Visualizing Hidden Themes of Taxi Movement with Semantic Transformation

Ding Chu, David A. Sheets, Ye Zhao, Yingyu Wu
Kent State University

Jing Yang
UNC Charlotte

Maogong Zheng, George Chen
Shenzhen Institute of Advanced Technologies

ABSTRACT
A new methodology is developed to discover and analyze the hidden knowledge of massive taxi trajectory data within a city. This approach creatively transforms the geographic coordinates (i.e. latitude and longitude) to street names reflecting contextual semantic information. Consequently, the movement of each taxi is studied as a document consisting of the taxi traversed street names, which enables semantic analysis of massive taxi data sets as document corpora. Hidden themes, namely taxi topics, are identified through textual topic modeling techniques. The taxi topics reflect urban mobility patterns and trends, which are displayed and analyzed through a visual analytics system. The system integrates interactive visualization tools, including taxi topic maps, topic routes, street clouds and parallel coordinates, to visualize the probability-based topical information. Urban planners, administration, travelers, and drivers can conduct their various knowledge discovery tasks with direct information. The taxi topics reflect urban mobility patterns and trends, which are displayed and analyzed through a visual analytics system. The system integrates interactive visualization tools, including taxi topic maps, topic routes, street clouds and parallel coordinates, to visualize the probability-based topical information. Urban planners, administration, travelers, and drivers can conduct their various knowledge discovery tasks with direct information.

Keywords: Taxi Trajectories, Semantic Transformation, Clustering, Topic Modeling, Latent Dirichlet Analysis

1 INTRODUCTION
In the metropolitan areas of big cities, such as New York, London and Beijing, a large amount of taxis move around streets transporting people among urban cores, business centers, tourist attractions, transportation hubs, and residential territories. A modern vehicle GPS device on a taxi can record its realtime moving path (i.e. trajectory) as a series of positions sampled with a small periodic interval. At each location, GPS information and meta information are stored such as time, geographical coordinates (latitude, longitude), speed, direction, and occupied/vacant status regarding whether the taxi has a customer. It remains a far-reaching goal to extensively decipher the information hidden in the complex, dynamic behavior of large populations. Moreover, domain experts and practitioners require effective visualization so that they can intuitively manipulate the iterative, exploratory process and derive pertinent insights.

In this paper, we propose a new visual analytics system that finds and manifests implicit patterns derived from the taxi movement data throughout a city. We creatively study each trajectory as a document consisting of the taxi-traversed street names. This is made possible by mapping GPS coordinates (i.e. latitude and longitude) from geometric positions to a meaningful text representation. This method is named semantic transformation since it can replace the spatio-temporal GPS sample with additional contextual information. Such transformation can add semantic knowledge from different aspects, such as street names, Points of Interest (POI) (e.g., restaurants and shops), and other functional information. The transformation can be flexibly manipulated such as applying it only on the segments when the taxi is occupied (i.e. with customers). In this paper we implement the transformation from taxi positions to street names. A street itself implies geographic and cultural information (i.e. semantics) to the citizen of a city. In the future we will study the transformation of more semantic information.

This approach enables analysis of massive taxi datasets as document corpora. Consequently, we are able to introduce the Latent Dirichlet Allocation (LDA) [4] to infer hidden patterns of moving taxi populations. We name such patterns taxi topics since LDA is widely used in topic modeling of document collections. Here a taxi topic is a cluster of streets within a probabilistic framework (see details in Sec. 4). This new scheme has several characteristics including:

- Discovering Group Movement Patterns: The topics reflect group movement patterns of taxis more than just the spatial division of road networks. For instance, one topic may consist of major streets of a city district and a highway connecting an airport to this district. The highway may concurrently be a part of another topic together with streets of another city district.
- Integrating Text Analysis Tools: The transformation facilitates the use of text mining techniques (in this paper we use LDA) on the raw GPS location data. It creates a new method for effective knowledge discovery of trajectories.
- Tolerating Error/Noise: A limited number of erroneous GPS mapping or missing GPS records are tolerated in topic detection based on the probabilistic topic models. It significantly reduces the need for data preprocessing and correction in traditional trajectory processing.
- Reducing Dimensions: The dimensionality of the trajectories, originally a large and varied number of GPS sampling locations, is reduced to the number of topics. This number is controlled and adjusted by users. A street is also represented by such a vector of probabilities over the topics. Therefore, investigative operations like clustering and outlier detection can be applied to the trajectories or streets via the probability vectors.

