Interactive Visual Discovering of Movement Patterns from Sparsely Sampled Geo-tagged Social Media Data

Siming Chen, Xiaoru Yuan, *Senior Member, IEEE*, Zhenhuang Wang, Cong Guo, Jie Liang, Zuchao Wang, Xiaolong (Luke) Zhang, *Member, IEEE* and Jiawan Zhang, *Member, IEEE*

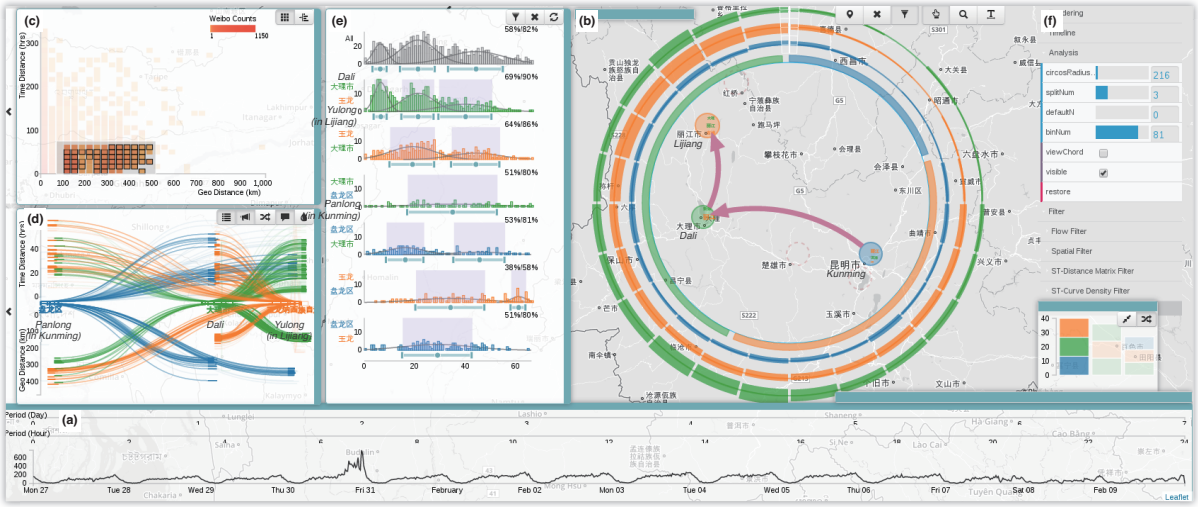


Fig. 1. System Interface: (a) Time line view, for selecting time range of the social media data. (b) Spatial temporal view, for aggregating movements based on spatial, temporal domain. (c) ST Matrix View, showing the distribution of movement in geo distance and time interval attributes. (d) ST Detail View, for analyzing pair-wised detail movement. (e) ST Model View, showing the output of uncertainty model and guiding users to filter reliable data for each derived movement category. (f) Parameter Control Bar.

Abstract— Social media data with geotags can be used to track people's movements in their daily lives. By providing both rich text and movement information, visual analysis on social media data can be both interesting and challenging. In contrast to traditional movement data, the sparseness and irregularity of social media data increase the difficulty of extracting movement patterns. To facilitate the understanding of people's movements, we present an interactive visual analytics system to support the exploration of sparsely sampled trajectory data from social media. We propose a heuristic model to reduce the uncertainty caused by the nature of social media data. In the proposed system, users can filter and select reliable data from each derived movement category, based on the guidance of uncertainty model and interactive selection tools. By iteratively analyzing filtered movements, users can explore the semantics of movements, including the transportation methods, frequent visiting sequences and keyword descriptions. We provide two cases to demonstrate how our system can help users to explore the movement patterns.

Index Terms—Spatial temporal visual analytics, Geo-tagged social media, Sparsely sampling, Uncertainty, Movement

1 Introduction

Social media data often contain rich information about people's lives. For example, microblogs not only allow people to share their thoughts and activities, but also offer such information as when and where a user posts the microblog. The availability of spatial and temporal information in social media can help us better understand people's activities [13, 21]. For example, if we know the locations and time of a user's microblogs, we can infer not only where the user has been, but also how the user moved from one place to another. Combined with the contents, we can learn what people do on a particular route or how a user's thoughts and actions have evolved through space and time.

While visual analytics of microblog contents has been studied recently [31, 46], it is still a challenge to conduct in-depth analysis of the movement patterns of microblog users with the spatiotemporal information extracted from microblogs. For example, while geotags embedded in microblogs enable us to easily learn where people have been, it is difficult to identify movement patterns, such as how people moved and how long some related trips may take. These movement patterns can help us in situations like planning trips, or understanding people's travels within certain regions.

The difficulty in discovering movement patterns is largely caused by the variability and uncertainty of the time when users post microblogs. Some users may post microblogs immediately before or after a trip, but some may delay posting for various reasons, and the de-

- *Siming Chen, Xiaoru Yuan, Zhenhuang Wang, Cong Guo, Jie Liang, Zuchao Wang are with Key Laboratory of Machine Perception (Ministry of Education), School of EECS, Peking University. E-mail: {siming.chen, xiaoru.yuan, zhenhuang.wang, cong.guo, jie.liang, zuchao.wang}@pku.edu.cn.*
- *Xiaolong (Luke) Zhang is with College of Information Sciences and Technology, Pennsylvania State University. E-mail: lzhang@ist.psu.edu.*
- *Jiawan Zhang is with School of Computer Science and Technology, and School of Computer Software, Tianjin University. E-mail: jwzhang@tju.edu.cn.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

lay could be random. Such randomness in the the posting time makes it hard to estimate travel time.

From the perspective of understanding movement patterns, the temporal data embedded in social media can be regarded as episodic data, as defined by Andrienko et al. [6], because the data are not acquired with a predictable sampling rate. Consequently, existing trajectory analytical tools, which are largely designed for those data sets that are sampled in a predicable manner and density (such as GPS data), cannot be used to analyze social media data to construct the movements of users.

Although the temporal data from an individual user might be random and episodic, data from many people can be aggregated and processed to help the analysis of movement patterns of social media users. The research reported in this paper explores an approach to analyze movement patterns of social media users. Based on the features of social media data, we first propose an analytical pipeline to guide the process and analysis of social media data. To address the uncertainty in extracted traveling time based on social media, we develop an uncertainty model to help understand of the distribution of relevant temporal data. Based on the pipeline and the uncertainty model, we design a visual analytics system that allows analysts to use the contents of microblogs, the locations, and posting time for in-depth analysis of movements of social media users.

Our main contributions can be summarized as follows:

- **Developing an interactive visual analytics procedure to analyze movement patterns of social media users** The novel visual analytics procedure is based on an uncertainty model that we have developed to estimate the temporal attributes of social media data. Through this procedure, users can extract more reliable information about traveling time from data and build reasonable movement categories to further explore movement patterns.
- **Developing a visual analytics system to analyze the semantics of the movement patterns of social media users** This method combines the spatial, temporal and textual information extracted from raw social media data. With this method, users can interactively explore, discover and analyze the categorized groups, transportation methods or purpose and frequent visiting patterns.

This paper is structured as follows. Section 2 reviews related work. We describe the data and design considerations in Section 3 and 4. After introducing the uncertainty model in Section 5, we present the visual analytics procedure and technical details in Section 6 and 7. We demonstrate the use of our tools in two case studies in Section 8. Finally, we discuss the limitations and future work.

2 Related Work

The related work is classified into four categories : movement visual analytics, analysis of geo-tagged social media, visual analytics of sparsely sampled trajectory data and movement semantic analysis.

2.1 Movement Visual Analytics

Andrienko et al. [3] summarized the analytics tasks in this field. There are two types of data, dense sampling trajectories and Origin-Destination (OD) data are two types of movement data. Trajectory data is usually multivariate, and can be visualized with various methods, such as 3D space-time cubes [20], stacking-based visualization with informative glyphs [35], and growth ring maps [8]. To help users better interact with and manipulate views and data in analysis, researchers have proposed tools to reduce trajectory clutter [5], as well as to filter [23], cluster [2] and aggregate spatiotemporal data [4].

