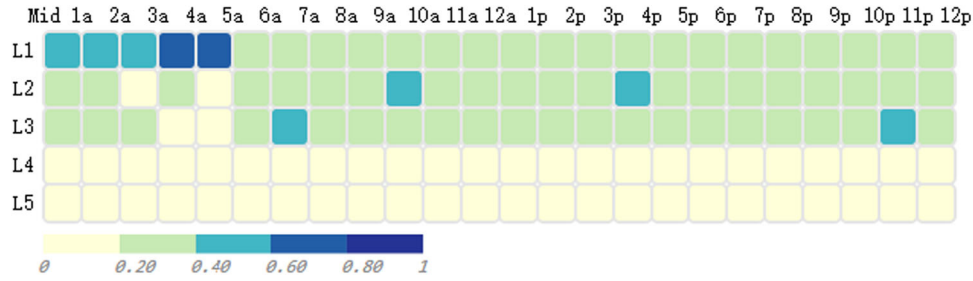


Fig. 7 The matrix heatmap showing speed changes over time

In Fig. 7 we show the change in speeds over time. The scale mapping proportion to color is shown at the bottom. The horizontal axis shows 24 h and the vertical represents five speed levels. Each cell in the heatmap corresponds to the ratio of the certain speed level within a time interval, wherein the darker the color is, the larger the proportion. By using this matrix heatmap, users can easily find how speed changes over time and the speed levels with large proportions in any time interval.

5.2.3 The annular heatmap

The annular heatmap represents the changes of the other three traffic attributes over time, i.e., traffic state $Tst(Gr_i, t_i)$, point visit $Pv(Gr_i, t_i)$ and traffic volume $Tv(Gr_i, t_i)$. In Fig. 8, the circular ring is equally divided into 24 parts angularly, representing 24 h. Each sector of the inner ring shows traffic state $Tst(Gr_i, t_i)$. The sectors of the middle ring and the outer ring represent point visit Pv and traffic volume Tv . The color of the sector indicates the proportion of these three values to the corresponding largest value during a day. The color scale is the same with Fig. 7.

5.3 The LDA topic analysis view

Figure 9 shows the LDA topic analysis view, where (a) is a spatial topic map showing the spatial distribution of representative words of the five topics, (b) is a matrix scatter plot showing the temporal rules of the five topics, and (c) is a group bar chart showing the POI information of the five topics.

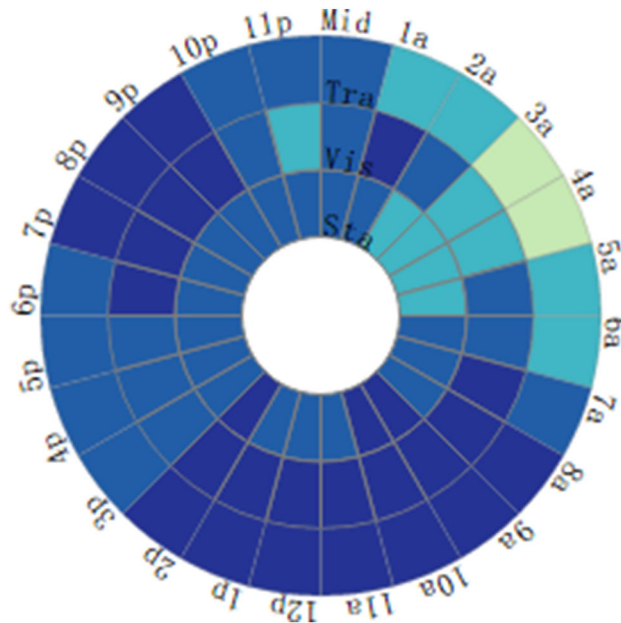


Fig. 8 The annular heatmap showing how three attributes change over time, namely traffic state Tst, the point visit Pv, and the traffic volume Tv

