Noname manuscript No.
(will be inserted by the editor)

Travel Topic Analysis: A Mutually Reinforcing Method for Geo-tagged Photos

Ngai Meng Kou · Leong Hou U · Yiyang Yang · Zhiguo Gong

Received: date / Accepted: date

Abstract Sharing personal activities on social networks is very popular nowadays, where the activities include updating status, uploading dining photos, sharing video clips, etc. Finding travel interests hidden in these vast social activities is an interesting but challenging problem. In this work, we attempt to discover travel interests based on the spatial and temporal information of geo-tagged photos. Obviously the visit sequence of a traveler can be approximately captured by her shared photos based on the timestamps and geo-locations. To extract underlying travel topics from abundant visit sequences, we study a novel mixture model to estimate the visiting probability of regions of attractions (ROAs). Such travel topics can be used in different applications, such as advertisements, promotion strategies, and city planning. To enhance the estimation result, we propose a mutual reinforcement framework to improve the quality of ROAs. Finally, we thoroughly evaluate and demonstrate our findings by the photo sharing activities collected from Flickr$^{\text{TM}}$.

Keywords Web Images · Travel Analysis · Regions of Attraction · Mixture Models

1 Introduction

With the popularity of GPS embedded devices, huge amount of geo-tagged data is produced in recent years and makes it possible to mine out travel related knowledge. This area has seen a significant increase in attention over the past decade, where the mining problems include pointing the next visiting location [6,26], planning a tour route [26,39], and suggesting interesting landmarks [24,28,34,36,44]. All these work aim to recommend or suggest trip knowledge for individuals, thus need to know users' personal information, such

Ngai Meng Kou, Leong Hou U, Yiyang Yang and Zhiguo Gong
Department of Computer and Information Science, University of Macau, Macau
E-mail: {yb27406,ryanlhu,ya97405,fstzgg}@umac.mo

as their profiles [6,24], historical locations [6,26,34,36,44], current location [26, 39,44], or their preferences [28]. However, such personal information may not be available due to privacy concerns [5,11] or cold-start issues (e.g., newly registered users). In this work, instead of giving trip recommendations for each individual user, we study how to summarize popular travel topics for a city by analyzing historical geo-tagged data of travelers. We argue that such knowledge is important for travel agencies to plan their topic-oriented tours, such as culture-oriented tours, shopping-oriented tours, or kids-oriented tours. In fact, the results are also important for the tourism department of a government to learn its main travel topics, thus well deploy its facilities in order to promote its tourism industry.

Typically there are two common approaches to summarize the main travel topics of a city: (1) expert-based, or (2) data-based. In the first approach, tourism experts are invited to subjectively classify the landmarks of the city into several categories. However, this method is not suitable to tour recommendations since a travel is not only decided by landmark categories but also other constraints (e.g., spatial and temporal). Thus we tackle this problem with the second approach (i.e., data-based) in this work. To our understanding, geo-tagged data can reveal the 'footprints' and 'behaviors' of travelers, thus can well extract the travel topics of a city.