OD analysis is often supported with flow map [29], OD matrix [40], and OD map [41]. A flow map uses a real map to show OD flows as a graph. It clearly represents the spatial features of OD flows, but large OD sets can easily make the view cluttered. Solutions to this problem include regionalization [16] and edge bundling [29] to reduce the clutter. OD matrix visualization separates OD pairs from a map, and uses a matrix to show the patterns of OD flows. While the matrix avoids the problem of cluttered views, the lack of spatial information often requires the presence of a map alongside the matrix [5]. OD map [41] improves the OD matrix method by overlaying a matrix on a map so that users can still see the OD flow patterns through a matrix without referring to a separate map [32].

However, these methods are not competent when applied in the social media data. Trajectory-based tools often require reliable and steady temporal information, which social media data lack. OD analysis tools demand clear distinctions of origins and destinations, which are difficult to be identifed from social media data.

2.2 Analysis of Geo-tagged Social Media Data

Massive research has been done to exploit geotags of social media data to infer and predict people's spatial and temporal behaviors. For example, geotags are used to improve situation awareness in emergency response [37, 43] and disease control [19], and to understand the dynamics of neighborhoods [12] and cities [36, 42]. Cao et al. [11] investigated how information flow among multiple places. Our review focuses on research using geotags to understand movement patterns.

Geo tags have been used to analyze the routes or trajectories. Kurashima et al. [24] proposed a travel route recommendation system based on geo-tagged photo sharing data. Azmandian et al. [7] used Twitter geotags to link people's movement patterns to transportation networks in cities and to classify people's weekly activities. Li et al. [26] showed that combined with computational tools, geotags can successfully reconstruct a complete road map of a city. Based on the content and geo-tagged information, semantics of movement can be explored from the social media data. Gabrielli et al. [14] used Twitter geotags to detect what types of locations have been involved in people's movements. Krueger et al. [21] used GPS and location-based service data to support the analysis of movement behaviors.

These methods, however, usually involve no or little temporal information. The tasks they support largely focus on the understanding of spatial features of movement. Fuchs et al. [13] incorporate temporal information significantly in analyzing movement patterns. Their research shows that statistical data of temporal information can improve the analysis results.

2.3 Visual Analytics of Sparsely Sampled Trajectories

Analyzing sparsely sampled trajectory data is a newer topic. Traditional methods for trajectory analysis and OD analysis cannot be directly applied. Andrienko et al. [6] named such data as episodic movement data, and classify the data into three categories: location based data (by static sensors), activity based data (e.g. Twitter, Mobile phone), and device based data. They argue that aggregation is an approach to reduce the uncertainties. This approach has been used to compare the difference in spatial temporal behaviors of photo-taking user groups [34], and to study user movements from different types of sensors, such as Bluetooth [27] and RFID [39].

This approach can be applied in social media data analysis. Users could be regarded as sensors that feed data, but users, as automous agents who can move around, are more complex than physical sensors. Thus research is needed to study how to aggregate spatial temporal data from social media users.

2.4 Movement Semantic Analysis

We assume movements with semantics are more reliable. Therefore we try to extract data with certain meanings and use that for deeper analysis. The semantics can be derived from *stops* at certain locations, and *moves* between consecutive stops [28]. For each stop, the semantics mainly concerns the function of the location and the activities that the person performs at that location. This can be derived from nearby keywords [22] or Points of Interest [21]. In this work we use the keywords in microblogs.

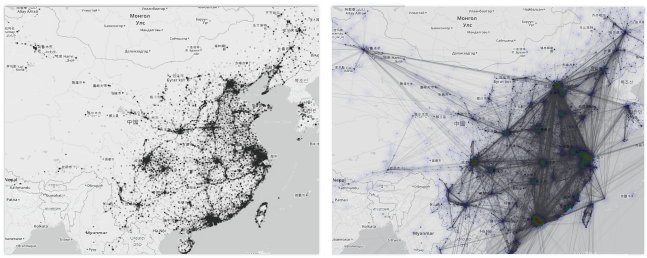
For each move, the semantics mainly concerns the transportation modes: walking, cycling, cars, trains, flights, etc. Most works on transportation mode detection are based on supervised learning. They derive rich features for each sampling point or short trip, and train classification models, including Logistic Regression [33], Decision Trees [47, 30], Support Vector Machines [47] and Hidden Markov

Model [30], etc. To generate such rich features, they require dense movement data, such as GPS [47] and GSM [33] data. They also require considerable amount of labelled data for classifier training. In contrast, our microblog data is sampled very sparsely. It is also expensive to gather the training data with labelled transportation modes between two consecutive microblog posts. Therefore we adopt an unsupervised learning method based on travel time clustering, where each cluster of travel time corresponds to one transportation mode. K-means clustering has been tested on call record data, attributing a transportation mode to every movement [38]. However, we find that in microblog data, many movements cannot be reliably attributed to known transportation modes. Thus, we use a fuzzy clustering method, Guassian Mixture Model, allowing some movements to be considered as noise.

In summary, while massive research has been carried out for human movements, visual analysis for geo-tagged social media is challenging due to the uncertainty and irregularity of spatial temporal data embedded in social media data. To fully leverage the richness of social media data in understanding people's behaviors, we need to explore methods for reliable extraction of relevant movement parameters.

3 Data Description and Statistics

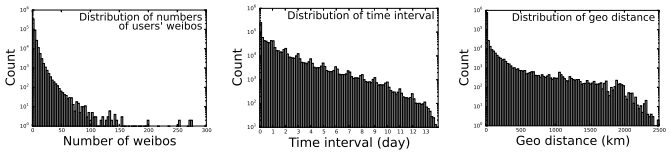
The data involved in this research were extracted from Sina Weibo, the most influential microblogging service in China. The primary services provided by Sina Weibo are very similar to those by Twitter. Each weibo is a short blog, like Tweet in Twitter.For each useful weibo, we extracted its content, geotag, time stamp, and user profile information.



(a) (b)

Fig. 2. Geo-tagged weibo data in two weeks. (a) Each weibo as a small dot on the map. (b) Movements with a heat map on the background.

We first conducted several preliminary analyses on raw weibo data. The data we chose were dated between January 27 and Februray 9, 2014, and include approximately 13.7 million weibos from 4.9 million distinct user accounts. For each individual user, we plotted each weibo on a map as a dot based on the location extracted from the weibo's geotag. All dots associated with a user are connected based on the temporal sequence to weibos (Fig 2). The time interval and geo distance are then derived for each consecutive movement pairs of the sequence. The trajectory is sparsely sampled, which indicates that users might travel to other places between two recorded locations.



(a) (b) (c)

Fig. 3. Data statistics, including the distribution of the numbers of users' weibos (a), time interval (b) and geo distance of the movement (c).

We analyzed the attributes of the movements (Fig 3). In Fig 3a, we can see that the majority of users posted less than 150 posts during that period of time, while a few accounts had significantly high frequency of activities (more than 1000 weibos in two weeks). After checking these accounts, we found that most of them are robot accounts, such as advertisement accounts, weather forecasting accounts, or traffic status broadcasting accounts. These types of accounts should be excluded in analysis, because of our interest in human movements.

We also analyzed the time interval distribution (Fig 3b). Generally, the smaller the time interval value is, the more movements we obtain. Also we observed an interesting phenomenon: there is a small peak at the end of each one-day period. By exploring the raw data, we found that some people tended to record their lives in a regular way (e.g. every morning posting a "morning", or checking in a regularly visiting cafe). When people travel, they might post a weibo when arriving in a new place and staying there for one day or more. As for the distribution of the users by geo distance in Fig 3c, there are still a fair amount of long distance movements ($>$ 100 km) worth attention.

