Mining Travel Patterns from Geotagged Photos

YAN-TAO ZHENG
Institute for Infocomm Research, Singapore, yantaozheng@gmail.com
ZHENG-JUN ZHA1, TAT-SENG CHUA
National University of Singapore, Singapore, {zhazj,chuats}@comp.nus.edu.sg

Recently, the phenomenal advent of photo-sharing services, such as Flickr and Panoramio, have led to voluminous community-contributed photos with text tags, timestamp and geographic references on the Internet. The photos, together with their time- and geo-references, become the digital footprints of photo shooters and implicitly document their spatiotemporal movements. This study aims to leverage the wealth of these enriched online photos to analyze the people’s travel pattern at the local level of a tour destination. Specifically, we focus our analysis on two aspects: (1) tourist movement patterns in relation to the regions of attractions (RoA), and (2) topological characteristics of travel routes by different tourists. To do so, we first build a statistically reliable database of travel paths from a noisy pool of community-contributed geotagged photos on the Internet. We then investigate the tourist traffic flow among different RoAs, by exploiting Markov chain model. Finally, the topological characteristics of travel routes are analyzed by performing a sequence clustering on tour routes. Testings on four major cities demonstrate promising results of the proposed system.

Categories and Subject Descriptors: H.4.0 [Information Systems Applications]: General

Additional Key Words and Phrases: travel pattern mining, geotagged photos

1. INTRODUCTION

The prevalence of photo capture devices, together with the advent of media-sharing services like FlickrTM, have led to voluminous digital photos with text tags, timestamp and geographical references on the Internet. In the era of Web 2.0, the text description and geo-references of photos are mostly socially generated and collaboratively contributed by web communities. Different from other online media, the geotagged2 photos connect geography, time and visual information together and provide an unprecedented wealth of data to accomplish the geographic-related multimedia and vision tasks that were unattainable in the past.

In this study, we focus on analyzing the tourist mobility and travel behaviors based on the geotagged photos. When people take photos at a geographic location, the photos become digital footprints marking their physical presence [Girardin 2008]. The time- and geo-references associated with a sequence of photos then manifest the spatio-temporal movements of the photo shooter. Figure 1 shows the travel paths generated from geotagged photos in London plotted on Google Earth. As shown, in spite of noise, the travel paths from geotagged photos reveal much information about people travel behavior, spatial movement patterns, tourist density and their common travel trajectories. Understanding these mobility patterns can

1Corresponding author.
2Geotagging, or georeferencing, here refers to associating a media resource with geographical/location information.
be tremendously useful to many mobile and location based social and multimedia applications, such as tour guide and recommendation systems.

A variety of methods are available to acquire people travel data, such as survey on people’s location histories [Lau 2007; Lew and McKercher 2002; McKercher and Lew 2004] and automatic location sensing devices like GPS [Zheng et al. 2009; Mckercher and Lau 2008]. The cost, scalability and privacy issues, however, hinder the effectiveness of these methods. In this paper, we show that the community-contributed geotagged photos on the Internet can tackle these issues and provide an efficacious solution to automatic tourist mobility analysis. Publicly available on the Internet, the geotagged photos make the acquisition of travel itinerary data a trivial task. More importantly, the travel itinerary database can be easily augmented by downloading more photos, enabling tourist mobility analysis to scale up to a multitude of tour destinations.

To mine the travel patterns, we focus our analysis on two aspects: (1) tourist movement patterns in relation to the regions of attractions (RoA) within a local destination, and (2) topological characteristics of travel routes by different tourists. The tourist movement patterns characterize the transition traffic among different RoAs, namely how tourist traffic flows within the local destination. The traffic analysis also helps to indicate the topological relation and connectivity among different RoAs. For example, two RoAs are coupled together, if the tourist traffic always flows from one to other. Similarly, a RoA is identified as a centric one, if it
has influx of tourist traffic from many other RoAs. At other end of the spectrum, the topological characteristics of travel routes reflects the group patterns of tourist travel behavior. Specifically, it signifies the choice of the crowd on travel routes visiting different RoAs.

To perform the two travel analyses above, several challenges, however, need to be tackled. First, online photos are noisy, and thus, it is critical to build a statistically reliable database of travel paths from geotagged photos. Second, tourist movement is, in essence, a sequential event of visiting different places; and a proper model is required to analyze such sequential data. Third, to investigate the topological characteristics of tour routes of different people, a robust similarity metric is needed to measure the similarity of different travel routes. To tackle the first issue, we build the tourist travel path database from geotagged photos, based on the concept of mobility entropy. The rational is simple: the mobile nature of sightseeing renders the photos of a true tourist to be spread over a large spatial extend within the tour destination. Then, a significance test is applied to ensure that the resulting photo path are statistically reliable. Though outliers and noise might still exist, the tourist travel patterns are expected to be statistically significant, when the number of travel paths is large.

To analyze the tourist traffic statistics, we formulate the movement of a tourist as a visit sequence of RoAs in the framework of Markov chain model [Diaconis 2009]. The Markov chain model is widely used in various disciplines to analyze the trend of spatio-temporal movement and outcomes of sequential events [Ishikawa et al. 2004; Upton and Fingleton 1989]. Based on the first-order dependence in Markov chain, we can estimate the statistics of visitors traveling from one region to another. To investigate the topological characteristics of tour routes, we perform sequence clustering on travel routes, in which the modified Longest Common Subsequence (LCSS) [Vlachos et al. 2003] is adopted as similarity metric for better robustness to noise.