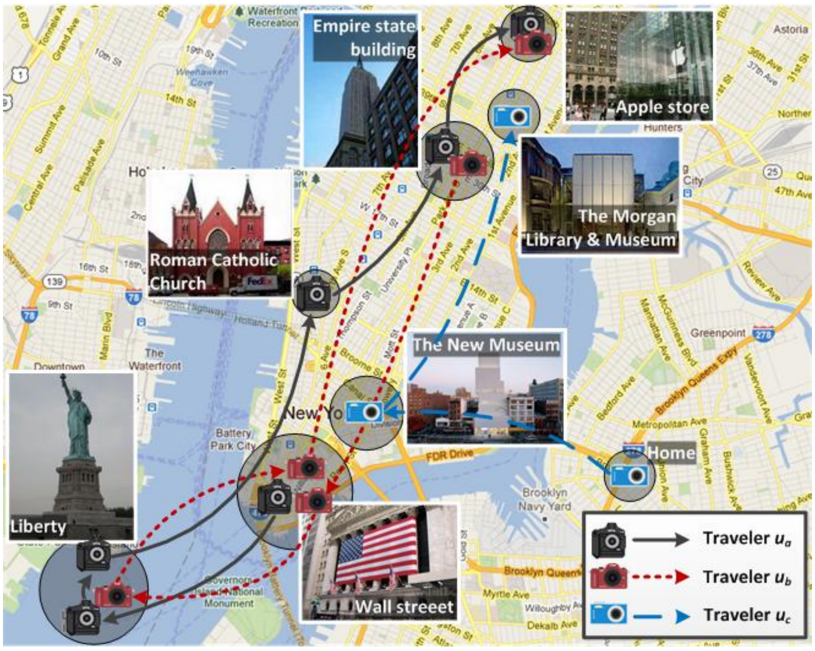


Fig. 1 Tour trajectories in NYC

We illustrate our idea by the following example. Suppose that we collect a set of geo-tagged photos from three travelers u_a, u_b,and u_c at NYC. Figure 1 plots their tour sequences based on the taken time and geo-location of the photos, where a camera icon represents a set of photos taken in a region of attraction and an arc represents a transit from one region to another region. For instance, traveler u_a's tour starts at `Wall Street` and subsequently passes through `Statue of Liberty`, `Roman Catholic Church`, and `Empire State Building`. Finally, u_a ends her trip at `Apple Store`. From the high co-occurrence of the visited regions (four common regions), travelers u_a and u_b should have some common travel interests. On the other hand, the travel interests of u_c should be different from the others since she only visits museums in her trip.

In practice, the geo-tagged data can be extracted from traveler activities in several ways, such as metro card records [30], mobile phone signals [2], etc. However, most of them are not public available due to privacy concerns. Fortunately, there are an increasing number of people sharing their geo-tagged data over public social network services, such as Flickr$^{\text{TM}}$, Picasa$^{\text{TM}}$, and Photobucket$^{\text{TM}}$, which enables to analyze the user behaviors using their geo-tagged data.

Intuitively, the trip (i.e., regions of attraction (ROAs)) of a traveler is often driven by her interests. If a group of travelers visits similar ROAs, we simply conclude that they have common interests (i.e., a travel topic). For example, a group of travelers may prefer to visit natural landscapes such as forests, parks, mountains and rivers, while another group of travelers may prefer to take photos at famous landmarks such as `Times Square` and `United Nations`.

Suppose $r_1, r_2, ..., r_N$ are the ROAs of a city, a travel topic $\overrightarrow{\theta}$ can be viewed a vector of probability, $(p(r_1|\overrightarrow{\theta}), p(r_2|\overrightarrow{\theta}), ..., p(r_N|\overrightarrow{\theta}))$, where $p(r_i|\overrightarrow{\theta})$ indicates the conditional probability to visit an ROA r_i under topic $\overrightarrow{\theta}$. Our **first mission** in this work is to extract a set of travel topics based on the information of geo-tagged data. We add a note that our problem is different from trajectory mining work [1,13,18,20,27] where their analysis are mainly based on the visiting order of ROAs. However, travelers who share similar travel interests may have completely different routes to visit the same set of ROAs so that the trajectory based solutions are not applicable to our problem.

Obviously the quality of a travel topic $\overrightarrow{\theta}$ is highly related to the quality of ROAs (i.e., $\mathbb{R} = \{r_1, ..., r_N\}$). However, the shape and size of ROAs cannot be defined easily. For instance, the region of `Apple Store` may only be a small area while the `Wall Street` is a large region since travelers often visit the entire district instead of any single area of it. To define a good quality set of ROAs, a straightforward solution is to select $\mathbb{R}$ from the landmarks of a city. However, this approach may omit some travel topics. As an example, travelers may visit some regions out of landmarks. Another approach to define $\mathbb{R}$ is to cluster the data [21,22,35,40] based on their geographical locations and data density using standard clustering approaches. Given the clustering result as an input, our **second mission** in this work is to further enhance the identification of ROAs using a mutual reinforcement framework. This idea has the same intuition of [43,44] where ROAs and travel trajectories are mutually reinforced.

We summarize our main contributions in this manuscript as follows.