4 Design Consideration

We have identified some unique features of weibo data in order to design visual analytics tools on weibo user movement. These features are as follows:

- **Large amounts of people and wide geographic coverage**. The number of Weibo users is massive. They can access social media services and post weibos wherever they want. However, it is challenging to analyze the movement patterns of different regions within different time periods. Thus, it is desirable to have flexible interaction methods to support dynamical exploration of weibos' spatial temporal characteristics **(R1)**.
- **Irregular sampling of trajectories**. We have observed that, between two places, the movement time interval of different people varied significantly, making a complicated distribution of time intervals. The variation may be caused by different reasons, including different transportation methods (such as by air or train), different habits of weibo users (for example, delayed vs. prompt posting), and whether visiting other places. It is important to understand the difference in such time interval distribution in order to understand the movement behaviors. Therefore, users should have tools to examine the distribution of time intervals **(R2)**.
- **Uncertain data**. We have observed unreasonable movements, such as a movement time interval of seven days for a short distance travel, or a movement of 100 km in five minutes. Filter functions are needed in order to help people select useful data and exclude outliers. Furthermore, only parts of the geotagged samples can be regarded as meaningful movements, such as those without unreasonable delay. Therefore, we need new algorithms or models to help users conduct data filtering and then construct more reliable movement categories based on the contents **(R3)**.
- **Movement with semantic information**. The activity that people post in geo-tagged social media data can provide semantic meanings. The movement time interval, geo-distance and visiting sequences, as well as textual information, could involve the motivation or characteristics of the movement. Different features can be categorized by different travel purpose or methods. Thus, it would be beneficial to have tools that help users explore such categorical semantic meanings **(R4)**.

Hence, our goal is to develop a system to help users obtain movement patterns and semantics from the following perspectives.

- **Categorized movement and possible transportation Methods**. Although not all people directly mentioned what forms of transportation they take, we could infer the information, with a higher probability belonging to some traffic methods.
- **Frequent visiting patterns**. With reliable categorized movement, the weibo data reflect social behaviors such as travel sequence.
- **Keywords and content of the trips**. Combined with the keywords view of each place and movement, we can generate hypotheses on the trip goal or interesting behaviors.

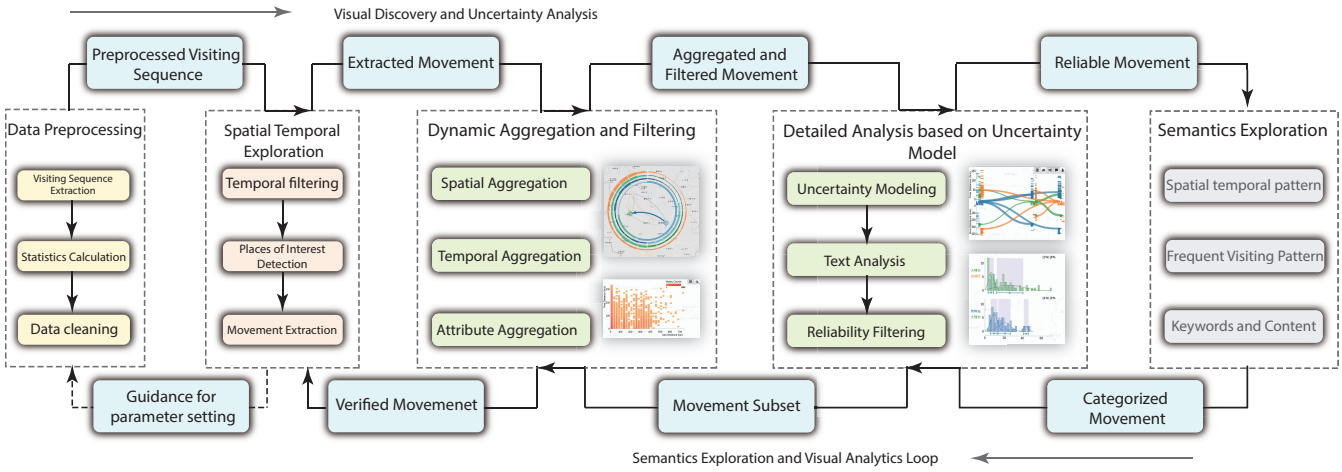


Fig. 4. System pipeline. Our system supports the visual discovery process based on uncertainty model, by interactively filtering and drilling down to the details. Throughout the process, semantics of the movements would be derived from the categorized reliable movement.

To achieve the goal, we have proposed a visual analytical pipeline that integrates the automatic algorithms and interactive filtering (Fig 4), with the following components.

- **Data preprocessing**. In this step, visiting sequences of each user from the original input data can be extracted. Based on statistical data, users can set the parameters to clean the data in this step.
- **Spatial temporal visualization**. In this step, users can apply various spatial temporal visualization tools to interactively explore the data **(R1)**. For example, they can filter the places of interest based on density maps, and extract movements among these places for further analysis.
- **Dynamic filtering and aggregation**. To support the analysis of the movement patterns from large amounts of people, aggregation would be performed in different aspects, including locations, periodicity, and attributes. Users can apply multiple filters across these aspects. **(R2)**.
- **Detailed analysis based on uncertainty model**. With the filtered movement, users can rely on the uncertainty model to identify multiple movement categories with semantics meanings. The model provides a guidance on the confidence interval of travel distance and time and lets users choose more reliable movements for each category **(R3)**. With text analysis, users can also interactively adjust the confidence interval of each category that better fits into a real world situation. The outputs are the categorized and reliable movements for semantics exploration.
- **Semantics exploration and iterative visual analytics**. Based on the filtered movements, users can look for spatial temporal patterns and frequent visiting patterns. By analyzing the patterns in linked views, users can iteratively verify the results and extract semantics of the movements **(R4)**.

5 Uncertainty Model

Sparsely sampled geo-tagged social media data can produce high variability and uncertainty, which may lead to misleading results. We propose a model to guide users to filter reliable movement for subsequent analysis. We are aware that movement with certain semantics, especially with clear transportation modes, are more likely to produce reliable results. We classify the movements according to different transportation modes.

5.1 Gaussian Mixture Model (GMM) based Uncertainty Calculation

The geo-tagged social media data cover many transportation modes (Fig 5). Traveling by different modes takes different lengths of time. Therefore, we used a time interval between two consecutive posts (roughly the travel time) to classify the movements. We first removed the movements with a time interval that was outside reasonable bounds. This corresponded to a time interval that was either much

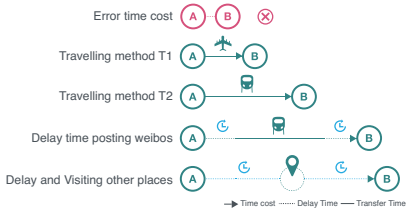


Fig. 5. Illustration of the movement categories derived from Geo-tagged Social Media. Besides the time error, the delay and visiting other places lead to the uncertainty in time interval of the movement.

shorter than that achieved with the fastest transportation means, or significantly longer than what is reasonable.

After that we classified the remaining movements according to time interval. We observed that some active users tended to publish their locations at their departure from or arrival in a new city. In such case the travel time between the two cities is close to the time interval between weibo posts. However, the time interval most of the time should be slightly longer than the travel time, because users do not post weibos promptly at departure and arrival. The time interval in these movements also included two other parts: the time from the they post weibo to their departure from the first city, and the time from their arrival in the second city to the next weibo post. Another complexity is that travel time itself varies due to different planes/trains and potential delays.

We used a Gaussian mixture model (GMM) to describe the time-interval distribution due to various transportation modes. In the areas of transportation research, although various models have been developed to estimate travel time for complex scenarios [17], it has been recognized that for a particular kind of traffic, travel time generally follows a Gaussian distribution [9] and mixture models of travel time fit data well [15]. We further assume that the time interval can also be approximated as a mixture of Gaussian models. With such a method, we can not only extract different transportation modes, but also infer the proportion of different modes.

K denotes the number of transportation modes, α_k denotes the proportion of people choosing mode k, and y denotes the time interval to post Weibo. The distribution of y is:

$$P(y|\theta) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(y|\mu_k, \sigma_k^2), \quad (1)$$

in which $\alpha_k \geq 0$, $\sum_{k=1}^{K} \alpha_k = 1$, μ_k and σ_k are the mean and standard variance of the time interval in type k, respectively.

This model defines the confidence interval as ($[\mu_k$-σ_k, μ_k+$\sigma_k]$). Movements with time interval in this range are considered more re-

liable to extract semantics information.

In order to determine the parameter K, we use a semi-automatic method. We first use the Bayesian Criterion Technique (BIC) [25] to automatically determine the default K value [44, 18]:

$$BIC(\theta) = -2\log p(y|\theta) + d\log N, \quad (2)$$

where θ represents the model parameters, d is the number of parameters in the model ($= 3 * K$ in our situation), and N is the number of input data. The model that minimizes BIC will be selected.