Overall, this study proposes to leverage the community-contributed geotagged photos to mine the tourist travel patterns within a local destination. To the best of our knowledge, this is the first approach that substantiates the tourist mobility analysis based on geotagged photos. The main contributions are:

— we propose a statistical approach to build a real-life tourist movement trajectory database from geotagged photos.

— we develop a scheme to analyze the tourist travel patterns in two aspects: (1) tourist movement patterns in relation to RoAs, and (2) the topological characteristics of travel routes.

We demonstrate the proposed scheme on four major cities, including Paris, London, San Francisco and New York City. Experiments show that the proposed approach can deliver promising results. The rest of the paper is organized as follows. We first review related literature work in Section 2 and elaborate on the details of proposed approach in Section 3. Section 4 and 5 presents the experiments. Finally, Section 6 gives the conclusive remarks along with discursion of future work.
2. RELATED WORK

In recent years, the advent of media-sharing services, such as Flickr\textsuperscript{TM} and Youtube\textsuperscript{TM}, has led to a vast amount of community-contributed photos and videos available on the Internet [Zha et al. 2010]. Together with socially generated textual and spatiotemporal metadata, these enriched multimedia data have spurred much research on discovering knowledge and patterns of our human society. Crandall et. al [2009] first investigate ways to organize a large collection (∼ 35 million) of geotagged photos and predict locations of photos from visual, textual and temporal features. Kalogerakis et. al [2009] attempted to estimate the geo-locations of a sequence of photos. Kennedy et. al [2007] proposed to discover aggregate knowledge of a geographical area, by analyzing the spatiotemporal patterns of tags of Flickr photos in the area. Similarly, Rattenbury et. al [2007] and Yanai et. al [2009] analyzed the spatiotemporal distribution of photo tags to reveal the inter-relation between word concepts (namely photo tags), geographical locations and events. Li et. al [2008] and Zheng et. al [2009; 2009] learned the geographical and visual appearance knowledge of tourist landmarks from community contributed photos on the Internet. Similarly, Li et. al [2009] explored ways to perform large-scale landmark image classification. The commonality between the aforementioned work and this study is that they all aim to extract some knowledge and patterns from photos with textual and spatiotemporal metadata, while the difference is that this study focuses on mining travel patterns of tourists. In this aspect, this study shares similar vision with [Kurashima et al. 2010] and [Lu et al. 2010]. The major different is, however, that this work builds a database of tourist itineraries out of noisy Internet photos, while the approaches in [Kurashima et al. 2010] and [Lu et al. 2010] assume that all geotagged photos are pertaining to tourists and tourist itineraries are readily available. Choudhury et. al [2010a] [2010c] explored the construction of travel itineraries from geotagged photos too. The itinerary is defined as the list of landmarks that a person visit, which is generated by mapping photo geospatial coordinates to the latitude/longitude of the given landmarks. In contrast, the itinerary in the proposed work is defined to be the spatio-temporal movement trajectory with much finer granularity.

The study on tourist travel pattern within a tour destination has been a popular geographic research topic. Mckercher and Lau [2008] attempted to identify the movement patterns and styles of tourists within an urban destination. Asakura and Iryo [2007] investigated the topological characteristics of tourist behavior in a clustering approach. Lewa and Mc Kerchera [2006] explored the urban transportation and tourist behavior modeling to identify explanatory factors that influence tourist movements. Compared to the work above, this study differs mainly in two aspects. First, the travel information in the previous work is mainly acquired via a manual survey with a limited number of tourist respondents. Consequently, these studies covered only one or two tour destinations. In contrast, the proposed approach mines the travel information from Internet photos, which renders the data acquisition highly efficient, and thus, allows the travel analysis to easily scale up to a multitude of destinations. Second, this study analyzes both traffic flow pattern and topological characteristics of travel routes, while most existing work focus on traffic pattern only.
The proposed system is also closely related to tour route recommendation systems [Lewa and McKercher 2006; Asakura and Iryo 2007; De Choudhury et al. 2010b; Zheng and Xie 2011]. Choudhury et al. proposed a system to construct intra-city travel itineraries for vacation planning. Elias and Sester [2006] attempted to recommend a navigation route that traverses along a set of tourist landmarks with minimal complexity of route description, while Zhang et al. [2008] focused on searching tourist routes to visit a few tourist attractions with the shortest traveling distance. Similarly, Kawai et al. [2009] developed a personalized tour recommendation system to visit tourist sights around a tour destination.

Characterized by spatiality and temporality, tourist movement trajectory can be regarded as one kind of spatio-temporal data, which connects this study to the spatio-temporal data mining in Knowledge Discovery in Databases (KDD) community. Peng and Chen [2003] attempted to mine the movement patterns of a mobile user and devised a data allocation method in a mobile computing environment. Verhein and Chawla [2008] discovered the spatio-temporal association rules that described how mobile objects move between regions over time. Giannotti et al. [2007] proposed to represent movement trajectory pattern in terms of “region of interest”, in which aspect our approach shares similar vision. Nevertheless, in contrary to the aforementioned work, the proposed approach focuses on delivering a tourist travel behavior analytic mechanism that can be readily applied on a multitude of tour destinations.

This work is based on and extends our previous work [Zheng et al. 2011], which focuses merely on analyzing tourist movement patterns in relation to RoAs. Compared to [Zheng et al. 2011], this work extends it in the following aspects. First, besides the travel patterns among RoAs, it investigates the topological characteristics of travel routes. To do so, a sequencing clustering method is developed. The popular tour routes, relax-trips and busy-trips are also analyzed. Second, it explores two different methods for validating the statistical significance of tourist photo paths, which are based on the Poisson and Normal distributions respectively.