– We transform the travel topic discovery problem into a multinomial mixture probability model such that the travel topics can be estimated by Expectation-Maximization. To the best of our knowledge, we are the first work to adopt this model for travel interest analysis.
– We propose a mutual reinforcement framework that iteratively refines the travel topics and regions of attraction. Our experiments demonstrate that the reinforcement process improves the overall identification quality.

The rest of the paper is organized as follows. We provide the formal definitions and adopt a probabilistic mixture model to estimate the travel topics in Section 2. The solution of travel topics extraction is discussed thoroughly in Section 3. In Section 4, we experimentally evaluate our methods using real data collected from Flickr^{TM}. Section 5 discusses related work. Finally, Section 6 concludes the paper.

2 Preliminaries

In this section, we formally define the fundamental elements of our travel topic discovery problem, such as tour trajectory, region of attraction, and travel topic. Furthermore, we discuss how the discovery problem can be viewed as an optimization problem such that the best travel topics can be estimated by Expectation-Maximization algorithm (EM).

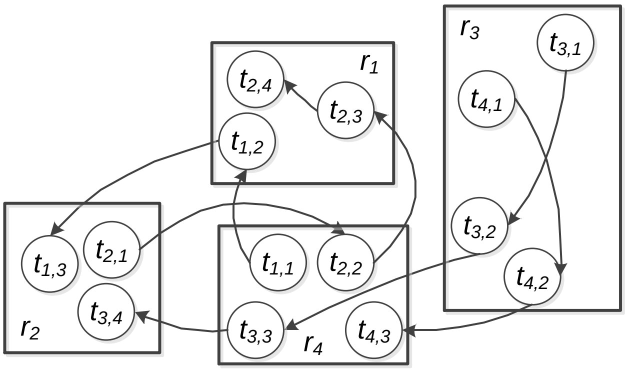


Fig. 2 Travel trajectories, regions of attraction, travel topics, and background topics

We first define an ROA and a tour trajectory as follows, where these two concepts are thoroughly used in our solution.

Definition 1 (Region of attraction, ROA) A region of attraction, r, is the convex polygon of a set of geo-tagged photos.

Definition 2 (Tour trajectory) A trajectory, $t_i = \{t_{i,1}, ..., t_{i,n}\} \in \mathbb{T}$, is a bag of geo-locations of photos taken by the same user with corresponding taken-time, where $t_{i,j}$ donates the j-th geo-location in t_i.

Given a geo-tagged photo dataset, a tour trajectory t_i (see Def. 2) can be extracted from the photos taken by a traveler. According to the taken time of the photos, we can construct a tour trajectory $t_i = \{t_{i,1}, ..., t_{i,n}\}$ where $t_{i,j}$ indicates the geo-location of the j-th taken photo in t_i. Typically, a tour trajectory $t_i = \{t_{i,1}, ..., t_{i,n}\}$ should be decomposed into $\{t_{i,1}, ..., t_{i,j}\}$ and $\{t_{i,j+1}, ..., t_{i,n}\}$ if the time interval between $t_{i,j}$ and $t_{i,j+1}$ is too long (48 hours) since they are not likely in the same tour.

As introduced in Section 1, the set of *regions of attraction* (ROAs) (see Def. 1) visited by a traveler is driven by her travel interests. Figure 2 illustrates 4 different tour trajectories where these trajectories may reveal their travel interests. For example, the travel interests of t_4 are unlikely similar to t_1 and t_2 as they visit different set of ROAs, $\{r_3, r_4\}$ vs $\{r_1, r_2, r_4\}$.

Given the trajectories and ROAs, our mission is to summarize a group of *travel topics* that reveal the travel interests of travelers. In this work, a travel topic is represented by the probability of visiting each ROA. We formally define a travel topic $\overrightarrow{\theta}$ as follows.

Definition 3 (Travel topic, $\overrightarrow{\theta}$) A travel topic $\overrightarrow{\theta}$ represents the probability distribution over ROAs denoted as $(p(r_1|\overrightarrow{\theta}), p(r_2|\overrightarrow{\theta}), ..., p(r_N|\overrightarrow{\theta}))$, where $p(r_i|\overrightarrow{\theta})$ is the probability of visiting ROA r_i under topic $\overrightarrow{\theta}$, subject to $\sum_{i=1}^{|\mathbb{R}|} p(r_i|\overrightarrow{\theta})$=1.