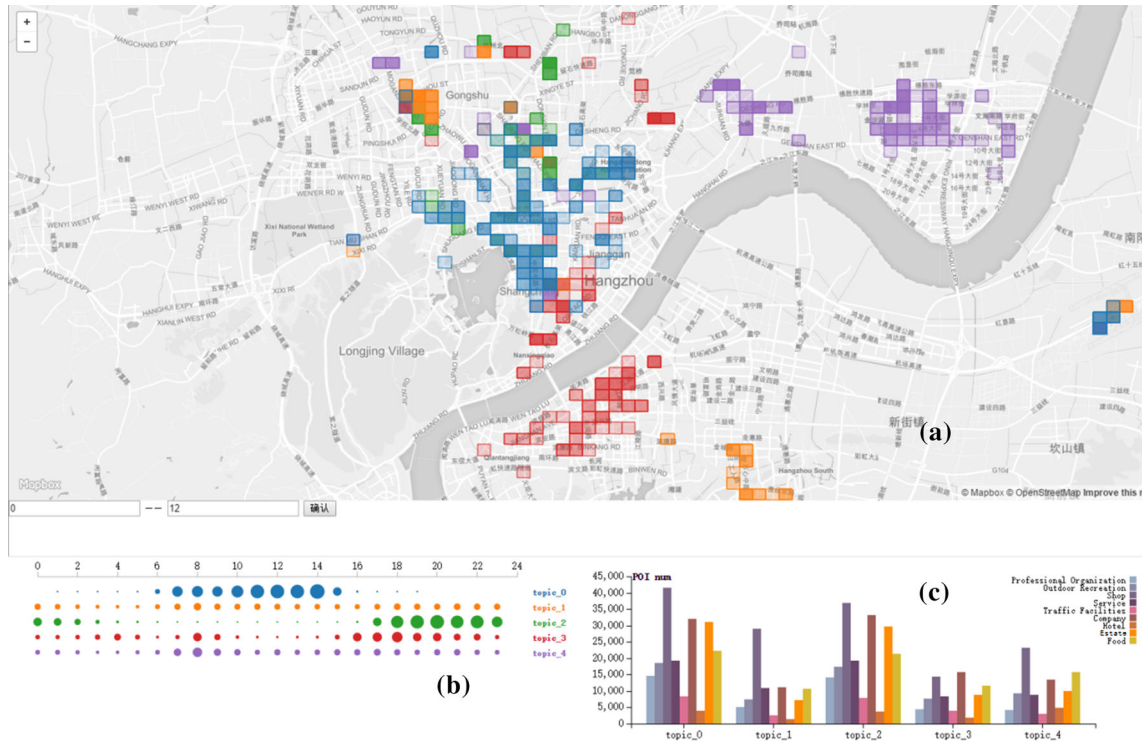


Fig. 9 The LDA topic analysis view. a The spatial Topic Map; b the temporal matrix scatter plot; c the group bar chart

The temporal matrix scatter plot

To find the temporal characteristics of topics, we count the number of words belonging to each time interval within the 2500 most related words of the five topics. The results are presented in Fig. 9b. The horizontal line shows 24 time intervals and the vertical are the five topics. The size of the point indicates the number of words during specific time interval for a certain topic. The color of the point is the same as the corresponding topic color. From Fig. 9b, we find each topic has its own temporal features. For example, topic 0 (in blue) is more active during daytime, which is related to people’s activities during the day, whereas topic 2 (in green) is active mainly at night, thus related to people’s night-time activities. The active time of topic 3 (in red) coincides with people’s commuting time, and is more probably related to people’s working activities. Meanwhile, topic 1 (in orange) and topic 4 (in purple) have similar statistical distributions over time without obvious emphasis. These five topics are all inactive between 3 a.m. and 6 a.m. since most people are resting/sleeping during this time period.

5.3.1 The spatial topic map

We analyse the spatial distributions of different topics during certain time periods. We draw the top 2500 words as rectangles on the map for different time periods (Fig. 10), in which the color of the rectangle is set to be the same as the most relevant topic. As shown in Fig. 10c–f, the words of Topic 0 (blue) are mainly distributed in the downtown part, and it shifts to topic 2 (green) during the night. From Fig. 10a–f, topic 1 (orange) is always in the area of the Fuyang, Xiaoshan, and Gongshu districts, which are of urban border areas, whereas topic 3 (red) is mainly along the Qiantang river, especially Binjiang district. The words of Topic 4 (purple) mainly locate at the Xiasha higher education zone and Jianggan District. In Table 1 we summarize the temporal, spatial and POI features of five LDA topics. Based on these three features, the semantic information implied in LDA topics is analysed and shown in Table 1.

5.3.2 The group bar chart

In Fig. 9c, the five groups of bar charts correspond to five topics. Each group contains nine bars, representing the number of POIs in nine categories. We find that the “Shop” category stands out from any other