We design a prototype VATT (Visual Analytics of Taxi Topics) system aimed at supporting visual analytics tasks. It integrates taxi topic maps with several visualizations including street clouds, street/trajectory parallel coordinates, and topic routes. The system helps users effectively explore the discovered hidden patterns. We illustrate our scheme with several case studies on a large data set acquired from more than twenty thousand taxis in Shenzhen, China.

2 RELATED WORK
2.1 Trajectory Mining
Our method involves using semantics transformation and LDA for mining taxi mobility patterns. Trajectory data mining has been extensively studied [35]. A data-centric framework is presented in [18] to investigate the periodic behaviors along time intervals. Reference spots and Fourier transform are used to enable periodicity discovery [17]. Grouping moving objects according to spatio-temporal closeness defines flock/swarm/convoy patterns or path/sub-trajectory clusters [32].

Clustering and classifying long trajectories are mostly performed over distance measures using Euclidean or shortest-path over networks. For example, partitions of trajectories are used to cluster
them in rectangular regions, creating fine-grain groups [14]. TRACLUS [15] uses modified DBSCAN, a popular clustering method based on density reachability [8], to group trajectories by the density of their segments in the space. NEAT [22] studies trajectory clustering over road networks, where a similarly modified DBSCAN is used by further considering traffic flows together with segment densities and connectivity. The method finds clusters by looking for contiguous road segments with large traffic flows. Our taxi topics can also be considered clusters of streets but with distinct features described here. First, the topics are not strictly confined by the connections in the road network but rather created by the probabilities that sets of taxi trajectories follow similar streets.

Second, our method can naturally incorporate a variety of parameters over the transformed names into the clustering process. We explore this with approaches such as filtering repeated (or unwanted) street names or augmenting the trajectory documents based on taxi speed (i.e. speed compensation). The existing techniques mostly account for spatial-temporal values where such operations cannot be easily applied or need geometric computations. Finally, using LDA, our topics overlap with a probability distribution. A road can belong to multiple topics to better reflect taxi movement patterns within and across the topics. For instance, an airport highway is an important component of several topics (with different probabilities), each linking to one city district. This naturally reflects the taxi groups traveling within roads of one district and the airport. Using existing techniques, the cluster containing the highway connecting the airport might include zero, one, or several districts but does not directly exhibit the fact that the highway is shared by them.

Trajectories are used to extract geographical borders [24]. This method assumes non-overlapping regions which is expected for administrative/spatial borders. In contrast, our technique finds topics without the constraint of specific spatial areas, where the overlap is quantified by a probability defining the strength of multiple topics in a common area.

Semantic trajectories [22] has been studied where trajectories are enriched with semantic meanings. Preprocessing and manual (or semi-automatic) annotation are needed to add semantics to the whole trajectories or their segments. These methods do not directly annotate GPS positions since it “is not efficient as it may generate a large number of repetitive annotations” [22]. In contrast, our direct transformation from GPS samples to street names and the novel use of document mining techniques can overcome this problem and promote automatic processing.

Based on similar taxi dataset, Yuan et al. developed T-Drive [33], which makes recommendations of fastest paths for taxi drivers through a routing algorithm based on historical data. Gao et al. [10] presented a system that can help identify factors affecting taxi drivers income. Our approach based on semantic transformation and dimension reduction can be integrated with such existing methods for better suggestion and planning.