For the ease of discussion, we define a mapping function $\mathbb{R}(t_{i,j})$ that indicates the corresponding ROA containing the j-th location of tour t_i, e.g., $\mathbb{R}(t_{2,3}) = r_1$ in Figure 2. Accordingly, we can replace $p(r|\overrightarrow{\theta})$ by $p(\mathbb{R}(t_{i,j})|\overrightarrow{\theta})$ if $r = \mathbb{R}(t_{i,j})$. Inspired by the topic discovery problems [4,16,19,42], given a topic $\overrightarrow{\theta}$, the probability of independently selecting each ROA in a tour t_i can be defined as follows.

$$p(t_i|\overrightarrow{\theta}) = \prod_{j=1}^{|t_i|} p(\mathbb{R}(t_{i,j})|\overrightarrow{\theta}) \tag{1}$$

Given a mixture of K travel topics $\overrightarrow{\theta}_{1:K} = \{\overrightarrow{\theta}_1, ..., \overrightarrow{\theta}_K\}$ and the corresponding mixture coefficients $\overrightarrow{\pi}_{t_i} = \{\pi_{t_i,1}, ..., \pi_{t_i,K}\}$ of a tour t_i, the probability of t_i given the mixture of travel topics $\overrightarrow{\theta}_{1:K}$ can be defined as

$$p(t_i|\overrightarrow{\theta}_{1:K}) = \prod_{j=1}^{|t_i|} \sum_{k=1}^{K} \pi_{t_i,k} \cdot p(\mathbb{R}(t_{i,j})|\overrightarrow{\theta}_k) \tag{2}$$

where $\sum_{k=1}^{K} \pi_{t_i,k} = 1$ and $\forall_{1 \leq k \leq K}\ \pi_{t_i,k} \geq 0$.

To estimate the mixture of K travel topics $\overrightarrow{\theta}_{1:K}$ for a tour trajectory set $\mathbb{T}$, we can apply Expectation-Maximization algorithms (EM) [12,16] to find the maximum likelihood estimation of the variables (i.e., $\overrightarrow{\theta}_{1:K}$ and $\overrightarrow{\pi}_{t_1}, ..., \overrightarrow{\pi}_{t_{|\mathbb{T}|}}$).

To further enhance the estimation quality of EM, we should remove **background topic** $\overrightarrow{\beta}$ from the estimation process. In this work, a background is an ROA that is visited by abundant travelers (i.e., overly popular), e.g., `Statue of Liberty` in NYC. Such popular ROAs likely appear in every travel topics such that they are not helpful to distinguish travel topics. A background topic $\overrightarrow{\beta}$ gives the popularity degree of each ROA which is formally defined as follows.

Definition 4 (Background Topic, $\overrightarrow{\beta}$) The probability of visiting an ROA r_c under the background $\overrightarrow{\beta}$ is described as:

$$p(r_c|\overrightarrow{\beta}) = \frac{\sum_{i=1}^{|\mathbb{T}|} c(t_i, r_c)}{\sum_{i=1}^{|\mathbb{T}|} \sum_{j=1}^{|\mathbb{R}|} c(t_i, r_j)} \tag{3}$$

where $c(t_i, r_j) = 1$ if and only if tour t_i passes ROA r_j; otherwise, $c(t_i, r_j) = 0$.

In the example of Figure 2, ROA r_4 has the highest background probability, where $p(r_4|\overrightarrow{\beta}) = \frac{4}{11}$, as all four tours pass r_4. To leverage the effect of the background topic $\overrightarrow{\beta}$ in the identification process, we replace $p(t_i|\overrightarrow{\theta}_{1:K})$ in equation 2 with a linear combination of $p(t_i|\overrightarrow{\theta}_{1:K})$ and the background probability of the ROAs:

$$p(t_i|\overrightarrow{\theta}_{1:K}; \overrightarrow{\beta}, b) = \prod_{j=1}^{|t_i|}((1-b)\sum_{k=1}^{K} \pi_{t_i,k} \cdot p(\mathbb{R}(t_{i,j})|\overrightarrow{\theta}_k) + b \cdot p(\mathbb{R}(t_{i,j})|\overrightarrow{\beta})) \tag{4}$$

where b is a tunable mixture coefficient for leveraging the effect of background topic $\overrightarrow{\beta}$ in this work. An ROA can be generated from a travel topic or the background topic when we involve $\overrightarrow{\beta}$ into the model. According to the probability in $\overrightarrow{\beta}$, it is more likely to generate these ROAs from $\overrightarrow{\beta}$ instead of other travel topics which can significantly avoid hottest ROAs dominate all travel topics.

For clarity, we use an example to explain how Equation 4 reduces the effect of over-popular landmarks. Suppose r_a is an over-popular landmark (e.g., 10% photos were taken in r_a), r_a is likely ranked as the most important ROA in every travel topic when these topics are estimated by Equation 2. When taking the background topic $\overrightarrow{\beta}$ into consideration (cf. Equation 4), r_a is likely generated from the background topic $\overrightarrow{\beta}$ (given) instead of other travel topics $\overrightarrow{\theta}_{1:K}$ (estimated). In other words, r_a is included into a travel topic $\overrightarrow{\theta}$ only if r_a is strongly related to $\overrightarrow{\theta}$, i.e., a certain number of similar trajectories ($> 10\%$) passes r_a.

Finding the estimation of the variables (i.e., $\overrightarrow{\theta}_{1:K}$ and $\overrightarrow{\pi}_{t_1}, ..., \overrightarrow{\pi}_{t_{|\mathbb{T}|}}$) in Equation (4) can be viewed as the following optimization problem as follows.

$$\underset{\overrightarrow{\theta}_{1:K}, \overrightarrow{\pi}_{t_1}, ..., \overrightarrow{\pi}_{t_{|\mathbb{T}|}}}{\operatorname{argmax}} \prod_{t_i \in \mathbb{T}} p(t_i | \overrightarrow{\theta}_{1:K}; \overrightarrow{\beta}, b) \tag{5}$$

Discussion. Similar to other topic discovery problems [4,16,19,42], the optimization (Equation (5)) can be solved by the EM algorithms. However, discovering travel topics is more challenging since the set of ROAs $\mathbb{R}$ and the number of travel topics K are enormously varying from city to city. As discussed above, the quality of travel topics $\overrightarrow{\theta}_{1:K}$ is sensitive to the size and shape of ROAs. Suppose that there are 100 ROAs in $\mathbb{R}$ but 4 of them are very large which almost cover the entire map. The discovery result should be poor with such a skew data distribution. Although $\mathbb{R}$ and $\overrightarrow{\theta}_{1:K}$ can be defined by clustering [40,44] and incremental topic discovery [3,19] respectively, there is no prior work to study how to optimize the quality of these two techniques in a unified framework. In this work, we propose a novel solution, named *Mutual Reinforcing Travel Topic Discovery* (MRTD), that iteratively refines $\overrightarrow{\theta}_{1:K}$ and $\mathbb{R}$ until their identification quality is converged.

3 Mutually Reinforcing Travel Topic Discovery

3.1 Incremental Travel Topic Construction

In this section, we discuss how to construct travel topics incrementally based on a set of pre-defined ROAs $\mathbb{R}$. The motivation of this incremental process is providing an automatic way to decide a suitable number of topics for different cities. When constructing a new travel topic, we should consider 2 conditions, (1) if the new travel topic is a common interest shared by a group of travelers (i.e., level of confidence) and (2) if the new travel topic is different from the other travel topics in the result (i.e., level of duplication). Based on these concerns, we propose a 3-steps framework to construct the travel topics.