We performed sensitivity analysis on the selection of K. We tested different Ks and found that when BIC is lower, the modeling effect is better. However, when K value is slightly changed by adding one or subtracting one, the modeling effect does not change much. Although the default K value is reasonable in most cases, there are situations in which it does not fit well, or in which users want to set their own value. In such cases, we also allow users to manually choose a K value.

The model inputs are the filtered movements by users, with the exclusion of unreasonable time intervals. The error time interval was sometimes estimated as a Gaussian distribution, so we needed users to verify the matching result and filter the data with time interval less than fastest transportation. Section 5.2 shows a typical scenario. The model outputs are the confidence intervals of each categorized distribution. Users can filter the reliable data based on the distribution. Considering the variability, users could also adjust the parameters of the output confidence interval range and K value in order to better match the movements if they have prior knowledge. Further users need visualization interfaces to filter the categories and analyze the text.

5.2 Experiment Results

We conducted an experiment to illustrate our model with the geo-tagged Weibo data. In the experiment, we loaded two weeks data illustrated in Section 3. We searched for two cities, say A and B, and then filtered people who sent Weibos in both cities. There were 4,791 Weibos posted in A and B. The distance between A and B is about 2,000 km. We extracted 996 (A to B) and 1,201 (B to A) trajectories. There were 361 and 393 unique user counts. In Fig 6, A to B is blue, and B to A is yellow.

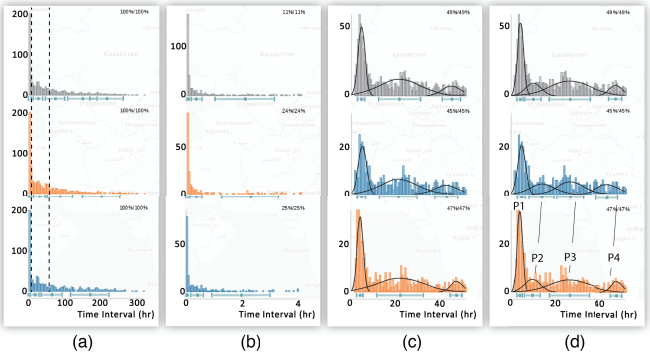


Fig. 6. Experiment results. (a) Overall time interval distribution. (b) Error time distribution. (c) Model output within reasonable time interval range and the $K = 3$. (d) Different settings of the $K = 4$.

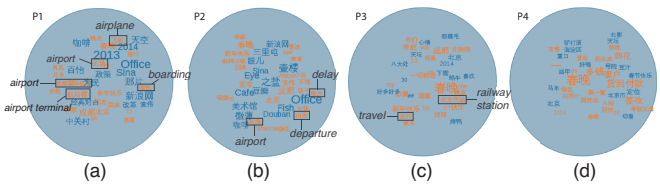


Fig. 7. Keywords visualization for each derived categories from Fig 6d.

Category	Ground Truth	Mean	Sigma	Confidence Interval	Weight
$Plane(M1)$	3.1	4.24	0.91	[3,33,5.15]	0.31
$Plane(M2)$	3.1	5.13	1.62	[3.51,6.75]	0.32
$Train1(M1)$	12-14	11.55	4.26	[7.29,15.81]	0.15
$Train1(M2)$	12-14	8.55	2.87	[5.68,11.42]	0.22
$Train2(M1)$	22-30	24.87	5.43	[19.44,30.3]	0.38
$Train2(M2)$	22-30	25.46	9.14	[16.32,34.6]	0.29

Table 1. The comparison of ground truth and the estimated results from the uncertainty model. Each column is time interval and the unit is hour. $M1$: A to B, $M2$:B to A. $Train1$ and $Train2$ represent fast and normal speed train, respectively.

We initially saw the distribution of the time interval as a long-tail distribution, which means a small number of movements with a large time interval (Fig 6a). We removed these movements based on the threshold of an experience number (60 hours), as a too large time interval doesn't provide clear semantics information. We then investigated the remaining part of the movements, and found there was a peak in a very short time range. We filtered out the error time interval (less than the fastest airplane's time interval (3.1 hours), in Fig 6b). After this operation, the data were within the range of time interval from three hours to 60 hours, which accounted for 45 percent and 47 percent of the total data for the two routes, respectively. according to the BIC, the suggested K values were 3 and 4. K values outside this range led to either insufficient estimation or several duplicated groups. By applying the model, we found three peaks, with an estimated $K = 3$. The mean time intervals of these routes from A to B are 5.13 hours, 21.09 hours, and 42.69 hours. The means from B to A are 4.24 hours, 21.67 hours, and 46.5 hours (Fig 6c).

The observed time ranges are reasonable according to the real situation. We assumed that the first peak matched travelling by plane, and the second peak matched travelling by train. Then we use the time table to verify the hypotheses. The flying time was approximately 3.1 hours. Considering the boarding and departing times, 4.24 and 5.13 hours as the means of flying were reasonable. Train-riding time varied from 12.2 hours to 40 hours. By setting $K = 4$, we got two peaks split from the original one, with mean times of approximately 11.55 hours and 24.87 hours from A to B, and 8.55 hours and 25.36 hours from B to A (Fig 6d). We checked the train time tables, and found two 2 types of trains: High-speed trains with an average travelling time of 12.2 hours, and normal trains with an average travelling time of 26.98 hours. The travel times by normal trains were clearly indicated by the distribution. However, travel times with high-speed rails did not match well. The last peaks of 42.69 and 46.5 hours did not match normal trains and could be due to delayed Weibo posting. The summary is in Table 1.

The proposed data model is practical for analyzing such sparsely sampled data. We filtered the data from each confidence interval of the categorized movements and used text analysis to verify the hypotheses. We extracted the top 30 words from all the selected categories. We got the keywords of *airport*, *plane*, and *sky*, in the short time-period peak (four to six hours) from Fig 7a. We also got the keywords of *train*, *railway station*, and *travel* in the longer time-period peak (16 to 34 hours range) from Fig 7c. There were no primary words related to the travel methods for time intervals of more than 40 hours, from which we would not infer the semantics about the transportations (Fig 7d). We revised the hypotheses of fast-train category, as we found many people mentioned *delay* in this category. This indicated that some people in this category might experience flight delays. We guessed that fewer people went by fast train because of the high price, which was confirmed by the local news. Based on our model, we can reasonably categorize the trips and derive the confidence intervals and the semantics for more reliable movements in each category.

6 Visual Analytics Procedure

This section describes the visual analytics procedure and related visualization tools we used to analyze movements based on geo-tagged

social media data. This procedure supports filtering and aggregation of interesting movements from social media data. With the movement data as input, we utilized the uncertainty model to provide more reliable, categorized movement information that can combine with social media content for further analysis of user movement patterns.

6.1 Extracting Visiting Sequences

We first organize weibos based on user IDs, and extract geo-locations from each weibo. By lining up these locations chronologically, we can reconstruct users' movement sequences, which can be regarded as "trajectories". Different from traditional GPS-based trajectories, which have dense sampling points with a very fine granularity (e.g., at the level of city block), the trajectories in our research have sparse sampling points. The sampling interval varies from hours to days. As described in Section 3, robotic accounts and those with only one weibo are excluded in the preprocessing stage.

6.2 Filtering Places of Interest and Deriving Movement

Initially, in the spatial temporal workspace, users can filter a time range in the time line view as shown in Fig 1a and set a region with a customized spatial range in the map view in Fig 1b for movement analysis. To obtain movement patterns, first we extract places of interest. Based on the observation we conclude some criteria for choosing places.

- Places with at least a fair amount of people visiting, e.g. a big city or an interesting tourism place.
- Places with limited sizes and can be any geometric shape.

Considering the criteria, we adopt a density-map approach based hot place identification. By calculating the density map, our tool can extract top k (set by users) places with the largest amount of weibos. The extracted result is then presented inside a colorized circle. The circle covers the geographic region of selection. Users can add, resize, move, or delete places by manipulating the circles. Users can also draw a polygon as a filter to find places within the region specified by the polygon (Fig 8). The procedure also adopts K-means and other density-based clustering methods, such as DBScan to generate the places of interest for analysis. However, the K-means approach cannot be applied to only focus on specific regions (e.g., a region defined by a polygon). Although DBScan can generate results in any shaped regions with high density, the computation complexity of DBScan is too high for on-line interactive filtering. We combine density map calculation and flexible user manipulation to support the places of interest extraction. After extracting the places, the system can estimate location names from a map server, and assign one color to each selected place.