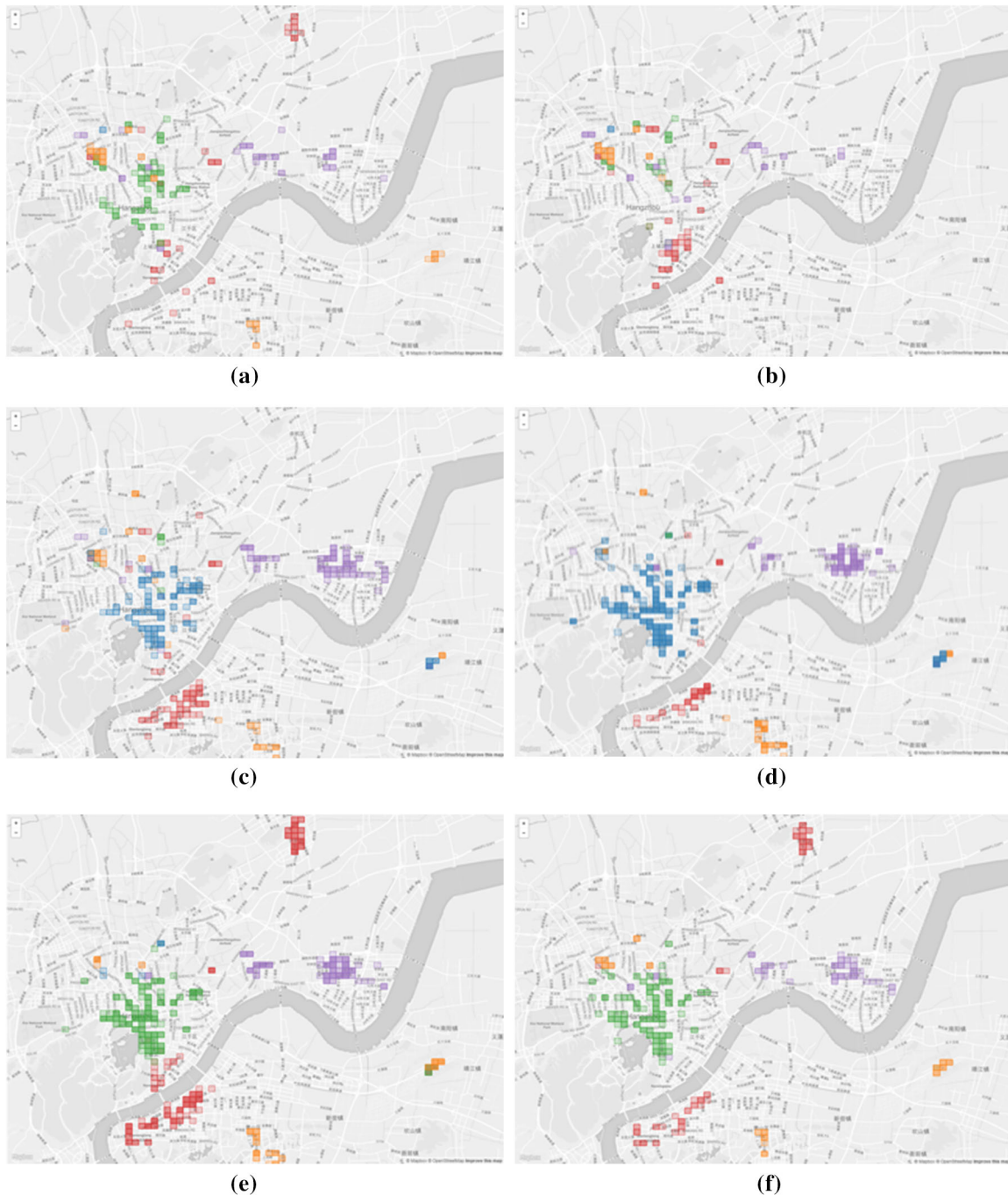


Fig. 10 Spatial distributions of topics in different time periods. **a** 0:00 a.m.–3:00 a.m. **b** 3:00 a.m.–6:00 a.m. **c** 6:00 a.m.–10:00 a.m. **d** 10:00 a.m.–4:00 p.m. **e** 4:00 p.m.–9:00 p.m. **f** 9:00 p.m.–12:00 p.m.

topics, indicating that shopping is booming in this city. The “company”, “estate” and “food” are also evident which is in line with people’s daily lives. In addition, each topic has its own characteristics in terms of the POI categories. For topic 0 and topic 2, the POIs are similar and the number in each category is generally high. Therefore, the areas covered by these two topics are comprehensive areas integrated with business, tourism, culture, and entertainment. The “Shop” category is the most prominent indicator in topic 1, followed by company, service, food, and estate. Other categories, such as traffic facilities and hotels, are few. This indicates that, in these areas, the traffic is relatively light and most are occupied by inhabitants. The areas covered by topic 1 can be regarded as being comfortable and relaxed. In topic 3, “Company” has

Table 1 The temporal, spatial, POI features and semantic information of different LDA topics

	Temporal feature	Spatial feature	POI feature	Semantic information
Topic 0	Active in daytime	Downtown	Large numbers in different categories	A comprehensive daytime mode integrated business, tourism, culture and entertainment
Topic 1	No obvious change	Fuyang, Xiaoshan, and other peripheral regions	Lack in traffic facilities and hotels, however living facilities with are complete	Most are inhabitants who prefer comfortable and recreational life
Topic 2	Active in nighttime.	Downtown	Large numbers in different categories	Integrated night mode for people who prefer to go out at night
Topic 3	Commuting time	Binjiang District and Yuhang District	It has the largest number in the category of “Company”	The mode of working people
Topic 4	No obvious change	Xiasha higher education campus	Mostly “Shopping” “food”, “hotel” accounted for a relatively large proportion	The entertainment mode of university students

the largest number, and the active time coincides with the rush hour, indicating that topic 3 is an area related to work-related activities. For topic 4, except “shop,” “food” is also the main focus, and compared with other topics, the category of “hotel” significantly accounts for a larger portion here. Hence, the topic 4 is more related to entertainment services.