3. APPROACH
3.1 Building the Travel Path Database

Given a set of geotagged photos $P = \{p\}$ in a tour destination, we first build the database of tourist travel paths. A photo $p$ is a tuple $(\theta_p, \wp_p, t_p, u_p, \varrho_p)$, contain-

![An example of a tourist movement](image)

**Fig. 2.** Toy example of a tourist movement trajectory.

ACM Transactions on Intelligent Systems and Technology, Vol. V, No. N, Month 20YY.
Yan-Tao Zheng et al.

ing the unique photo ID $\theta_p$, tagged GPS coordinates $\varphi_p$, in terms of latitude and longitude, time stamp $t_p$, when photo was taken, photographer/uploader ID $u_p$, and tagged text $\varphi_p$. Here, the tourist travel movement is modeled at a daily basis. According to photographer ID $u_p$, we organize photos of each photographer in a day in a chronological sequence $<p_0, ..., p_k>$. By representing the geographical calibration $\varphi_p$ of photo $p$ in ordinary Cartesian coordinates $(x_p, y_p)$, we define a photo trail as the movement of a photographer, in the notation of [Giannotti et al. 2007], as follows:

**Definition 1.** The **photo trail** of a traveler corresponds to a spatio-temporal sequence (ST-sequence) $S = <(x_0, y_0, t_0), ..., (x_k, y_k, t_k)>$, where $t_k$ is the time-stamp of photo $k$ and $(x_k, y_k) \in \mathbb{R}^2$.

Based on the Definition above, we construct the photo trail of a photographer/uploader, by concatenating photos in the order of their time-stamp in a daily basis. We then classify these spatio-temporal sequences to tourist and non-tourist travel paths. The premise for classification is that the mobile nature of sightseeing renders the photos of a true tourist to be spread over a large spatial extent within the tour, as shown in Figure 2.

**Entropy based Mobility Measure.** In a probabilistic perspective, the mobility complexity leads to a geospatial distribution of photos with reasonably high entropy. We, therefore, exploit this **mobility entropy** to discriminate the tourist and non-tourist movement trajectories, by utilizing the concept of Shannon entropy in Information Theory. Let $p(x, y)$ denote the geospatial density of photos with geospatial coordinates $(x, y)$ pertaining to the photographer/uploader. The mobility entropy $H_{mob}(S)$ of a movement trajectory $S = <(x_0, y_0, t_0), ..., (x_k, y_k, t_k)>$ is computed as follows.

$$H_{mob}(S) = - \sum_x \sum_y p(x, y) \log p(x, y),$$

where $p_{ij}(x, y)$ is a discrete geospatial distribution of photos in grid $(i, j)$. By partitioning the tour destination into $n \times m$ grids$^3$, $p_{ij}(x, y)$ is estimated by the counts of photos in grid $(i, j)$. To discriminate photo trails, we empirically set a mobility entropy threshold $\varepsilon_{mob}$. The photo trail $S$ is then classified as a tourist one, if $H_{mob}(S) \geq \varepsilon_{mob}$. $\varepsilon_{mob}$ is empirically set to 0.2 in our experiments.

**Statistical Significance of Travel Paths.** We need to ensure that the resulting travel path database are statistically reliable. To do so, we perform statistical significance test on the resulting photo trails. Here, we characterize the photo trajectory of a photographer/uploader with the number $k$ of his/her visiting places. To some extent, the number of visit places indicates the mobility complexity of tourist. The number of visiting places is determined by the number of photos with unique geospatial coordinates. Geospatial coordinates within a small distance (30m in experiments) will be considered as one. We devise two methods to validate the statistical significance of discovered travel paths.

---

$^3$The values of $n$ and $m$ are determined in the way that the size of the grid is equal to 1km $\times$ 1km in the experiments.

ACM Transactions on Intelligent Systems and Technology, Vol. V, No. N, Month 20YY.
Method 1: If we model the number of visiting places $k$ in tourist itinerary as the number of events occurring in a fixed interval, $k$ will effectively become a Poisson random variable, which can be modeled by the Poisson distribution:

$$P(k; \mu) = \frac{\mu^k e^{-\mu}}{k!},$$

(2)

where $\mu$ is equal to the expected number (mean) of visiting places. In probability theory and statistics, the Poisson distribution is a discrete probability distribution that models the probability of a given number of events occurring in a fixed interval. In the setting of this work, the “number of events” corresponds to the number of visiting places in a tour itinerary. The validity of using Poisson distribution depends on the assumption that the number of visiting places $k$ in a tour is randomly distributed. In other words, individual tour itinerary samples should be independent from each other. This independence assumption is reasonable, in the way that a tour itinerary depends on many factors, including tourist personal preference, his/her prior visits, total tour duration, tour destination demography, etc.

With the Poisson model, we apply the statistical significance test on photo itinerary $k$. Here, we set the level of significance at 5%. In this Poisson experiment, a photo trajectory of $k$ visiting places might be generated by (1) a normal tour or, (2) by some sort of noise (such as camera clock, GPS error, etc). The probability of the photo itinerary happening by chance of noise must be less than the level of significance. To do so, we discard photo itinerary $k$, whose Poisson probability $P(k; \mu)$ is less than the level of significance of 5%. In other words, we discard photo itineraries that visit too few or too many places. The reason for discarding itineraries with too many place is that itineraries that move fast are highly probable to the results of camera clock and GPS noises.