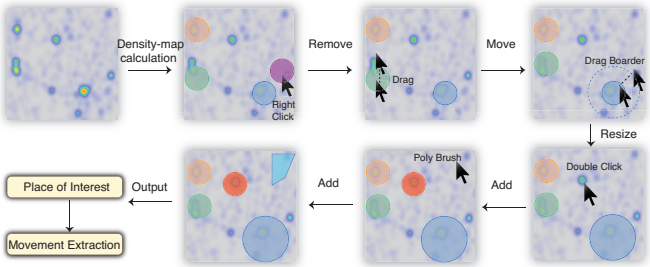


Fig. 8. Steps of automatic place suggestion and interactions for filtering the places of interest.

Movements derive from extracted places. There are two types of movement between two places A and B: Direct visiting, and indirect visiting, passing by other place(s). For indirect visiting, we accumulated the time interval and geo-distance along the stops from A to B. The intermediate places might be on the way from A to B, which are accurately accumulated. However, people might go to another place, say C, outside the route from A to B. This deviation may introduce complexity (Fig 5). These filtered places and movements become the inputs for aggregation analysis.

6.3 Dynamic Aggregation and Filtering

Aggregating the movements of many people can help us explore movement patterns [6]. Our system provides interactive tools for multiple attribute aggregations and filtering of movements. The arrows on the map show derived, aggregated movements between places. Tools allow users to easily filter movements (Fig 9).

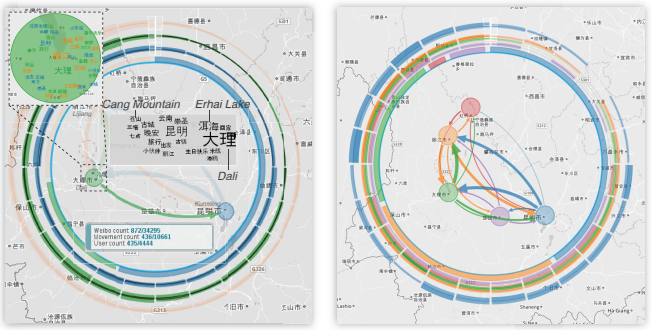


Fig. 9. Aggregation and interactions on the spatial view. (a) Highlighting the movement for details. (b) Compact circular layout, with level of details showing as enlarged histograms.

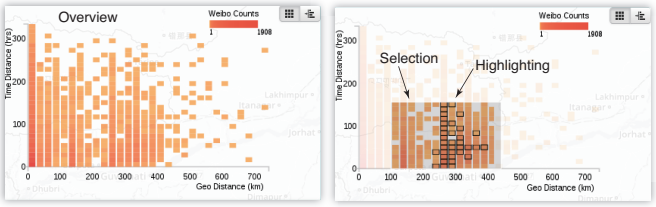
In our system, users can interactively control various parameters for movement aggregation.

- **Analysis based on places:** By aggregating the Weibos in one place, we can extract the number of visits, unique visitors and the average time interval of the movement in/out of the place. Users can get detailed information by hovering the cursor over the place. In the inner layer of the circle, the size of each arc band encodes the number of Weibos in each region (Fig 9). For example, the blue region has approximately twice as many Weibos as the green region.
- **Analysis based on periodicity:** As an overview, temporal distribution of Weibos is visualized in the time line (Fig 1a), for understanding the periodic behaviors and outliers. Surrounding the analyzed region, we encode the periodic temporal information along the circle (Fig 9a). Users can set the granularity (such as by days or by weeks). The histograms are arranged in a clockwise fashion, representing 24 hours of a day or seven days of a week. Users can observe the periodic distribution of the Weibos, peaks, or regular amounts of Weibos in different places. Along the radial direction, each band identical to each place has two histograms with opposite directions, encoding movements in and out movement of the place. Users can apply circular brushes in each band, to analyze the temporal behavior of people's visiting pattern. When the number of analyzed cities is larger, our system would use the compact layout which compresses each histogram to a thick line, with the opacity to indicate the number of Weibos (Fig 9b). Users can select clusters highlighted as histograms among the pixel band. Sorting by clicking on the band can aid the comparison tasks for each time period.
- **Analysis based on movement attributes:** Time-interval and geo-distance distribution describe the important features of movements. In Section 3, we saw that the distributions vary much from person to person. We provide users with the filtering functions for these attributes in the ST Matrix view (Fig 10). It encodes the number of Weibos by color in each aggregated bin of geo distance and time interval. The x axis is the geo-distance and the y axis is the time interval. To get reasonable ranges to analyze, users can control the x and y ranges based on the distribution (for example, ignoring the movements of 150+ hours).

Our system enables users to explore the aggregated data in a highly interactive fashion. Based on the extracted movements and places of interest, users can apply movement filters. Those movements that pass

at least one place of interest would be stored. This lays the foundation for exploring multiple combinations of visiting sequences and frequent patterns. Fig 1b (left bottom) shows a histogram encoding the movement count of each aggregated movement. Users can brush a subset of the aggregated movements.

With filtered movements based on various criteria, users can conduct detailed pairwise analysis.



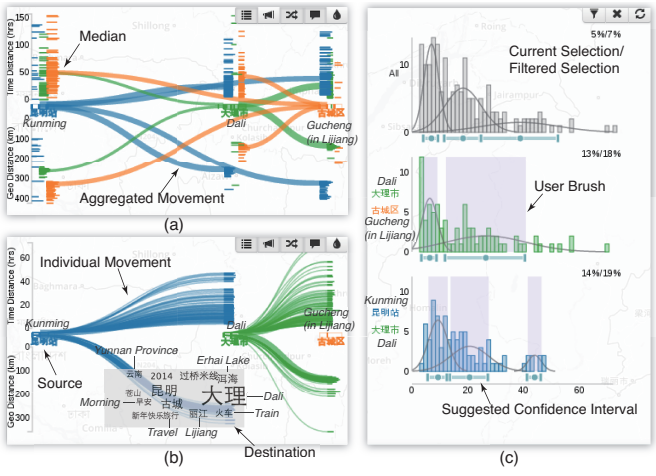
(a) (b)

Fig. 10. Aggregation of time interval and geo distance. (a) ST Matrix View. (b) Brushing and highlighting in the ST Matrix View.

6.4 Detailed Analysis based on Uncertainty Model

After users filter the aggregated movements in spatial, temporal and attribute domains, movement details are exposed in the ST-Detail view (Fig 1d) and ST-Model view (Fig 1e). The design goal of these views is to support time-interval and geo-distance filtering based on location pairs. After applying the uncertainty model, users can keep reliable movements of each category with the confidence interval.

As shown in the ST-Detail View (Fig 11a), the x axis is 1-D projected places and the y axis, with two directions, is time interval (hours) on the top and geo-distance (km) at the bottom. The x-axis default order is based on the distance of each place in the 2D space. After setting a starting point (default is the northwest corner), we used a greedy algorithm to add each place in a sequence with the smallest distance. This method can provide a reasonable projection sequence in a 1-D space, especially with not many places of interest. The x axis can be sorted by selected visiting sequences from left to right (Fig 11b).



(a) (b) (c)

Fig. 11. Detailed analysis based on Uncertainty Model. (a) ST Detail View. (b) Frequent movement pattern highlighting with X axis reordering. (c) ST Model View for uncertainty analysis and interactive filtering.

In the ST-Detail view, each movement is represented as a two-folded shape curve. It starts from the middle of the y axis within the source-places position, and curves to the two destinations along the y axis, encoding the time interval and geo-distance. Within each place in the x axis, the destinations of each movement are uniformly laid out, based on their sources. Users can compare the pairwise flow counts coming from different sources and the time-interval distributions. When users hold the cursor over the curve, the corresponding source and destination are highlighted, thus helping users to understand the speed distribution. Users can also hide all arcs and show aggregated movement with the median values to reduce the visual complexity. This view also provides heatmaps to show the distributions. The scale of both distance and time data in the view is adjustable, according to user's demands. When users brush the movement in the ST-Detail view, the ST-Model view is also updated. By default, it shows all the pair-wised movements. Users can control to show one or more of the above visual encodings by operating the buttons on the panel, to reduce the cognitive loadings. The curve used here has a bundling effect and reduces the clutter. We provide clear labels of location and highlighting effects, to avoid misleading users with the interpolation effects. Besides the time-interval variation, we also care about the geo-distance variation, because there may be different routes between the location pairs, which leads to the distance variation.