5.4 The dashboard view

The dashboard view (Fig. 11) is used for parameter setting, which includes the time interval, the buttons to change to the other three views, and the ratio of key nodes to display. Once the time interval is selected, the three linked visual views display the information of the key nodes during the corresponding time period.

6 Case study

In Fig. 12 we show the user interaction in our system. Users can deepen their understanding of the city through convenient interactions so as to select a satisfactory store location. First, users select the time period and adjust the ratio of key nodes displayed on the map in the Dashboard View. The system draws the regional traffic flow in the General Traffic Flow view, and the user observes the distribution of the key nodes and the traffic flow directions. Then, users switch to the LDA Topic Analysis View by clicking the LDA button in the Dashboard View. According to the nature and requirements of the planned store, users select the appropriate time period for topic semantic analysis and determine the matching LDA topic. Then users return to the General Traffic Flow View to zoom in the area covered by the selected topic to narrow down the selection range and find out the candidate areas. After that, users switch to the Attributes View to observe and compare the attribute characteristics of the candidate areas. After selecting an area of interest, the system displays heatmaps of the area attributes to help users understand the attribute distribution over time of the area. Through the comparison of the attributes, the user determines the area they select. Finally, the users can study the trajectories in the selected area by displaying the topic-related trajectories on the map

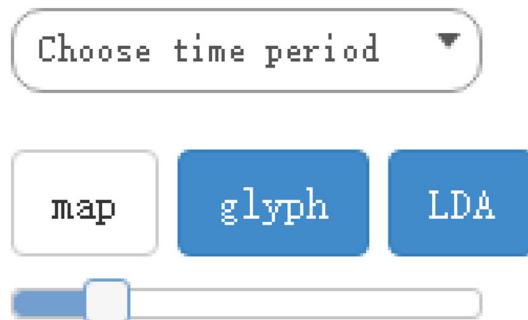


Fig. 11 The dashboard view

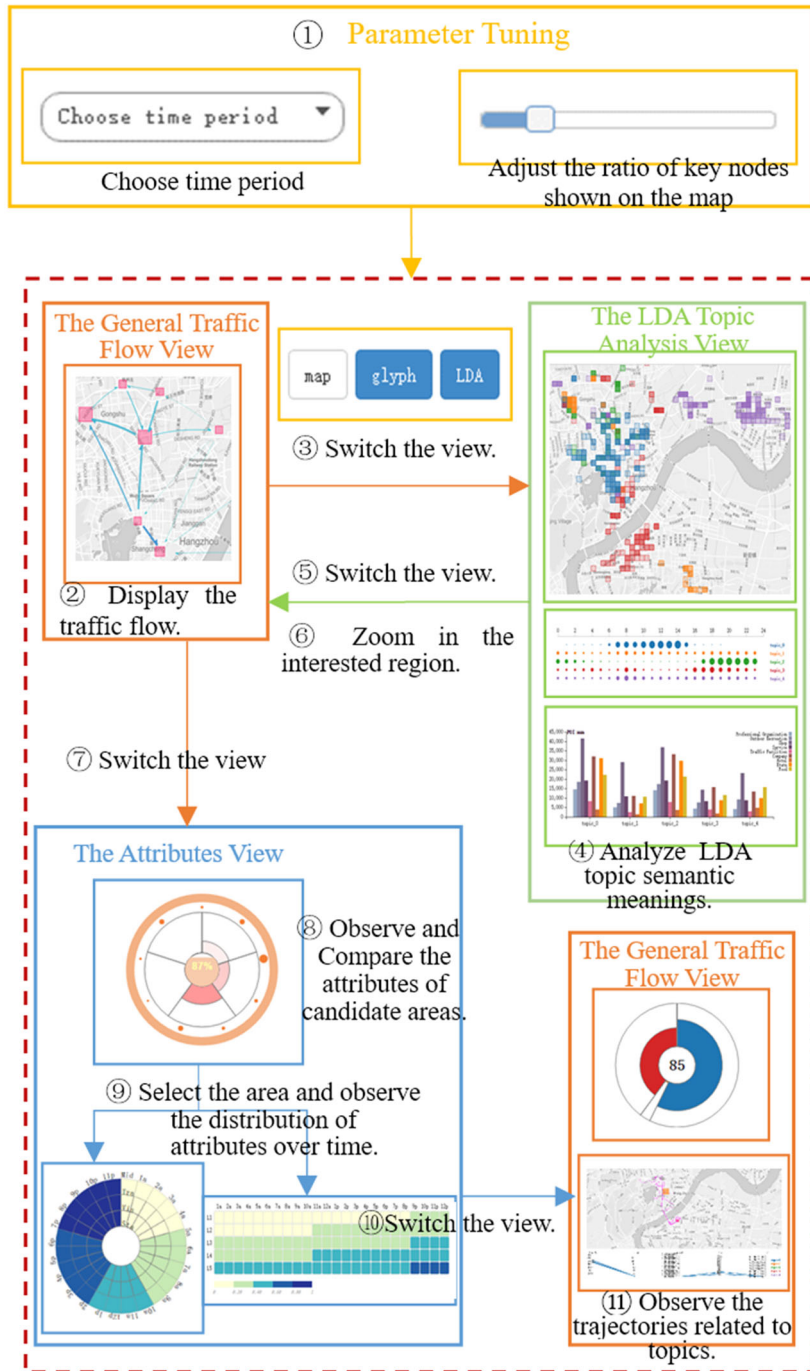
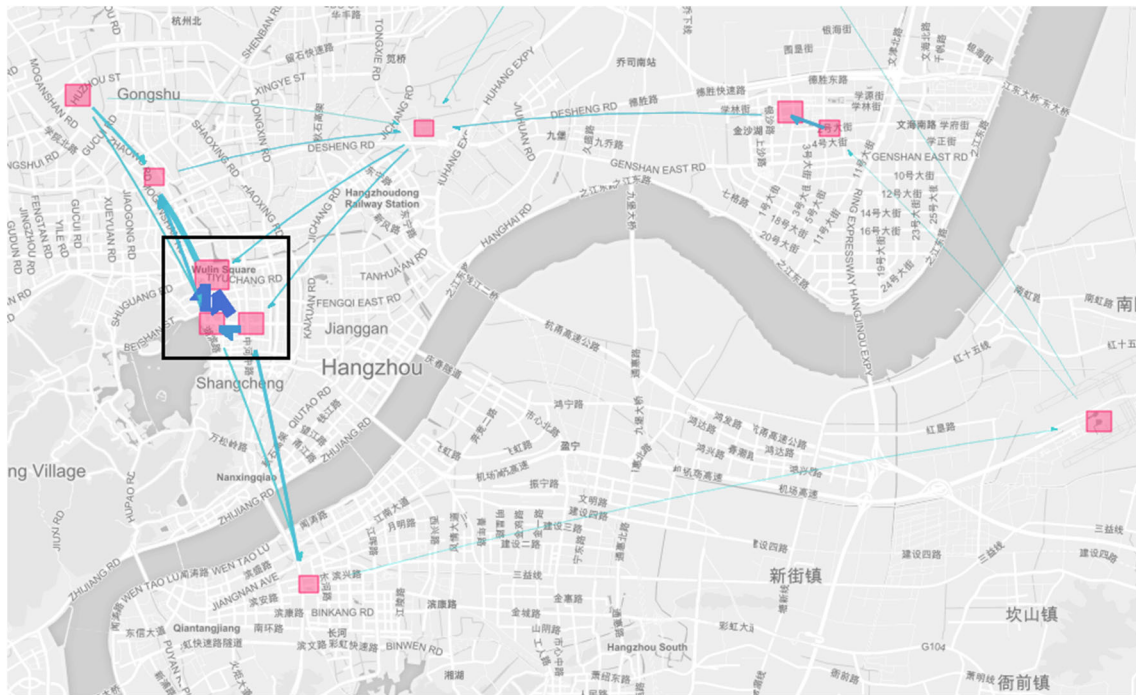