2.2 Trajectory and Taxi Visualization

Many approaches have presented visual tools for exploring geographical data. Andrienko et al. presented a review of the key issues and approaches [2, 3]. The techniques involve map-based displays and use information visualization techniques to visualize the spatial attributes of the data over temporal changes. Object trajectories are a key type of movement data where many techniques have been developed, such as GeoTime [13], TripVista [11], FromDaDy [12], vessel movement [30], etc. Wang et al. [28] studied point pattern analysis methods and built an interactive visualization system of hot spots based on a mashup technique. Cnrovvasan et al. [5] visualizes movement trace data based on the proximity to points of interest. The method limits its study on important locations and the specific movement patterns (e.g. concurrence, convergency, fluctuation) are computed on an abstraction space. These methods mostly visualize trajectories using density-based map and spatio-temporal aggregation to provide data overview. Spatial and temporal dimensions are discretized into regions/windows for levels of detail and clutter reduction techniques are employed.

Tominiski et al. tackled the problem of visualizing trajectory attributes together with the individual points comprising the trajectories [26]. Landesberger et al. interactively visualized trajectory data with categorical changes over time and spatial data displays [27].

The visual study of taxi data has been conducted in a variety of applications. Liu et al. presented a visual analytics system of route diversity [19]. They developed several visual encoding schemes to display the statistical information for different routes and their importance with a ring map. In particular, entropy is computed for a source/destination pair of locations, and visualized to reveal the diversity of routes. Pu et al. [23] developed a system for users to monitor and analyze complex traffic situations in big cities for regions, roads and from vehicle views. Historical statistical information (named fingerprints) are encoded and placed on the map as ring-map-based radial layouts. Wang et al. presented an interactive system for visual analysis of urban traffic congestion [29] using traffic jam propagation graphs that join spatially and temporally related jam events. These graphs are explored by multiple views for the whole city, together with details of road segments. Ferreira et al. allowed users to visually query taxi trips [9]. The model supports origin-destination queries, different aggregations and visual representations for exploration and comparison of query results. An adaptive level-of-detail rendering strategy can generate clutter-free visualization for large results. Hidden details are shown as summarized overlay heat maps.

The VATT system aims at supporting visual analysis based on the discovered taxi topics with several major differences from previous methods. First, we need to visualize the discovered patterns of the group behavior, which is implemented as multiple colored layers of clusters over maps. Street line transparencies are used to show the associated probabilities for overlapped topics over streets. Second, our method introduces semantics to spatial-temporal data so we use tag cloud to help users directly find the information. Third, the LDA-computed probability and entropy are used in visualizing diversity of streets and taxi trajectories. Parallel coordinates and clouds are used on the reduced topical dimensions. The VATT visualizations need to be further refined. We will provide more visualization functions inspired by the related work, such as better clutter reduction and levels of detail, temporal semantic information with ring maps, and interactive query support.

2.3 LDA and GPS Data

LDA [4] has become a widely-used machine learning technology in organizing and finding patterns in different textual corpora. The complexity of detecting and processing GPS data into a usable form has been widely studied [20]. To the best of our knowledge, our method is the first to use LDA in GPS sampling data processing.

3 Taxi Data Transformation

Taxi Trajectory Data The trajectory data used in our case studies contains daily trajectories of 21,360 taxis in Shenzhen, a big city in southern China bordering with Hong Kong. Shenzhen has over fifteen million residents in a condensed area and taxis are a major means of passenger transportation. Each taxi reports nearly three thousand GPS sample positions per day. Each sample consists of taxi plate, time, status, speed, direction, and latitude and longitude, e.g., (BXXXXX, 06/27/2012 23:59:33, 0, 50, 180, 22.533, 114.044). The accumulated taxi trajectories create a very large data set for investigation with a total of 59,087,230 samples recorded in one day.
4 STUDYING TAXI DATA WITH TOPIC MODELING

4.1 Topic Discovery with LDA

LDA (Latent Dirichlet Allocation) is an effective (semi-)automated probabilistic topic modeling tool that can discover hidden thematic structure in a large corpus of documents [4]. LDA statistically groups keywords into potential topics by studying their occurrences among a large collection of documents. Consequently, each keyword is assigned a probability of its appearance in each topic. Visualizing the topics has been investigated to present the themes inside news, tweets, and other textual documents [6, 1]. To the best of our knowledge, we are the first to apply this powerful tool to geographic trajectory data, enabled by the semantic transformation. In our study we implement the LDA computation using the Stanford Topic Modeling Toolbox. Table 2 and Table 3 illustrate the LDA results of eight taxi topics, which are generated from 21,360 taxi trajectories between 9am to 12pm, Jun. 27, 2012.