Method 2: As the number of photo trajectory samples is relatively large, we can also approximate the tourist itinerary in terms of number of places $k$ with normal distribution $\mathcal{N}(\mu_k, \sigma_k)$ [Degroot and Schervish 2001]. The mean $\mu_k$ and standard deviation $\sigma_k$ are estimated from the trajectory samples. Similar to Method 1, this normal model of tourist itinerary implicitly assumes individual tour itinerary samples are independent from each other.

With the normal model, we apply Z-test to evaluate the statistical significance of a photo trajectory $k$. The null hypothesis $H_0$ in the test is defined as follows.

$$H_0 : k \sim \mathcal{N}(\mu_k, \sigma_k),$$

(3)

where $H_0$ states that the itinerary $k$ generated from geotagged photos is statically significant, rather than generated from noisy photos by chance. The z-score $z$ is then computed as below:

$$z = \frac{(k - \mu_k)}{\sigma_k/\sqrt{n}},$$

(4)

where $n$ is the number of photo itinerary samples. By looking up the z-score in a table of the standard normal distribution, the corresponding p-value can be obtained. A lower p-value indicates a lower probability that the null hypothesis $H_0$ holds. If p-value is less than a threshold $\tau$, the null hypothesis $H_0$ is rejected.
and the photo itinerary is deemed to be statistically insignificant or unreliable and discarded subsequently. Same as method 1, we set the level of significance $\tau$ at 5%.

3.1.1 Discovering Regions-of-Attractions (RoA). As our focus is on tourist mobility analysis at macro-level, a comprehensive list of regions of attractions within a tour destination is needed. Here, we define the region of attraction (RoA) as follows.

**Definition 2.** A region of attraction $r$ is a spatial extent in the geographical feature space of Cartesian coordinates $(x, y)$, where a considerable number of tourist movement trajectories pass through. RoA can be modeled as a spatial neighborhood function $F(x_i, y_i) : \mathbb{R}^2 \rightarrow \{0, 1\}$.

In the spirit of our previous work [Zheng et al. 2009], we develop a density-based model to discover regions of attractions, by analyzing the geospatial distribution of geotagged photos. As stated in Definition 2, a region-of-attractions is a communal and interpretable spatial concept shared by a multitude of people. In other words, a RoA corresponds to a spatial extent, where many tourists visit and photograph. Clustering on geotagged photos then become an intuitive solution to discover the list of regions of attraction.

Here, we adopt DBSCAN algorithm [Ester et al. 1996] to perform the geospatial clustering on geotagged photos for the following reasons. First, DBSCAN is a density-based clustering algorithm. Intuitively, it tends to identify regions of dense data points as clusters. This density driven approach just fits our task well, as the high density of photos implicates the popular appeal of the region. Second, DBSCAN algorithm supports clusters with arbitrary shape. This is critical to our task, as shapes of RoA can be spherical, linear, elongated etc. Third, DBSCAN is demonstrated to have good efficiency on large-scale data. (cf. [Ester et al. 1996] for more details of DBSCAN.)

After obtaining clusters of photos, we then determine the name and spatial extent of RoA, by examining the GPS coordinates and text title of component photos. We compute the frequency of n-grams of all photos text titles in each cluster. The name of RoA is determined to be the photo title with highest frequency. The geospatial extent of RoA is the area defined by the GPS coordinates of its member photos. Similar to [Zheng et al. 2009], the resulting RoA is validated by the number of unique photographers/uploaders. This is to further ensure the popular appeal of RoA.

3.2 Transition Traffic between RoAs

Based on the concept of RoA, we define the transition statistics between RoAs as below.

**Definition 3.** The transition statistics between RoAs depicts how tourist traffic flows from one RoA to another. It is defined as transition probabilities among different RoAs.

By defining the tourist travel as a sequence of RoA, we investigate how tourists move from one RoA to another using the Markov chain model, in the spirit of [Xia et al. 2009].
Mining Travel Patterns from Geotagged Photos

Fig. 3. Examples of similar tour routes.

Here, we like to point out that travel path is the actual moving trajectory of a tourist, which is represented by a set of photos with timestamp and GPS coordinates. RoA (region of attraction) is a region where many tourists visit (and take photos). Different from travel path, a travel route consists of a sequence of RoAs.

In a statistical perspective, we model the movement of a tourist as an independent stochastic random process. The state space of the stochastic process is the set of RoA \( \{ r \} \) in the tour destination. Let \( T = \{0, 1, 2, \ldots \} \) denote the time index of the moves of a stochastic process. The stochastic process representing tourist movement \( \{ R_t \}_{t \in T} \) is referred to as a Markov chain (MC), if the value of next state does not depend on any previous states, given the value of current state, as defined below.

\[
P(R_{t+1} = r_{t+1} | R_t = r_t, R_{t-1} = r_{t-1}, \ldots, R_0 = r_0) = P(R_{t+1} = r_{t+1} | R_t = r_t),
\]

where \( R_t \) is the random variable of RoA, \( r_t \) is a value of \( R_t \) and \( r_t \in \{ r \} \). In Markov chain model, each move in the state space \( \{ r \} \) is called a step. As each movement occurs after one unit time step, the stochastic process of tourist movement is modeled by a stationary discrete Markov chain. The transition probability \( P(r_j | r_i) \) from RoA \( r_i \) to \( r_j \) can then be estimated by counting the tourists moving from RoA \( r_i \) to \( r_j \). Accordingly, the RoA transition can be represented by a directed graph \( G = (V, E) \), in which vertex \( V \) corresponds to RoA and edge \( E \) represents the transition statistics.

3.3 Topological Analysis of Travel Routes

To investigate the topological characteristics of travel routes by different tourists, we define the travel route as below.