The ST-Model view (Fig 11c) combines with small multiples of the time-interval histograms for movements between two places. The time interval between two regions indicates the different categories of movement patterns, as we discussed in Section 5. Users can adjust the time-interval range of the histograms from the ST Matrix view and the ST Detail view. After the user filters out the outliers and focuses on the most probable travel time range, our system calculates the GMM-based uncertainty model, drawing the distributions that match the histograms. The confidence interval of each distribution is visualized as a blue bar beneath each histogram. Because the distribution might vary and not be fully approximated, users can drag the edge of the confidence interval to fine-tune the results. User can also directly filter all the movements with the time interval falling into the confidence interval (Fig 1e). Some cases that match a Gaussian distribution may not have practical meanings. Users can eliminate these by manipulating the filtering brushes. Users can filter the confidence intervals with multiple brushes in each time-interval histogram. Furthermore, users can verify the categories and interactively explore the semantics of the movement in other views.

6.5 Text Analysis and Semantics Exploration

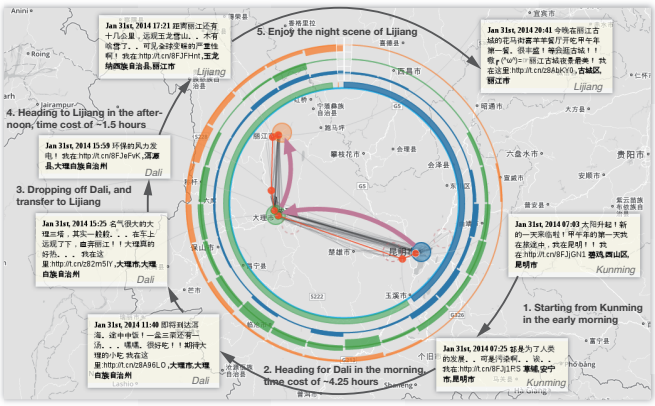


Fig. 12. Detail exploration for semantics. Individual movement (red) selected from frequent pattern (pink) are highlighted with details.

Geo-tagged social media data has the advantage of textual content over traditional GPS trajectories. Users can analyze the text information to further verify the movement categories and explore the semantics. As discussed in Section 2.4, we supported the text analysis for *stops* at certain locations, and *moves* between consecutive stops [28]. First, users can extract keywords from each place. The keywords belonging to the same origin have one color. (Fig 9a). Second, when users brush the movement categories in the ST-Detail view, the keywords of selected Weibos pop up in the filtering window and layout as a wordle (Fig 11b). After the keywords extraction, users can click the keyword or raw data on the map to trace the original Weibo contents (Fig 12). Through these interactions, users can verify or reject the categorized movement derived from the uncertainty model. There-

fore, combined with the contents data, users can derive the semantics of categorized transportation, and infer the trip goals or interesting movement behaviors.

Based on the reliable movement of each category, users can extract frequent visiting patterns among places of interest (Fig 1b). Afterward, users can filter parts of the frequent patterns to explore the spatial, temporal, and attribute aggregations. The filtering and verification process work together to complete the visual analytics loop.

7 System Implementation

Our system users client-server architecture. The client is built with HTML5/Javascript, and the server-side services are implemented with Python and MongoDB. The client is implemented with multiple web techniques, including Leaflet Maps [1], D3.js [10], and WebGL. The spatial data is organized in a quadtree for efficiency. Our algorithm for mining frequent movement patterns is based on the work by [45]. Weibo data-gathering and pre-processing are the two of the primary services. We crawled Weibo data through the open APIs by Sina Weibo. Our crawler can gather one million new geo-tagged Weibos per day, occupying 2.5 percent of the overall Weibos. We have accumulated more than 50 million Weibos with geo-tags in two months.

8 Case Studies

This section demonstrates how our system can be used to find reliable movement patterns with categorized semantics from geo-tagged Weibo data.

8.1 Case Study 1: Movement Analysis in Yunnan

We first studied the movement of tourists in Yunnan province, a popular tourists destination in southern China. We extracted 116,704 valid Weibos with geotags in Yunnan from 26,571 unique users. The Weibos were posted between Jan. 27th and Feb. 9th, 2014.

After selecting a circular area, covering the majority area of Yunnan, the system identified three major cities/regions, based on the popularity. As shown in Fig 9a, the cities were Kunming, Dali and Lijiang. By highlighting aggregated movements, we knew that there were 436 movements going from Dali to Kunming. The top keywords aggregated from the Weibo contents indicated people's footprints, such as *Cang Mountain* and *Erhai Lake*, which are famous sites in Dali. Furthermore, we could filter the in/out time of the movements (such as starting from Kunming in the morning and arriving in Dali in the afternoon or night). The filtered movements were also highlighted in the ST Matrix view (Fig 10). We filtered the ST Matrix view with a time interval of 0 to 150 hours and a geo-distance of 100 km to 450 km.

As shown in Fig 1e, we set the filter according to the suggested confidence intervals of the extracted categories. The filtered movements are approximately 58 percent of the original movement data. Based on the filtered results, we can generate the frequent visiting patterns on the map, by setting the least-visited places as 3. We found that the most frequent pattern started from Kunming, then to Dali, and finally to Lijiang (Fig 1b). We clicked the route, and found that 40 instances actually took this route in the selected two weeks. Their movements made up approximately 18 percent of the total 223 movements, which is a meaningful number worths further exploration.

Fig 11c shows two categories in the movement from Dali to Lijiang and three categories from Kunming to Dali. The mean times of the first peak were five hours from Dali to Lijiang and seven hours from Kunming to Dali. These time intervals matched the elapsed time of one-way transportation by train or bus. We regarded these movements as real-time sampling movements. The other three matches generally indicated how long people stayed stay in each city: One day in Dali, one day in Lijiang, and two days in Kunming, respectively.

By selecting the confidence intervals in the first category, we can drill down to the details and search the patterns. Fig 12 shows a typical visitor's route, which started from Kunming in the early morning, stopped at some sites in Dali, and then headed to Lijiang at night. Such a discovery can be confirmed by other sources, such as travel websites indicating this route as one of the most suggested travel routes in this region.

In this case, we emphasized the application of proposed visual analytics procedure in Section 6 for finding interesting places, aggregating spatial temporal data, and filtering data for the model calculation.

8.2 Case Study 2: Movement Exploration in Taiwan

Our second case involves analyzing movements in Taiwan, which is a popular travel destination. We emphasized the use of our tools, which were powered by our uncertainty model, in investigating the semantics of movements. In particular, tools such as brushing for the model output confidence intervals allow users to find interesting patterns and examine their reliability. Combined with other views showing keywords and time-interval distribution, these tools help users discover new information, develop new hypotheses, and verify hypotheses through exploring different spatial and temporal levels.

To understand how people move around in Taiwan, we first observed the overall movements in and out Taiwan. We found that in the Chinese lunar new year period, there were many movements related to Taiwan. Beijing, Shanghai, Xiamen and Guangzhou (Fig 13a) were among the cities most related to these movements. By examining the keywords of filtered Weibos, we learned that Taipei, Kenting, Kaohsiung, Hualien, and Sun Moon Lake were among the most popular places people visited (Fig 13b, 13c). We choose these places as the places of interest, and filtered Weibos based on them. We had 60,240 Weibos from 10,840 unique accounts. These Weibos were posted between from Jan. 27th to Feb. 9th, 2014.