**Table 1: Examples of mapping GPS to streets.**

<table>
<thead>
<tr>
<th>Latitude</th>
<th>Longitude</th>
<th>Street Name</th>
<th>Street ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.543</td>
<td>113.991</td>
<td>Quancheng East Rd</td>
<td>38819</td>
</tr>
<tr>
<td>22.543</td>
<td>113.997</td>
<td>Quancheng West Rd</td>
<td>38849</td>
</tr>
<tr>
<td>22.568</td>
<td>114.066</td>
<td>Beihuan Ave</td>
<td>10182</td>
</tr>
<tr>
<td>22.568</td>
<td>114.067</td>
<td>Beihuan Ave</td>
<td>10208</td>
</tr>
</tbody>
</table>

**Table 2: LDA results: Total frequency of all taxis' appearance on a topic and the corresponding probabilities, p(w|z), of this street's appearance in multiple topics.**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Topic ID</th>
<th>Beihuan Ave</th>
<th>Binhe Ave</th>
<th>Fuqiang Rd</th>
<th>Fuqiang Rd2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic1</td>
<td>758188</td>
<td>35247</td>
<td>0.39</td>
<td>37432</td>
<td>0.29</td>
</tr>
<tr>
<td>Topic2</td>
<td>717219</td>
<td>54810</td>
<td>0.50</td>
<td>52</td>
<td>0.00</td>
</tr>
<tr>
<td>Topic3</td>
<td>408980</td>
<td>111172</td>
<td>0.18</td>
<td>618</td>
<td>0.02</td>
</tr>
<tr>
<td>Topic4</td>
<td>298811</td>
<td>0</td>
<td>0.00</td>
<td>5</td>
<td>0.00</td>
</tr>
<tr>
<td>Topic5</td>
<td>1441009</td>
<td>81191</td>
<td>0.38</td>
<td>165</td>
<td>0.00</td>
</tr>
<tr>
<td>Topic6</td>
<td>218869</td>
<td>12</td>
<td>0.00</td>
<td>13</td>
<td>0.00</td>
</tr>
<tr>
<td>Topic7</td>
<td>1777306</td>
<td>0</td>
<td>0.00</td>
<td>49</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**Table 3: LDA results: Probability contribution, p(z|d), of taxi topics to taxi trajectories.**

<table>
<thead>
<tr>
<th>Trajectory</th>
<th>Topic1</th>
<th>Topic2</th>
<th>Topic3</th>
<th>Topic4</th>
<th>Topic5</th>
<th>Topic6</th>
<th>Topic7</th>
<th>Topic8</th>
</tr>
</thead>
<tbody>
<tr>
<td>BXXXX1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.23</td>
<td>0.69</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>BXXXX2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>BXXXX3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.15</td>
<td>0.29</td>
<td>0.20</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>BXXXX4</td>
<td>0.31</td>
<td>0.31</td>
<td>0.00</td>
<td>0.38</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**4.2 Entropy Information**

Entropy can quantify the uncertainty of a street or trajectory over the topics to measure the variation or diversity of the variable:

\[
H(w) = -\sum_z p(z|w) \log p(z|w).
\]

A street with high entropy is shared by multiple topics. Very likely, it is a backbone street connecting different parts of the city. For example, in Table 2, Binhe Ave. has a higher entropy than Fumin Rd. From Shenzhen map, Binhe Ave. is a major street passing city hubs, including the Convention Center and the Huanggang port to Hong Kong.

**Trajectory Entropy:**

\[
H(d) = -\sum_z p(z|d) \log p(z|d).
\]

A specific taxi’s trajectory has high entropy if it receives significant contributions from multiple topics. For example, this taxi may travel through a long path in the city across various districts (Fig. 8).

In summary the achieved probabilities and entropies are used to visualize the knowledge hidden in massive trajectories.