**Definition 4.** The travel route is the traveling course of a tourist, which is constituted by a visit sequence of RoAs \( R = < r_0, r_1, \ldots, r_i, \ldots, r_m > \), where \( i \) is the index of RoA and \( m \) is the length of the tour route. The cardinality \( |R| \) of a tour route is the number of RoAs visited in the route.

Topological analysis of travel routes is a task of group pattern mining on tourist mobility behavior. Clustering is an intuitive solution. However, travel route is a visit chain of different RoAs; and a principled metric is needed to estimate the similarity of two sequences. This metric needs to be robust to noise, as similar travel routes can always have slight differences, as shown in Figure 3.
Algorithm 1 RouteSimilarity(Travel route $R_x$, $R_y$)

**input:** $R_x = \langle r_{x_0}, r_{x_1}, \ldots, r_{x_n} \rangle$, $R_y = \langle r_{y_0}, r_{y_1}, \ldots, r_{y_m} \rangle$ and minimum bounding envelope $\epsilon$

**output:** longest common subsequence length $|lcss|$ and similarity of $R_x$ and $R_y = |R'_x| - \max(|R_x|, |R_y|)$

if isempty($R_x$) or isempty($R_y$) then
    return $|lcss| = 0$
else if $r_{x_n} = r_{y_m}$ and $|n - m| \leq \epsilon$ then
    $|lcss| = 1 + \text{RouteSimilarity}(R^-_x, R^-_y)$
else
    $|lcss| = \max(\text{RouteSimilarity}(R^-_x, R_y), \text{RouteSimilarity}(R_x, R^-_y))$
end if

**comment:** $R^-_x = \langle r_{x_0}, r_{x_1}, \ldots, r_{x_{n-1}} \rangle$
**comment:** $R^-_y = \langle r_{y_0}, r_{y_1}, \ldots, r_{y_{m-1}} \rangle$

To derive the sequence similarity metric for clustering, we exploit the modified Longest Common Subsequence (LCSS) [Vlachos et al. 2003]. The main idea here is that the match of two routes allows some elements (RoA) to be unmatched within a minimum bounding envelope, but the order of elements in the match must remain. In other words, two tourist movement trajectories $R_x = \langle r_{x_0}, r_{x_1}, \ldots, r_{x_n} \rangle$ and $R_y = \langle r_{y_0}, r_{y_1}, \ldots, r_{y_m} \rangle$ are deemed to be similar, if there exists a long subsequence $R'_x$ of $R_x$ that can be mapped to a long subsequence $R'_y$ of $R_y$. $R'_x$ (or $R'_y$) here does not necessarily consist of consecutive elements from $R_x$ (or $R_y$), but the elements in $R'_x$ (or $R'_y$) must be in the same order as in $R_x$ (or $R_y$) [Das et al. 1997]. Moreover, the minimum bounding envelope allows for some noise and outliers in the tour route matching. The similarity of tourist movement $R_x$ and $R_y$ is then defined as:

$$\text{Sim}(R_x, R_y) = \frac{|R'_x|}{\max(|R_x|, |R_y|)},$$  \hspace{1cm} (6)

where $|R'_x| = |R'_y|$ is the cardinality of the longest common subsequence. The procedures of estimating the similarity of two travel routes are summarized in Algorithm 1. The algorithm can be implemented efficiently via dynamic programming.

After computing the pair-wise similarity of tourist travel routes, we utilize the hierarchical agglomerative clustering to discover groups of similar movement trajectories. We use average link inter-cluster distance to define the distance of two sub-clusters, which measures the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any data of the other cluster [Bishop 2006]. Each resulting cluster is regarded as one potential tour route pattern. Specifically, we define the tour route pattern $\mathcal{G}$ as below.

**Definition 5.** A tour route pattern $\mathcal{G}$ refers to a group of tour paths that are topologically similar in terms of the visit sequences of RoAs.
The confidence score of a tour route pattern $\mathcal{G}$ is estimated by its cluster size and density, specifically as follow:

$$\text{Score}(\mathcal{G}) = \frac{1}{|\mathcal{G}|^2} \sum_{X \in \mathcal{G}} \sum_{Y \in \mathcal{G}} \text{RouteSimilarity}(X, Y) + \alpha|\mathcal{G}|,$$

(7)

where $\alpha$ is a tradeoff factor for the number of tour routes and their similarity. $\alpha$ is set to 0.5 empirically in this study. Intuitively, the tour pattern with high confidence score will represent popular tour routes, as its large number of similar travel routes reflects the choice of the tourist crowd.

4. DATA

Geotagged photos used in this study were downloaded from Flickr, by using its publicly available API. To download photos, the name of a tour destination, such as Paris, London, etc, is fed in as query to retrieve a set of seed photos. Then, the ids of the owners of these seed photos are retrieved. Based on the owners’ id, we download the entire collection of his/her shared photos to ensure the completeness of the generated photo movement trajectory.