Our system can automatically identify the popular regions and also allow users to change the focus places. We set four places of interest, including Taipei, Kaohsiuing, Kenting and Hualien (Fig 13d). We set the time-interval constraints as 0 to 80 hours, considering the normal staying time. The geo-distance was less than 700 km, as the round-trip between north and south of Taiwan is approximately 550 km. In the ST-Model view, we grouped the time-interval distributions based on the source places, and found several frequent movement patterns, including Hualien to Taipei (red to blue) and Kaohsiuing to Kenting (green to yellow). The time-interval variance of the movement from Taipei to each other city was the overall highest. We can brush the reliable movement of each category for further exploration. When selecting reliable data within all the confidence intervals of all extracted categories, interestingly, we found people usually visited cities in Taiwan in a counter-clockwise direction (Fig 13d).

By brushing the reliable movement data within the short time categories, which were generally the first peak with a mean between 4 to 15 hours (Fig 13-e1), we found the most frequent route A: Kaohsiuing, Kenting, Hualien, and Taipei in a sequence (Fig 13-e3). By clicking this pattern on the map, we saw that Kaohsiuing to Kenting (green to yellow) had the lowest mean and time-interval variance (Fig 13-e2). According to the keywords and detailed contents of Weibos, we found that people enjoyed the snacks at night in Hualien (red). These red histogram peaks at night were also demonstrated by the circular ring (Fig 13d). This pattern seemed to be a classic tourist path, as most of the keywords were about tourism and sightseeing. The average times of filtered movement from Kaohsiung to Kenting and Hualien to Taipei, were relatively short: 4.1 and 8.7 hours, respectively. The time intervals of these places were more accurate, considering people's transportation methods. However, the average time interval from Kenting to Hualien was 23.6 hours, which could be a long-distance to travel or involve stops at other places. Compared with the previous two short routes, the information about this long route is not very reliable. For this pair, we further modified the K as 3, and the original Gaussian was split into two. We found a more reasonable average time interval of 10.2 hours for a trip.

We changed the filters for the confidence interval to the category of each movement pair with the largest weight (the largest histogram area size) calculated from the uncertainty model (Fig 13-f1). We observed the pattern separation in the visiting time (Fig 13-f2). The pairs for the time intervals, from shortest to largest, were: Kaohsiung to Kenting (one day in average), Kenting to Taipei (two days) and Taipei to Kaohsiung (three days). We hypothesized that people likely stayed in the cities or went to other places in between. The time-interval variance

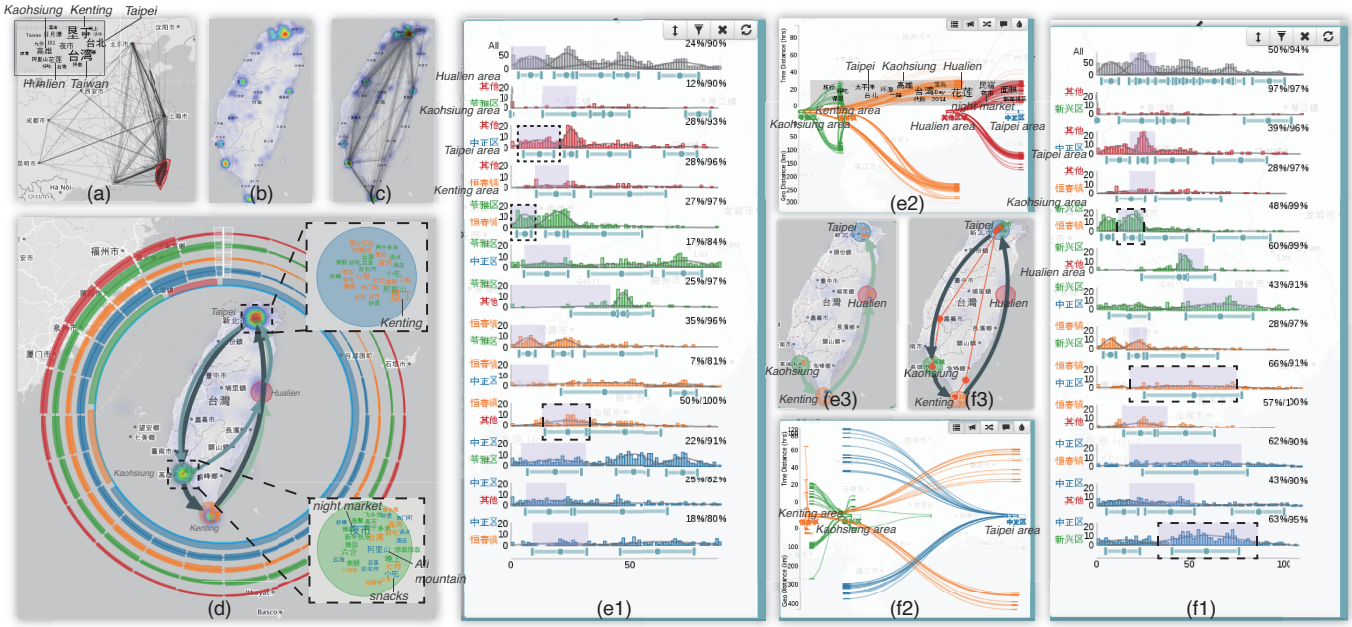


Fig. 13. Movement exploration in Taiwan. (a) Movements from mainland China to Taiwan. (b) Hot visiting places and (c) movements in Taiwan. (d) Top frequent visiting patterns in an anti-clockwise way. (e)(f) Selected movement in the (1) ST Model View, (2) ST Detail View, (3) Map View.

of movements from Taipei to Kaohsiung was high. By interactively exploring the details of related movements, we found that people who went from Taipei to Kaohsiung also visited places such as Taichung or Ali Mountain during the trip. The selected detailed movements are highlighted as a red line in Fig 13-f3. However, more people stayed in Kenting to enjoy scenery, and then went back to Taipei, although some people continued the journey to Hualien.

9 Discussion

In this paper, we have analyzed a special type of movement data that was derived from publicly accessed social media data. It has high population coverage, but the sampling is very sparse. Therefore, the data density and quality pose serious challenges in our research. Much information is inherently missing, resulting in considerable uncertainty. We have put a lot of efforts into dealing with such uncertainties, including performing various data filtering, applying the uncertainty model, and enabling user exploration. The case studies show that our system can give reasonable analysis result on inter-city movement patterns, which are difficult to be captured by traditional GPS data.

Much as we tried, we still need to face the data quality problems. Although large numbers of people and places are involved, we found that Weibo users are more likely to be young and live in developed regions. Therefore, our results could be biased toward specific population groups. We are considering comparing the analysis results with that derived from other datasets (such as Twitter data) in the future. There may be considerable GPS errors in the dataset. Language issues also exist, including slangs and misspellings[26]. To deal with these problems, users could to interactively filter out erroneous GPS data. By extracting top keywords, we removed most language issues. However, there may still be more systematic and automatic methods to deal with GPS and language problems.

Our uncertainty model can support finding semantics in terms of different transportation modes. We can filter reliable movement for subsequent detailed analysis. We used a clustering-based method to identify peaks in the time interval distribution, then tried to assign transportation modes to the peaks. Time-table analysis may be an alternative way. In such a method, users find all possible travel times based on time tables of different transportation modes and filter the movement accordingly. We argue that our uncertainty model is more general than a time-table based method. Our model can be applied in both situations in which time tables are unavailable (such as personal driving) or become inaccurate (such as flight delays). With the clustering-based method, we can observe some unexpected peaks that cannot be captured by time tables. For example, in Section 3, we observed the interesting daily movement time intervals (such as the one-day or two-day peak pattern in Fig 3b). These time intervals are shown as peaks in our distribution.

We envision extending the uncertainty model in several ways. First, we can consider location information and people's habits. If a series of Weibos come from the area of an airport or railway station, by somebody who always promptly sends Weibos, such data could be regarded as more reliable. We can try more advanced modeling in transportation research area, such as multi-state models [17]. They may improve the descriptive capacity under more complex situations.

In the interactive semantic exploration, users can inspect the meanings of movement by watching wordles of top keywords. We believe that other, more advanced, text visual analysis methods can be applied to reveal deeper semantic meanings. For example, we can show the wordles of appearing or disappearing keywords. We can show how the keywords evolve over time, and vary in the geographical space. We can also make visualizations at the Weibo topic level and Weibo text cluster level.