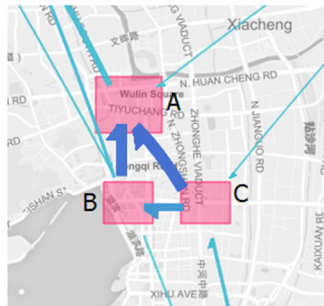
Fig. 12 The process of user interaction

through selection in the parallel coordinates of the General Traffic Flow View. The distribution of trajectories show busy streets for users to consider.

We illustrate how our visual analysis system helps users decide the store addresses through a case study—a user who wants to open a medium-sized middle-top-grade bar in Hangzhou uses our system to find the appropriate bar address. Bars are normally open at night, and the passenger flow peak always starts from 10 p.m. So, we first choose the period of 10 p.m. to 11 p.m. to observe the traffic flow in the general traffic flow view (Fig. 13a). The view shows the distribution of key nodes in the urban area of Hangzhou; however, the user needs more information to make a decision. Switching to the LDA view and observing the



(a)



(b)

Fig. 13 The general traffic flow map and the enlarged view of the three key nodes. **a** The general traffic flow map from 10 p.m. to 11 p.m. (with 60% key nodes displayed set by the dashboard); **b** the traffic details of the black box in (a)

distribution of LDA topic during this time period (Fig. 14), we can see that the active topics are topic 2 (green), topic 3 (red), and topic 4 (purple). According to the topics' temporal distributions shown in Fig. 9b and the POI distribution in Fig. 9c, we know that topic 2 represents a night activity pattern in the comprehensive areas, topic 3 shows the pattern of office workers in the working areas, and topic 4 shows the activity pattern in the entertainment area. We further observe that topic 4 is distributed in the Xiasha higher education park, so college students are the main force of consumption in this area. The middle-top-grade bar requires potential customers have higher consumption capacity, so topic 4 is not suitable.