4.1 Tourist Movement Trajectory Database

We download photos in four major cities: London, Paris, New York City and San Francisco. In total, we collected $\sim$769k geotagged photos from $\sim$23k Flickr users. We concatenate photos of a photographer into photo trails in a daily basis. We then apply the entropy based mobility measure to filter out photo trails from non-
tourists. The mobility entropy threshold $\varepsilon_{mob}$ classifies a photo trail into tourist or non-tourist ones. The value of $\varepsilon_{mob}$ is determined empirically in a set of experiments. First, we manually labeled a set of photo paths as the ground truth. We display photos pertaining to a travel path and manually decide whether it is tourist or non-tourist based on the visual content of tourism and sightseeing activities. For ambiguous photos, we simply regard them as non-tourist photos. As manual annotation is laborious, we only label 1000 photo paths, of which 446 are tourist and 554 are not. Then we then compute the mobility entropy of each photo trail and use the mobility entropy threshold to classify them into tourist and non-tourist trail. We change the threshold value from 0 to 0.5 to evaluate the parameter sensitivity over the classification accuracy. We observe that when $\varepsilon_{mob} < 0.3$, its sensitivity on accuracy is not obvious. However, when $\varepsilon_{mob} \geq 0.3$, the accuracy drops drastically. This is because the higher threshold misclassified many tourist photo paths as non-tourist ones. Based on the experiments, we set $\varepsilon_{mob}$ to 0.2. Furthermore, the significance test is carried out to ensure the movement trajectories from photo trails to be statistically significant.

The method 1 based on Poisson distribution identifies 8105 person-day travel paths taken by 5032 tourists in total, while the method 2 based on Normal distribution discovered 8047 person-day trips taken by 5010 tourists in total. To evaluate these two methods, we manually label a set of photo paths as the ground truth. We display the photos pertaining to a travel path and manually decide whether it is tourist or non-tourist based on the photos’ visual contents of tourism and sightseeing activities. For ambiguous photos, we simply regard them as non-tourist. As manual annotation is laborious, we sample a subset of 1000 travel paths for labeling. Based on the Table of Sample Size [Krejcie 1970], the evaluation on 1000 travel paths can estimate the accuracy of the total travel paths with a confidence interval of 3% and a confidence degree of 95%, according to the Table of Sample Size [Krejcie 1970]. Method Among the 1000 travel paths, 954 (or 95.4%) of them identified by method 1 and 966 (or 96.6%) of them identified by method 2 are found with obvious tourism and sightseeing characteristics in photos. This means that we have 95% confidence to state that method 1 and 2 can give the accuracy of 95.4% and 96.6% respectively, within the range of ±3%. This demonstrates that the proposed mobility entropy-based method is effective in identifying tourist travel paths. Observation shows that method 1 and 2 yield very similar results. This is because effectively both methods try to filter out photo paths with too few or many visiting places. Some negative photo paths are examined. We observe that photos of these travel paths are geospatially scattered, however, litter or no information can be perceived on touristic attractions and sightseeing activities.

As method 2 gives slightly better result, we will adopt its resulting movement trajectory database for the subsequent process, which consists of 8047 person-day trips taken by 5010 tourists in total. In average, each city has $\sim$2000 person-day trips. This significantly outnumbers the manually collected tourist movement datasets of existing tourist mobility analysis works [Lau 2007; Lew and McKercher 2002; McKercher and Lew 2004], not to mention that the database can be easily augmented by downloading more geotagged photos. Figure 1 and 4 show the
movement trajectories generated from geotagged photos in New York City, San Francisco, Paris and London plotted on Google Earth.

4.2 Regions of Attractions

By taking geotagged photos as input, we discover the regions-of-attractions (RoAs) in a density-based approach, as presented in Section 3.1. In total, we discover 80 RoAs with 18 in London, 19 in Paris, 23 in New York City and 20 in San Francisco. Among them, only 1 out of 80 RoAs is false, which is "San Francisco Pride Parade". This event is misclassified as a RoA, as it gives rise to voluminous photos with strong geospatial pattern. Figure 5 shows the RoAs in New York City and San Francisco.

For comparison purpose, we next apply the mean-shift clustering to identify RoAs, which is the approach used in [Kurashima et al. 2010] and [Lu et al. 2010]. Compared to the DBSCAN clustering, the mean-shift method used give a quite similar list of RoAs. The mean-shift discovers a total of 84 RoAs in the four cities, of which 77 (or 97.4%) RoAs are identical to the ones found by DBSCAN. We attribute this to the fact that a RoA is usually a dense cluster of photos, which can be reliably identified.

5. EXPERIMENTS

5.1 Tourist Traffic Analysis

**Popularity of RoA.** The popularity of a RoA can be estimated by its tourist traffic volume, namely the number of people that have photographed in the region. Table I summarizes the top 3 most visited (most popular) RoAs in the four cities. For each RoA, the percentage of tourists that visit it is also computed. We compare Table I against the list of top 3 attractions in Yahoo!Travel\(^5\) and found that two lists share 42% identical RoAs. The attraction popularity in Yahoo!Travel is estimated based on the feedback scores provided by Yahoo users. This overlap of popular RoAs suggests that despite of different background, people tend to agree on the most popular attractions to some extent. Figure 5 shows the popularity of RoAs in San Francisco and New York City. The height of the bar indicates the volume of tourist visits of the RoA.

**Transition Traffic between RoAs.** We utilize Markov chain model to estimate the transition probability \(P(r_j | r_i)\). \(P(r_j | r_i)\) indicates how tourist traffic moves from one RoA to another. A reasonably high value of \(P(r_j | r_i)\) suggests that RoA \(r_j\) and \(r_i\) are coupled in the way that tourists tend to visit RoA \(r_j\) right after \(r_i\). Figure 6 displays the RoA transitions with probability \(P(r_j | r_i) > 0.2\) in downtown area of Paris. As shown, the coupled RoAs are usually geographically adjacent to each other. Moreover, it is also observed that people tend to prefer certain direction when visiting two coupled RoAs. For example, the transition probability \(P(\text{Eiffel Tower | Arc de Triomphe})\) that tourists move from "Arc de Triomphe" to "Eiffel Tower" is 0.31, while the transition probability \(P(\text{Arc de Triomphe | Eiffel Tower})\) in the opposite direction is only 0.12. This suggests that tourists might share similar

---

\(^5\)http://travel.yahoo.com

The possible reason could be that data similarity results in the similar movement trajectories.
Fig. 5. Popularity of RoAs in downtown San Francisco and New York City. The height of the bar indicates the popularity of its corresponding RoA. For better viewing, please see the original color pdf file.