**4.3 Topic Modeling with Trajectory Data Adjustments**

The massive trajectory data include more information than the sequence of geographic coordinates, such as the vehicle’s speed and occupancy status. Such information can be integrated into the semantic transformation process to expose more knowledge in the downstream topic modeling results. In Section 3, we have discussed that a long street can be divided into several segments to increase geographic granularity. Next, we show other approaches to preprocessing the data before the transformation and topic modeling.

**Taxi Occupancy**

Applying the semantic transformation to original trajectories discovers topics without considering taxi occupancy. In addition, we can apply the transformation only on those trajectory sections a taxi is occupied by passengers. The resulting topics reflect moving behaviors of taxi riders. For example, they imply the mobility between a city’s business and residential divisions. Such knowledge can potentially contribute to new public transit routes or traffic rules over specific streets. Moreover, the topics of uncoupled taxi trajectories manifest moving patterns of drivers seeking passengers, e.g., patterns related to busy transportation hubs. The administration can arrange bus/subway stations accordingly.

**Speed Compensation**

The GPS positions that are acquired with regular time intervals create bias toward slower moving roads with
more samples. This results in high importance computed for such roads in the topics. This is useful in determining traffic jams and bottlenecks. On the other hand, users may want to reveal and analyze the mobility patterns without such bias. The bias can be removed by injecting pseudo-sampling points based on the taxi speed. In implementation, we set a speed threshold \( S \) and a speed interval \( \Delta S \). If a taxi’s speed \( S \) on a street exceeds \( S \), we add a series of pseudo points, the number of which is \( \frac{(S - S) \Delta S}{\Delta S} \).

### 4.4 Taxi Topics Evolution

Taxi topics are created for a given time window. At different periods of a day they reflect different patterns. For example, people travel from suburbs to city centers in the morning and reversely in the afternoon. It is important and interesting to find the evolutionary trends of the topics along a time line. Thus, we define a topic similarity between two topics to find similar topics between consecutive time windows. Given two topics \( i \) and \( j \), the similarity is computed as:

\[
S_{ij} = \frac{\text{Size}(T_i \cap T_j)}{\text{Min}(\text{Size}(T_i), \text{Size}(T_j))},
\]

where \( T_i \) and \( T_j \) are the sets of streets with a high probabilities \( p(w|z) > c \) in each topic. Based on the similarities, closely related topics are identified in consecutive time windows, which reflect temporal evolution of topics. Such relationships are visualized to identify the variation of topical contents, the emergence of new topics, and their fading out.

### 5 VATT System

Our VATT system integrates multiple visualizations supporting visual analytics of the hidden knowledge. Figure 1 shows the system interface consisting of several components: (a) taxi topic maps; (b) street clouds; (c) parallel coordinate plot (PCP) of streets or trajectories over topics; (d) topic routes of temporal evolution. These visualizations are orchestrated in a coordinated manner for effective user exploration.

#### 5.1 Taxi Topic Maps

Taxi topics, consisting of multiple streets, are visualized over geographic maps. Various backgrounds, such as topographical, satellite, and transportation maps, can be used to provide visual cues and context of geographic and cultural information. Features such as tourist attractions, subway stations, and government buildings can be placed on the map. The system supports direct use of Google Maps, OpenStreetMaps (openstreetmap.org), and user customized maps. Overlayed with the background, multiple layers are used to present different topics. To visualize one topic, streets belonging to it (defined as \( p(w|z) > c \) are drawn with a given topic color. The line width of a street is determined by the street importance in a particular topic. Here a street may be drawn by multiple overlapped lines with different topic colors. To improve visual effects, the transparency of a line can be set (1) by the street importance, so that a street is shown obviously with the color of the topic that has the largest \( p(w|z) \); or (2) by the entropy \( H(w) \) of the street to the topics such that multiple overlapped lines of this street representing each topic are better depicted, especially with a zoomed in view. Figure 1(a) displays eight LDA-generated topics with distinct colors on a terrain map of Shenzhen. Users can manipulate the appearance of the topics by setting layers visible/invisible and changing the color and transparency.