Meanwhile, the area where a middle-top-grade bar is located should contain a large number of consumers, and be close to commercial centers, cultural centers, or tourist attractions where consumer spending is generally higher. It would be better if some entertainment places can be found nearby, since different types of customers can be attracted. Thus, the areas covered by topic 2 are more suitable for the above requirements.

As shown in Fig. 13a, the area framed out by black rectangle have greater traffic flow, indicating heavy traffic. However, area A and area B have many in-flows, whereas area C mostly has outflows (Fig. 13b). Thus, the bar in area C has the risk of losing customers.

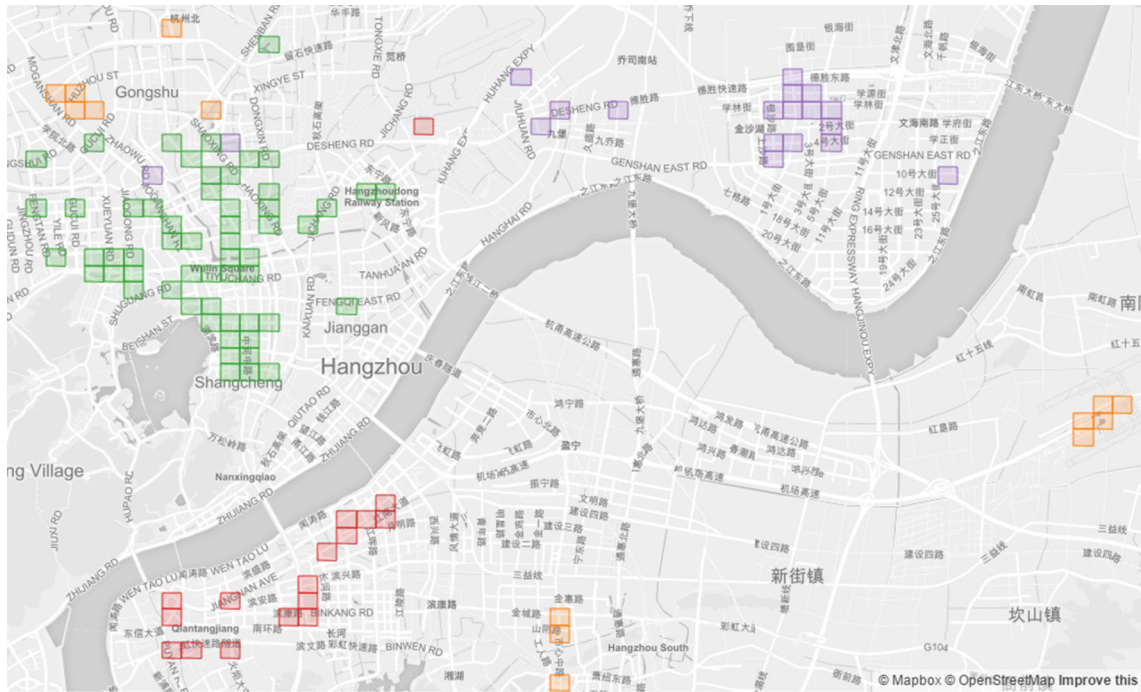


Fig. 14 Spatial distribution of topics from 10 p.m. to 11 p.m

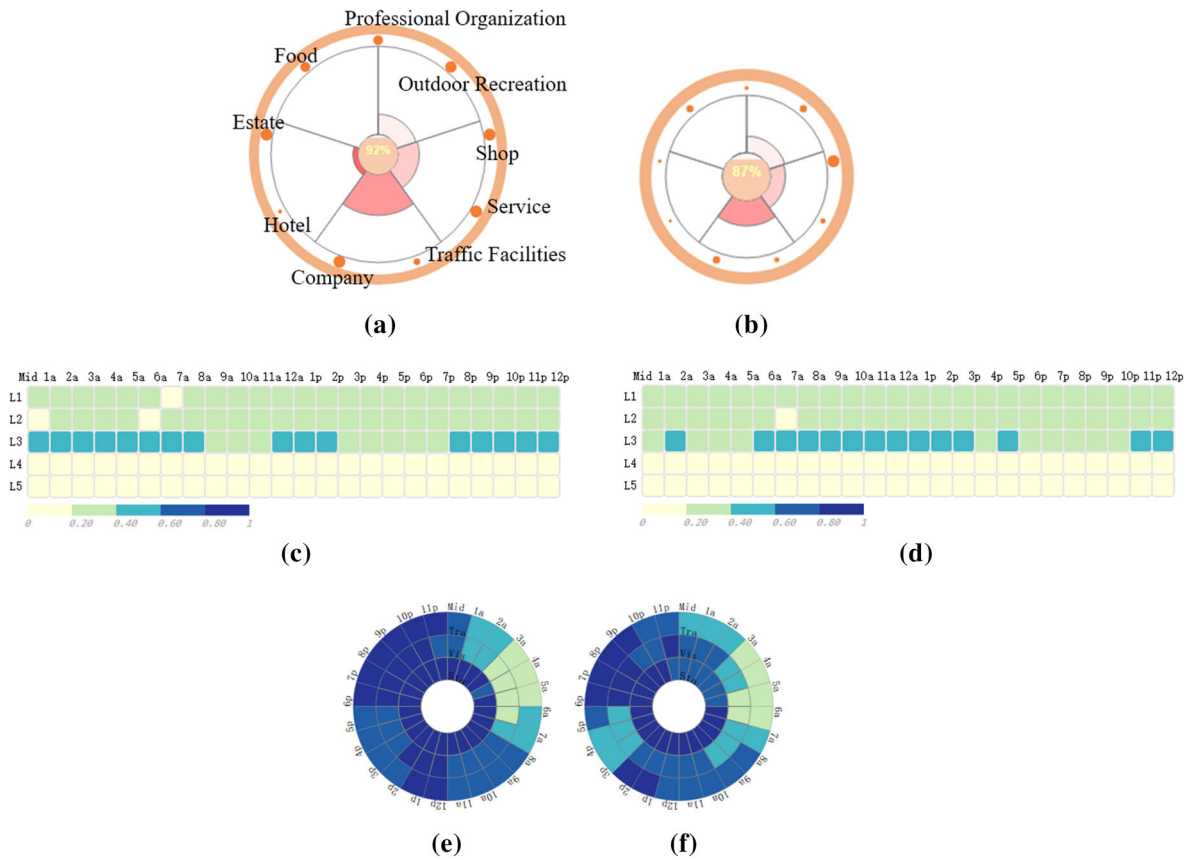


Fig. 15 Comparison of the attributes of node A and node B. a The glyph of node A. b The glyph of node B. c Matrix heatmap of node A. d Matrix heatmap of node B. e Annular heatmap of node A. f Annular heatmap of node B

We further observe the attributes views of the two nodes A and B, as shown in Fig. 15. The speed changes over time of the two nodes are shown in Fig. 15c and d. We see that traffics in both areas are similar between 10 p.m. and 2 a.m. The majority cars are in L3 speed levels which indicate normal traffic condition without congestion. From Fig. 15e and f we see two nodes are relatively busy between 6 p.m. and 1 a.m. since these sections have darker colors. Hence, both are good locations for opening a shop for night-time activities. The biggest differences between the two areas are the point visit and POIs, as shown in Fig. 15a and b. The radius of the glyph of A is larger than that of B, however since node A covers a larger area, more point visits does not necessarily mean larger customer density. As for POIs, node B has significantly less estates and professional organizations than node A. Given that bars are generally noisy at night, these are likely to generate complaints from nearby inhabitants. All things considered, node B is a more suitable location for bars than node A.