Table I. Top three most visited RoAs and percentage of tourists in the four cities. SF: San Francisco

<table>
<thead>
<tr>
<th>RoAs</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF</td>
<td></td>
</tr>
<tr>
<td>1. Golden Gate Bridge</td>
<td>27.6</td>
</tr>
<tr>
<td>2. Pier 39</td>
<td>22.9</td>
</tr>
<tr>
<td>3. Union Square</td>
<td>20.3</td>
</tr>
<tr>
<td>New York City</td>
<td></td>
</tr>
<tr>
<td>1. Times Square</td>
<td>35.6</td>
</tr>
<tr>
<td>2. Rockefeller Center</td>
<td>29.3</td>
</tr>
<tr>
<td>3. Brooklyn Bridge</td>
<td>22.9</td>
</tr>
<tr>
<td>Paris</td>
<td></td>
</tr>
<tr>
<td>1. Notre Dame</td>
<td>38.7</td>
</tr>
<tr>
<td>2. Eiffel Tower</td>
<td>31.0</td>
</tr>
<tr>
<td>3. Arc de Triomphe</td>
<td>30.5</td>
</tr>
<tr>
<td>London</td>
<td></td>
</tr>
<tr>
<td>1. London Eye</td>
<td>43.6</td>
</tr>
<tr>
<td>2. Trafalgar Square</td>
<td>34.5</td>
</tr>
<tr>
<td>3. Tower Bridge</td>
<td>29.9</td>
</tr>
</tbody>
</table>

Table II summarizes the centric RoAs in the four cities. Figure 6 shows that "Eiffel Tower" and "Cathedrale Notre Dame" are centric RoAs in Paris, as they receive influx of tourists from several adjacent RoAs. The centric RoA might be determined by several factors, including popularity, geographical location, trans-

Centric RoA. Tourist traffic tends to flow from several RoAs to a central one. We denote this central RoA as centric RoA. Specifically, we define centric RoA as the one with transition probability $P(\text{centric RoA} \mid r_i) > 0.15$ for more than 3 RoA $r_i$. Table II summarizes the centric RoAs in the four cities. Figure 6 shows that "Eiffel Tower" and "Cathedrale Notre Dame" are centric RoAs in Paris, as they receive influx of tourists from several adjacent RoAs. The centric RoA might be determined by several factors, including popularity, geographical location, trans-

ACM Transactions on Intelligent Systems and Technology, Vol. V, No. N, Month 20YY.
Fig. 6. Transitions among RoAs in downtown, Paris, with transition probability \( P(r_j|r_i) > 0.2 \). For better viewing, please see the original color pdf file.

Table II. Centric RoAs in the four cities.

<table>
<thead>
<tr>
<th>City</th>
<th>Centric RoAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>San Francisco</td>
<td>Union Square, Chinatown</td>
</tr>
<tr>
<td>New York City</td>
<td>Time Square, Brooklyn Bridge</td>
</tr>
<tr>
<td>Paris</td>
<td>Eiffel Tower, Cathedrale Notre Dame</td>
</tr>
<tr>
<td>London</td>
<td>London Eye, Trafalgar Square</td>
</tr>
</tbody>
</table>

Table III. Average number of RoA visits per day in the four cities.

<table>
<thead>
<tr>
<th>RoA #</th>
<th>New York City</th>
<th>San Francisco</th>
<th>Paris</th>
<th>London</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.7</td>
<td>3.4</td>
<td>3.6</td>
<td>3.2</td>
</tr>
</tbody>
</table>

In a way, the centric RoA is the place where people congregate and meet each other.

5.2 Topological Analysis of Travel Route

In our database of travel routes \( \{ R \} \) in four cities, the average cardinality of tour routes is \( |R| = 3.5 \). Namely, tourists visit an average of 3.5 RoAs per day. This number is similar to the average visit of 3.7 RoAs per day in the tourism study [Mckercher and Lau 2008]. Table III summarizes the average number of RoA visits per day in the four cities.

Popular Tour Routes. We then perform sequence-clustering presented Section 3.3 to discover popular tour routes, based on the confidence score in Eq. 7. The clusters of travel routes with highest scores tend to comprise many similar tour routes. We identify these clusters as group patterns of tourist travel routes. It reflects the choice of the crowd on popular tour routes. Specifically, we take the centroid route of the top three cluster as representative routes and list them in Table IV for the four cities. From Table IV, two observation draws our attention. First, different popular routes usually visit a distinct set of RoAs. Second, the popular tour routes mostly comprise of a small number (2 or 3) of RoAs. This
Table IV. Top 3 most popular tour routes in the four cities. NYC: New York City, SF: San Francisco