#### 5.2 Street Cloud

We use a street cloud to display semantic information and the relationships between the streets. The size of a street name reflects the street importance, i.e., the frequency of its occurrence in one topic. Moreover, the probability distribution over multiple topics, \( p(w|z) \), of a street is drawn as a graph line and overlaid on the lower part of the street name. This graph line allows users to identify whether the street is solely included in one topic or diversely shared by others. More importantly, the street names are not randomly positioned. Instead, the layout reflects their relationships by introducing attracting and repelling forces computed from pairwise cosine similarities:

\[
sim(w_i, w_j) = \frac{\sum_{z=1}^{Z} p(w_i|z)p(w_j|z)}{\sqrt{\sum_{z=1}^{Z} p(w_i|z)^2\sum_{z=1}^{Z} p(w_j|z)^2}},
\]

where \( w_i, w_j \) are two streets and \( Z \) is the number of topics. Streets with similar topic distributions are positioned near each other, and dissimilar streets are positioned further apart.

In the street cloud of a given topic, some streets are aggregated together if they are visited by taxis traveling much inside this topic. Meanwhile, different groups are placed far away from each other. Street cloud can also show the most important streets among all topics, or those selected by users. Figure 1(b) shows the top thirty streets with the largest frequencies in all topics. The street names are thus grouped according to their distribution over different topics.

#### 5.3 Parallel Coordinate Plots of Streets and Trajectories

The PCP visualization helps users to interactively discover knowledge from probability distributions over topics, which was used in exploring document topics [7]. Figure 1(c) visualizes streets over a PCP, where each dimensional coordinate is one topic and each polyline represents one street. To draw a street over multiple dimensions, its probabilities over the topics \( p(w|z) \) are used. A logarithmic scale is applied since many probabilities are very small and only a few of them are large. To overcome clutter, streets are not drawn if they have smaller probability values than a threshold \( P_i \) over \( K \) topics. \( P_i \) and \( K \) are user controllable parameters. Clicking a topic’s axis highlights all the streets which have high probabilities in that topic. In a similar manner, the PCP visualization can also be applied to taxi trajectories with each polyline representing a trajectory (here the probabilities \( p(z|d) \) are used). In general, PCP helps users identify distributive patterns of streets or taxis, as well as study outliers and salient features.

#### 5.4 Topic Routes of Temporal Evolution

Figure 1(d) demonstrates the routes of temporal evolution among multiple topics. Each column represents a time window and adjacent columns are consecutive time windows. Each circle stands for one topic at a time slot, whose radius is defined by the topic’s importance. For each column, we always show the topics with higher importance on top of lower ones. Two circles in adjacent columns are connected by a link if their topic similarity exceeds a given threshold. The width of the link is related to the similarity, whose value is also shown. Similar topics are given the same color. Figure 1(d) reflects the evolution of topics in a whole day. This visualization also plays a role as a control panel, where users can select time windows and topics so that they are shown in other coordinated views for further investigation.

### 6 Case Studies: Visual Analytics of Taxi Data

We use the VATT system to explore the massive taxi trajectories of one day (June 27, 2012) in Shenzhen city. The number of topics and time windows are easily controlled by users. We have tried different numbers of topics. The selection should depend on specific knowledge of the city. Meanwhile, the number of time windows can be adjusted by users for interactive analysis. In the following examples, taxi topics are generated at each of the eight time windows of the whole day, i.e., 0am-3am, ···, 9pm-12am, etc.
6.1 Taxi Topics and City Mobility Patterns

The topics reveal typical traveling patterns of city cabs. Fig. 2 displays the eight taxi topics created for the time window 9am to 12am. Here the speed compensation is applied and the topics are created from full taxi trajectories including occupied and vacant sections. The streets of each topic with a high probability are drawn on the map in a distinct topic color. Street clouds of four topics are also shown in the figure. They display the top 30 streets of each topic in the corresponding topic color. The topics are ranked and labeled by their importance from Topic1 (high) to Topic8 (low). An immediate finding is that the topics are approximately distributed following the city’s administrative divisions. The visualization implies that the districts of Shenzhen are functionally set up: a large portion of taxis can accomplish their movement inside a district. Next we introduce more findings with examples from Fig. 2.