In order to verify our result, we investigate the regions corresponding to these two nodes. Node A is the site of Wulin square, the largest shopping district and cultural square in Hangzhou. People often come here for recreation, so the area is crowded and the operating cost is very high. Area B covers the famous Hu Bin LiuGongYuan bar street, which is home to many popular bars, such as Feibi Bar, Feiwen Bar, Baidu Bar, etc. In addition, it is very close to the business center of area A. Thus, many complementary places and facilities can be found in nearby regions. This indicates that the area is a potentially good location to open a bar. However, in order to stand out, the drinks and services provided should be distinctive.

7 Conclusion

This paper proposes a method that can assist business users in choosing store/business locations through the analysis of trajectory data. This approach proposes an interactive visual analysis system. Using the massive data available, we effectively extract the characteristics of specific areas in the city through LDA-based topic modelling. Then, we design a variety of visual methods to display the results intuitively and clearly.

Now we determine an area for users to open a shop, which provides a range instead of an accurate site address; in our future study we need to incorporate the trajectories and street names to generate more accurate and concrete solutions to users. Besides, to make this system more scalable, we expect to determine the number of topics through the accurate measurement of “perplexity” proposed by Blei et al. (2003). We also plan to provide users with more interactions, such as the choice and adjustment of attribute metrics.

Acknowledgements The authors wish to thank anonymous reviewers for their pertinent and insightful reviews, which were of great importance in improving the quality of this work. This work was supported by the National Science Foundation of China (Grant No. 71571160, 61672462).