<table>
<thead>
<tr>
<th>Popular tour routes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF</td>
</tr>
<tr>
<td>1. Union Square &gt;&gt; Golden Gate Bridge</td>
</tr>
<tr>
<td>2. AT&amp;T Park &gt;&gt; SFMOMA</td>
</tr>
<tr>
<td>3. Ferry Building &gt;&gt; Alcatraz</td>
</tr>
<tr>
<td>NYC</td>
</tr>
<tr>
<td>2. Brooklyn Bridge &gt;&gt; Empire State Building</td>
</tr>
<tr>
<td>3. Statue of Liberty &gt;&gt; Times Square</td>
</tr>
<tr>
<td>Paris</td>
</tr>
<tr>
<td>1. Pont Alexander &gt;&gt; Arc de Triomphe</td>
</tr>
<tr>
<td>2. Sacre Coeur &gt;&gt; Notre Dame</td>
</tr>
<tr>
<td>3. Centre Georges Pompidou &gt;&gt; Musee d’Orsay</td>
</tr>
<tr>
<td>London</td>
</tr>
<tr>
<td>1. Brick Lane &gt;&gt; Lloyd’s of London</td>
</tr>
<tr>
<td>2. Trafalgar Square &gt;&gt; Coven Garden</td>
</tr>
<tr>
<td>3. London Eye &gt;&gt; Trafalgar Square</td>
</tr>
</tbody>
</table>

observation motivates us to further investigate the number of RoAs in the routes.

Relax-trips v.s. Busy-trips. Based on the number of visited RoAs in travel routes, the tourist daily trip can be classified into two types: busy-trip and relax-trip. Busy-trip refers to a one that covers RoAs more than |R|, while relax-trip is a one that has RoA visits less than |R|. Out of 8038 day trips in total, 3949 (49.1%) trips belong to busy-trips; and 4098 (50.9%) trips are relax-trip. The number of RoA visits in a day trip reflects, in a general sense, the activeness of tourist travel behavior. Tourists of busy-trips tend to seek tour experiences of traveling many varied RoAs, while tourists of relax-trip prefer to have relaxing journey by visiting only a few places. Another observation is that the number of RoAs in a trip is not necessarily associated with its geospatial distance or temporal duration. A relax-trip can have a long geospatial distance, if it visits 2 RoAs far apart, while a busy-trip may be short geospatially, if it visits the RoAs that are close to each other in downtown area. There do exist special cases where a tourist may have hectic schedule in visiting a single large RoA, such as a museum.

After examining the tour routes of relax-trips, we observe that 3557 (90%) out of 3949 relax-trips share the same routes with another trip. This suggests that relax-trips tend to be homogenous and regular. We attribute the homogeneity of relax-trips to 2 reasons. First, a relax-trip has a lesser number of RoAs, and therefore, less variation of RoAs combinations. Second, when people take up a relax-trip of only 2 or 3 RoAs, they tend to visit the most popular RoAs, as shown in Table IV. This preference further reduces the variation of relax-trips.

As busy-trips consist of more RoA visits, the travel itineraries of busy-trips are fairly diverse. Among 3949 tour itineraries in the dataset, only 512 (13%) of them are exactly the same as another. This suggests that tourist travel behaviors become miscellaneous, when people visit many places. This heterogeneity is attributed to many factors, including human nature of individuality, personal preferences, journey duration, topological characteristics of the tour destination etc. Though people travel differently, they do not travel randomly. This is because, in general, tour itinerary is a result of joint effect of sightseeing strategy, transportation con-
Fig. 7. Two groups of similar tour routes in San Francisco. For better viewing, please see the original color pdf file.

We observe that despite of the diversity of tour routes, some groups of tourists do share similar routes, especially when they visit a similar group of RoAs. For example, Figure 7 shows examples of two groups of similar tour routes in San Francisco. The group of tour routes in Figure 7 (a) goes from downtown ("Chinatown"/"Ferry Building") area to "Golden Gate Bridge", by passing "Ghirardelli Square" area. Similarly, Fig 8 depicts a group of similar tour routes in New York City. The direction of all the routes in this group is from "Grand Central Station"/"Empire State Building" area to "Statute Liberty" area. The similar tour routes can be resulted from several factors. First, the tourists in the same group may share similar sightseeing preferences, which makes their routing strategy alike. Second, such travel pattern might be a result of a topological relation, transportation convenience among various RoAs. For example, tourists usually visit "Ellis Island" right after the "Statute of Liberty" in Liberty Island. This is because there exists a convenient ferry service from Liberty Island to Ellis Island and most tour package includes both destinations.

6. CONCLUSION

Understanding tourist travel behavior and mobility patterns is important to many mobile and location-based social and multimedia applications. In this study, we
approach this task by exploiting the socially generated and community-contributed geotagged photos on the Internet. The photos, together with their time- and geo-references, implicitly mark the spatio-temporal movement trajectories of their photographers. We demonstrate that the geotagged photos can provide an efficient, effective and scalable solution to automatic tourist mobility analysis. Specifically, we focus our analysis on two aspects: (1) tourist movement patterns in relation to the regions of attractions (RoA), and (2) topological characteristics of travel routes by different tourists. We first built a statistically reliable tourist movement trajectory database from geotagged photos, by utilizing an entropy-based mobility measure. A list of regions of attraction (RoA) in a tour destination is then built, based on the frequency of tourist visits. We then investigated tourist traffic flow among different RoAs, by exploiting markov chain model to interpret tourists traffic transition. Finally, tourist travel patterns were analyzed by performing a sequence clustering on tour routes. Testing on four major cities, including San Francisco, New York City, Paris and London, demonstrated that the proposed approach can deliver promising results.

Several issues are worthy of further investigation. First, the personal information of photographers, such as nationality, age, gender, etc, can reflect the travel preferences of different communities. The challenge is, however, that only a small percentage of users share their personal information publicly. Second, based on the mined travel patterns and preferences, personalized travel guide can be recommended to cater to the needs of travelers of different background.

REFERENCES


ACM Transactions on Intelligent Systems and Technology, Vol. V, No. N, Month 20YY.