10 Conclusion

In this paper, we presented a visual analytics system to support the analysis of people's movements based on sparsely sampled movement data from geo-tagged social media. With the introduction of uncertainty modeling, users can explore spatial temporal aggregation of movements and filter reliable data. We also provided a set of interactive visualization tools for extracting semantic patterns, by incorporating human's cognitive power and machine's computation. Integrating the statistics results and interactive exploration would empower the analysis of the movement of such sparsely sampled data.

Acknowledgments

The authors wish to thank Ziteng Wang and Siqi Tu for the discussions and evaluations, and the anonymous reviewers for their valuable comments. This work is supported by NSFC No. 61170204, and partially funded by NSFC Key Project No. 61232012 and the National Program on Key Basic Research Project (973 Program) No. 2015CB352500. This work is also supported by PKU-Qihu Joint Data Visual Analytics Research Center.

REFERENCES

[1] leaflet.js, http://leafletjs.com/.

[2] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. Interactive visual clustering of large collections of trajectories. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 3–10, 2009.

[3] G. Andrienko, M. Jern, J. Dykes, S. I. Fabrikant, and C. Weaver. Geovisualization and synergies from infovis and visual analytics. In *Information Visualization*, pages 485–488, 2007.

[4] G. L. Andrienko and N. V. Andrienko. Spatio-temporal aggregation for visual analysis of movements. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 51–58, 2008.

[5] N. Andrienko and G. Andrienko. Spatial generalization and aggregation of massive movement data. *IEEE Transactions on Visualization and Computer Graphics*, 17(2):205–219, 2011.

[6] N. Andrienko, G. Andrienko, H. Stange, T. Liebig, and D. Hecker. Visual analytics for understanding spatial situations from episodic movement data. *KI - Knstliche Intelligenz*, 26(3):241–251, 2012.

[7] M. Azmandian, K. Singh, B. Gelsey, Y.-H. Chang, and R. Maheswaran. Following human mobility using tweets. In *Agents and Data Mining Interaction*, volume 7607 of *Lecture Notes in Computer Science*, pages 139–149. Springer Berlin Heidelberg, 2013.

[8] P. Bak, F. Mansmann, H. Janetzko, and D. A. Keim. Spatiotemporal analysis of sensor logs using growth ring maps. *IEEE Trans. Vis. Comput. Graph.*, 15(6):913–920, 2009.

[9] J. Benjamin and C. Cornell. *Probability, Statistics, and Decision for Civil Engineers*. McGraw Hill, 1970.

[10] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[11] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2649–2658, Dec 2012.

[12] J. Cranshaw, R. Schwartz, J. I. Hong, and N. M. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2012.

[13] G. Fuchs, G. Andrienko, N. Andrienko, and P. Jankowski. Extracting personal behavioral patterns from geo-referenced tweets. AGILE, 2013.

[14] L. Gabrielli, S. Rinzivillo, F. Ronzano, and D. Villatoro. From tweets to semantic trajectories: mining anomalous urban mobility patterns. In *Citizen in Sensor Networks*, pages 26–35. Springer, 2014.

[15] Y. Guessousa, M. Aronb, N. Bhourib, and S. Cohenb. Estimating travel time distribution for reliability analysis. *Transportation Research Procedia*, 3:339348, 2014.

[16] D. Guo. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1041 –1048, 2009.

[17] F. Guo, H. Rakha, and S. Park. Multistate model for travel time reliability. *Transportation Research Record Journal of the Transportation Research Board*, 13(2188):46–54, 2010.

[18] C. Hennig. Methods for merging gaussian mixture components. *Advances in data analysis and classification*, 4(1):3–34, 2010.

[19] M.-H. Hwang, S. Wang, G. Cao, A. Padmanabhan, and Z. Zhang. Spatiotemporal transformation of social media geostreams: a case study of twitter for flu risk analysis. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on GeoStreaming*, pages 12–21. ACM, 2013.

[20] T. Kapler and W. Wright. Geotime information visualization. *Information Visualization*, 4(2):136–146, 2005.

[21] R. Krueger, D. Thom, and T. Ertl. Visual analysis of movement behavior using web data for context enrichment. In *IEEE Pacific Visualization Symposium*, pages 193–200, 2014.

[22] R. Krüger, S. Lohmann, D. Thom, H. Bosch, and T. Ertl. Using social media content in the visual analysis of movement data. In *Proc. Workshop on Interactive Visual Text Analytics*, 2012.

[23] R. Krüger, D. Thom, M. Wörner, H. Bosch, and T. Ertl. Trajectorylenses - A set-based filtering and exploration technique for long-term trajectory data. *Computer Graphics Forum*, 32(3):451–460, 2013.

[24] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura. Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 579–588, 2010.

[25] B. G. Leroux et al. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360, 1992.

[26] J. Li, Q. Qin, J. Han, L.-A. Tang, and K. H. Lei. Mining trajectory data and geotagged data in social media for road map inference. *Transactions in GIS*, 19(1):1–18, 2015.

[27] T. Liebig, G. Andrienko, and N. Andrienko. Methods for analysis of spatio-temporal bluetooth tracking data. *Journal of Urban Technology*, 21(2):27–37, 2014.

[28] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan. Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4):42:1–42:32, 2013.

[29] D. Phan, L. Xiao, R. Yeh, and P. Hanrahan. Flow map layout. In *Proceedings of IEEE Symposium on Information Visualization*, pages 219–224, 2005.

[30] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using mobile phones to determine transportation modes. *ACM Trans. Sen. Netw.*, 6(2):13:1–13:27, 2010.

[31] D. Ren, X. Zhang, Z. Wang, J. Li, and X. Yuan. Weiboevents: A crowd sourcing weibo visual analytic system. In *Pacific Visualization Symposium (PacificVis) Notes, 2014 IEEE*, pages 330–334, March 2014.

[32] A. Slingsby, M. Kelly, J. Dykes, and J. Wood. Od maps for studying historical internal migration in ireland. In *Poster Proceedings of IEEE Symposium on Information Visualization*, 2012.

[33] T. Sohn, A. Varshavsky, A. LaMarca, M. Y. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W. G. Griswold, and E. de Lara. Mobility detection using everyday gsm traces. In *Proceedings of the international Conference on Ubiquitous Computing*, pages 212–224, 2006.

[34] R. K. Straumann, A. Çöltekin, and G. Andrienko. Towards (re) constructing narratives from georeferenced photographs through visual analytics. *The Cartographic Journal*, 51(2):152–165, 2014.

[35] C. Tominski, H. Schumann, G. L. Andrienko, and N. V. Andrienko. Stacking-based visualization of trajectory attribute data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2565–2574, 2012.

[36] A. Vaccari, M. Martino, F. Rojas, and C. Ratti. Pulse of the city: Visualizing urban dynamics of special events. In *Proceedings of 20th International Conference on Computer Graphis and Vision*, 2010.

[37] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1079–1088. ACM, 2010.

[38] H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Proc. IEEE Conference on Intelligent Transportation Systems*, pages 318–323, 2010.

[39] Z. Wang, T. Ye, M. Lu, X. Yuan, H. Qu, J. Yuan, and Q. Wu. Visual exploration of sparse traffic trajectory data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1813–1822, Dec 2014.

[40] L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[41] J. Wood, J. Dykes, and A. Slingsby. Visualisation of origins, destinations and flows with od maps. *The Cartographic Journal*, 47(2):117–129, 2010.

[42] C. Xia, R. Schwartz, K. Xie, A. Krebs, A. Langdon, J. Ting, and M. Naaman. Citybeat: Real-time social media visualization of hyper-local city data. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 167–170, 2014.

[43] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 27(6):52–59, Nov. 2012.

[44] K. Yu. Generating gaussian mixture models by model selection for speech recognition. Technical report, School of Computer Science, Carnegie-Mellon University, 2006.

[45] M. Zhang, B. Kao, C.-L. Yip, and D. Cheung. A gsp-based efficient algorithm for mining frequent sequences. In *Proceedings of the International Conference on Artificial Intelligence*, 2001.

[46] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins. Fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE Transactions on Visualization and Computer Graphics*, 20, 2014.

[47] Y. Zheng, L. Liu, L. Wang, and X. Xie. Learning transportation mode from raw gps data for geographic applications on the web. In *Proc. international conference on World Wide Web*, pages 247–256, 2008.