References

- Al-Dohuki S, Wu Y, Kamw F, SemanticTraj et al (2017) A new approach to interacting with massive taxi trajectories. *IEEE Trans Visual Comput Graphics* 23(1):11–20
- Andrienko G and Andrienko N (2008) Spatio-temporal aggregation for visual analysis of movements. 2008 IEEE Symposium on Visual Analytics Science and Technology, pp 51–58
- Andrienko G, Andrienko N, Wrobel S (2007) Visual analytics tools for analysis of movement data. *SIGKDD Explor Newsl* 9(2):38–46
- Andrienko N, Andrienko G, Stange H et al (2012) Visual analytics for understanding spatial situations from episodic movement data. *KI-Kunstliche Intelligenz* 26(3):241–251
- Andrienko G, Andrienko N, Bak P et al (2013a) *Visual analytics of movement*. Springer, Berlin
- Andrienko N, Andrienko G, Fuchs G (2013) Towards privacy-preserving semantic mobility analysis. In: *Proceedings of International EuroVis Workshop on Visual Analytics*. Eurographics Association Press, pp 19–23
- Blei DM (2012) Probabilistic topic models. *Commun ACM* 55(4):77–84
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Cao L, Li F (2007) Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In: *Proceedings of the IEEE 11th International Conference on Computer Vision*. IEEE Computer Society Press, pp 1–8
- Chen W, Guo F, Wang FY (2015) A survey of traffic data visualization. *IEEE Trans Intell Transp Syst* 16(6):2970–2984
- Chen Z, Wang Y, Sun T et al (2017) Exploring the design space of immersive urban analytics. *Visual Informatics* 1(2):132–142
- Chu D, Sheets D. A, Zhao Y, et al. (2014) Visualizing hidden themes of trajectories with semantic transformation. In: *Proceedings of IEEE Pacific Visualization Symposium*. IEEE Computer Society Press, pp 137–144
- Deerwester S, Dumais ST, Furnas GW et al (1990) Indexing by latent semantic analysis. *J Am Soc Inf Sci* 41(6):391–407
- Ferreira N, Poco J, Vo HT et al (2013) Visual exploration of big spatial-temporal urban data: a study of New York city taxi trips. *IEEE Trans Visual Comput Graphics* 19(12):2149–2158

- Guo D (2008) Regionalization with dynamically constrained agglomerative clustering and partitioning (redcap). *Int J Geogr Inf Sci* 22(7):801–823
- Guo D (2009) Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Trans Vis Comp Graphics* 15(6):1041–1048
- Guo H, Wang Z, Yu B, et al. (2011) Tripvista: triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection. *Proceedings of IEEE Pacific Visualization Symposium*. IEEE Computer Society Press, pp 163–170
- Hofmann T (1999) Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 50–57
- Hong F, Lai C, Guo H et al (2014) FLDA: latent Dirichlet allocation based unsteady flow analysis. *IEEE Trans Visual Comput Graphics* 20(12):2545–2554
- Karamshuk D, Noulas A, Scellato S, et al. (2013) Geo-spotting: mining online location-based services for optimal retail store placement. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 793–801
- Krüger R, Lohmann S, Thom D, et al. (2012) Using social media content in the visual analysis of movement data. *Proceedings of 2nd workshop on interactive visual text analytics*
- Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society Press, pp 2169–2178
- Liao ZF, Li Y, Peng Y et al (2015) A semantic-enhanced trajectory visual analytics for digital forensic. *J Vis* 18(2):173–184
- Liu H, Gao Y, Lu L, et al. (2011) Visual analysis of route diversity. In: *Proceedings of IEEE conference on visual analytics science and technology*. IEEE Computer Society Press, pp 171–180
- Liu D, Weng D, Li Y et al (2017) SmartAdP: visual analytics of large-scale taxi trajectories for selecting billboard locations. *IEEE Trans Visual Comput Graphics* 23(1):1–10
- Salton G, Yang CS (1973) On the specification of term values in automatic indexing. *J Doc* 29(4):351–372
- Salton G, Wong A, Yang CS (1975a) A vector space model for automatic indexing. *Commun ACM* 18(11):613–620
- Salton G, Yang CS, Yu CT (1975b) A theory of term importance in automatic text analysis. *J Am Soc Inf Sci* 26(1):33–44
- Shneiderman B (1996) The eyes have it: A task by data type taxonomy for information visualizations. In: *Proceedings of IEEE symposium on visual languages*, pp 336–343
- Sun G, Wu Y, Liang R et al (2013) A survey of visual analytics techniques and applications: state-of-the-art research and future challenges. *J Computer Sci Technol* 28(5):852–867
- Sun G, Liang R, Qu H et al (2017) Embedding spatio-temporal information into maps by route-zooming. *IEEE Trans Visual Comput Graphics* 23(5):1506–1519
- Landesberger T von, Bremm S, Andrienko N, et al. (2012) Visual analytics methods for categoric spatio-temporal data. In: *Proceedings of IEEE Conference on Visual Analytics Science and Technology*. IEEE Computer Society Press, pp 183–192
- Wang X, Grimson E (2008) Spatial latent dirichlet allocation. In: *Proceedings of neural information processing systems*, pp 1577–1584
- Weng D, Zhu H, Bao J, et al. (2018) HomeFinder revisited: finding ideal homes with reachability-centric multi-criteria decision making. To appear in *Proceedings of ACM CHI 2018*
- Zeng W, Fu C et al (2017) Visualizing the relationship between human mobility and points-of-interest. *IEEE Trans Intell Transp Syst* 18(8):2271–2284
- Zhao J, Forer P, Harvey AS (2008) Activities, ringmaps and geovisualization of large human movement fields. *Inf Vis* 7(3–4):198–209
- Zheng Y, Capra L, Wolfson O et al (2014) Urban computing: concepts, methodologies, and applications. *ACM Trans Intell Syst Technol* 38:1–55