Topic1 in purple roughly overlaps over Futian District, which contains the business and administration center of Shenzhen. Topic1 has the largest importance which indicates this district is the most concentrated area of taxis from 9am to 12am. Observing the street cloud of Topic1, the important streets can be easily recognized according to the font size, such as the Xinzhou Rd., Fuqiang Rd. and Binhe Ave. The first two are in a large group of streets in the middle of the cloud. Meanwhile, Binhe Ave. is pushed away from this group by the similarity forces. Users can zoom in on the map to study Binhe Ave. (Fulu), which is a side road out of Binhe Ave. From its location in the street cloud, in conjunction with the map, we can derive that it is used by taxis connecting the local roads inside this district as well as by taxis entering Binhe Ave. aimed to other districts. Such revelation on road connectivity might indicate potential traffic improvements to the administration.
They are two major expressways of Shenzhen city, including the aforementioned Beihuan, Binhe, and Guangshen roads. They are easily accessed and have no major expressways of the city, including the aforementioned Beihuan, Binhe, and Guangshen roads. They are two major expressways of Shenzhen city, including the aforementioned Beihuan, Binhe, and Guangshen roads. They are easily accessed and have no major expressways of the city, including the aforementioned Beihuan, Binhe, and Guangshen roads. They are easily accessed and have no major expressways of the city, including the aforementioned Beihuan, Binhe, and Guangshen roads.

Fig. 4 shows a few examples of these results. For the time occupied taxi trajectories, and another eight using vacant trajectories. We further create eight taxi topics using the data of only passenger-occupied taxis. Obviously, most of those taxis seeking passengers at the very early morning. They cant and occupied topics, respectively. Fig. 4a reveals the activities over a slot from 3am to 6am, Fig. 4a and Fig. 4b plot the streets over various arrows. Users can zoom in to study the active streets. Such knowledge around two major ports of the Hong Kong border indicated by travel around the city's main districts while focusing on important centers, Buji (Topic8 in orange) and Longgang (Topic7 in blue). In particular, the variation of city traffic patterns at different time slots plays a significant role in finding important information and making smart decisions. For example, city administration can adjust the rule of ways (e.g. one-way) according to the time of day, to help avoid traffic jams at this time. In comparison, we draw the occupied topic routes of occupied taxis during the whole day.

As shown on its street cloud, the related streets constitute a ring shape. This means that these streets are not grouped into salient centers, another critical business area. On the street cloud, the main group of Luohu's internal streets. Binhe Ave. connects to different divisions. The long green path clearly shows the high-activity area may be used by police to design optimal patrol routes. Fig. 5: Temporal change over one topic. The taxi activities are very low during midnight and early morning and many tourists use this airport to visit Hong Kong. Domestic flights flying from Shenzhen airport within China are less servicing people from the border port to the airport, perhaps because sellers from Hong Kong use Shenzhen airport to fly around China; way (Guangshen Expressway) connecting the airport to Huanggang Road.
Individual topics not linked to others reflect important patterns shown in the street cloud within rectangular boxes. Furthermore, the two major expressways connecting two northeastern residential centers, Buji and Longgang, respectively. They appear at the afternoon time window from 6pm to 9pm, and at the morning time window from 9am to 12pm. Clear differences can be observed at G205 and S360, the Hong Kong border (similar to Topic 6 in Fig. 2 created from original data). In other time slots, these roads are accommodated in the map view in Fig. 4c, the bright green topic of occupied taxis.
We innovate semantic transformation to enable the use of text analysis techniques on massive GPS sampling data. The LDA model is applied to generate taxi topics, which reflect inherent patterns of taxi mobility. We have developed a visual analytics system to explore massive taxi trajectories data based on the discovered topics. The visualizations support various visual analytics tasks.

In the future, we will investigate advanced LDA models (e.g. temporal LDA) for taxi data. We will study semi-automatic parameter definitions for optimal topic number. We will also integrate more visualization techniques into our system.

Acknowledgements

This work is partially supported by NSF IIS-1352927 and IIS-1352893.

References