Constructing a travel topic. Given a trajectory t_i, we can construct a new travel topic $\overrightarrow{\theta}_{t_i}$ based on a set of similar trajectories, where the set of similar trajectories can be formed by the following equation.

$$\mathbb{T}^{t_i}_{sim} = \{t_j | t_j \in \mathbb{T}, sim(t_i, t_j) > \lambda_{sim}\} \tag{6}$$

where λ_{sim} is the similarity threshold and sim is defined as:

$$sim(t_i, t_j) = \frac{\sum\limits_{r \in \mathbb{R}} pass(t_i, t_j, r)}{\min(|t_i|, |t_j|)} \tag{7}$$

where $pass(t_i, t_j, r)$ return 1 if both t_i and t_j pass through r, and 0 otherwise, e.g., $pass(t_1, t_2, r_1)$=1 in the example of Figure 2. While the numerator can be viewed as the common interest of two trajectories, the denominator is a normalization factor. According to our experimental study, the definition of $\mathbb{T}^{t_i}_{sim}$ may be too strict in practice as the length of the trajectories is largely varying. This turns out that very few trajectories can fulfill the threshold due to the length variance. Thereby, we revise the definition of $\mathbb{T}^{t_i}_{sim}$ such that it considers not only the trajectory in its entirety but also its subsequences of length larger than 4. If there are more than one subsequences fulfilling the threshold, we only pick the longest subsequence into $\mathbb{T}^{t_i}_{sim}$. For instance, given $t_i = \{r_4, r_5, r_6, r_7, r_8, r_9\}$ and $\lambda_{sim} = 0.5$, the subsequences of $t_j(= \{r_1, r_2, r_3, r_4, r_5, r_6\})$ of length larger than 4 are $\{r_1, r_2, r_3, r_4, r_5, r_6\}$, $\{r_1, r_2, r_3, r_4, r_5\}$, and $\{r_2, r_3, r_4, r_5, r_6\}$. We only pick $\{r_1, r_2, r_3, r_4, r_5, r_6\}$ into $\mathbb{T}^{t_i}_{sim}$ since it is the longest one.

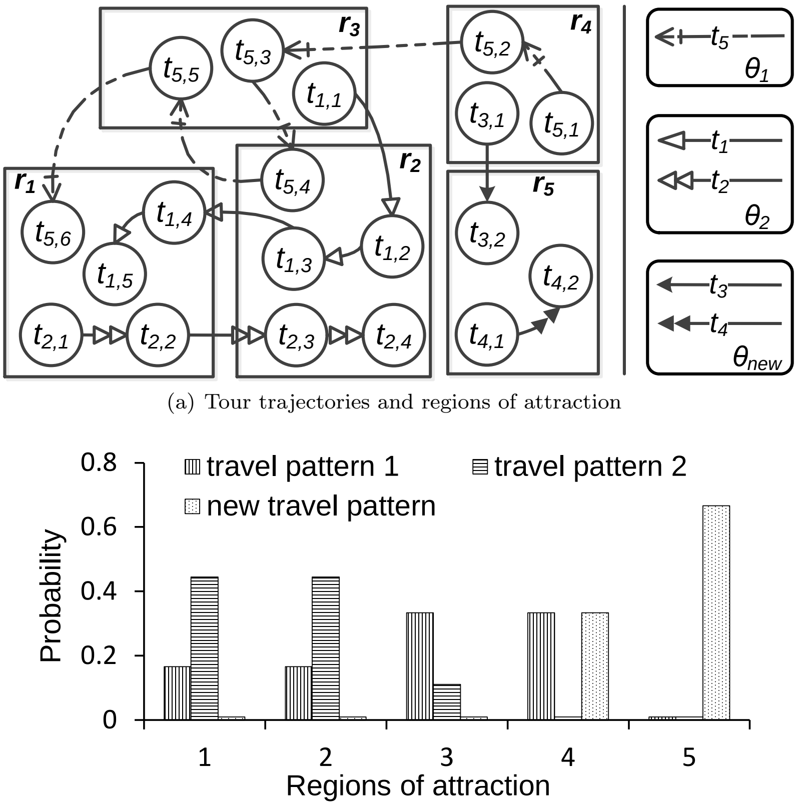
Given $\mathbb{T}^{t_i}_{sim}$, according to Equation (4) and (5), we can construct a travel topic based on $\mathbb{T}^{t_i}_{sim}$ using the following equation:

$$\begin{aligned}\overrightarrow{\theta} &= \underset{\overrightarrow{\theta}}{\operatorname{argmax}}\ p(\mathbb{T}^{t_i}_{sim}|\overrightarrow{\theta}) = \underset{\overrightarrow{\theta}}{\operatorname{argmax}} \prod_{t_i \in \mathbb{T}^{t_i}_{sim}} p(t_i|\overrightarrow{\theta}) \\ &= \underset{\overrightarrow{\theta}}{\operatorname{argmax}} \prod_{t_i \in \mathbb{T}^{t_i}_{sim}} \prod_{i=1}^{|t_i|} ((1-b)p(\mathbb{R}(t_{i,j})|\overrightarrow{\theta}) + bp(\mathbb{R}(t_{i,j})|\overrightarrow{\beta}))\end{aligned} \tag{8}$$

Note that Equation (8) is slightly different from Equation (4). The coefficients π is removed since there is only one travel topic ($\overrightarrow{\theta}$ instead of $\overrightarrow{\theta}_{1:K}$) under consideration.

Incremental construction. So far we only discuss how to form independent travel topics by Equation 8. Note that a travel topic is worth to construct only when it is *original* to the existing travel topics. Figure 3 illustrates a toy example when adding a new topic. Suppose that there are five tour trajectories in the dataset and two travel topics $\overrightarrow{\theta}_1$ and $\overrightarrow{\theta}_2$ have been constructed by trajectories $\{t_5\}$ and $\{t_1, t_2\}$, respectively. Their corresponding distributions are plotted in Figure 3(b). The first travel topic $\overrightarrow{\theta}_1$ indicates that some travelers are likely to visit ROAs r_3 and r_4 and the second travel topic $\overrightarrow{\theta}_2$ prefers to visit r_1 and r_2. Suppose that we are trying to construct a new travel topic $\overrightarrow{\theta}_{new}$ (by Equation (8)) using trajectory t_3 and the set of similar trajectories $\mathbb{T}^{t_3}_{sim} = \{t_3, t_4\}$ [1]. According to the distribution in Figure 3(b), $\overrightarrow{\theta}_{new}$ is original since it indicates a set of travelers treats r_5 as the most popular ROA which is different from $\overrightarrow{\theta}_1$ and $\overrightarrow{\theta}_2$.

[1] It should be noted that all trajectories (including t_1,t_2, and t_5) are taken into consideration. This is intuitive to the real world scenario, where a traveler may have multiple travel interests so that her tour can be partitioned into several travel topics.



(a) Tour trajectories and regions of attraction

(b) Travel topic distribution

Fig. 3 Example of originality checking

To assess the originality of a new travel topic, we introduce a function (first introduced in [19]), $ori(t_i, \overrightarrow{\theta}_{new}, \overrightarrow{\theta}_1, ..., \overrightarrow{\theta}_K)$, that calculates the probability divergence of generating one tour t_i by including or excluding the new travel topic $\overrightarrow{\theta}_{new}$. Formally, the originality function is defined as:

$$ori(t_i|\overrightarrow{\theta}_{new}, \overrightarrow{\theta}_1, ..., \overrightarrow{\theta}_E) = \log \frac{p(t_i|\overrightarrow{\theta}_{new}, \overrightarrow{\theta}_1, ..., \overrightarrow{\theta}_E)}{p(t_i|\overrightarrow{\theta}_1, ..., \overrightarrow{\theta}_E)} \tag{9}$$

where $\overrightarrow{\theta}_1, ..., \overrightarrow{\theta}_E$ are the existing travel topics. The value returned by function $ori(\cdots)$ can be viewed as the gain of generating tour t_i by introducing $\overrightarrow{\theta}_{new}$ in the travel topics. It is high only when $p(t_i|\overrightarrow{\theta}_{new}, \overrightarrow{\theta}_1, ..., \overrightarrow{\theta}_E) \gg p(t_i|\overrightarrow{\theta}_1,$ $..., \overrightarrow{\theta}_E)$. In this work, a travel topic is constructed only when the originality of every trajectory in $\mathbb{T}^{t_i}_{sim}$ fulfills a threshold, λ_{ori}.

Smoothing. In the above discussions, we have introduced how to discover

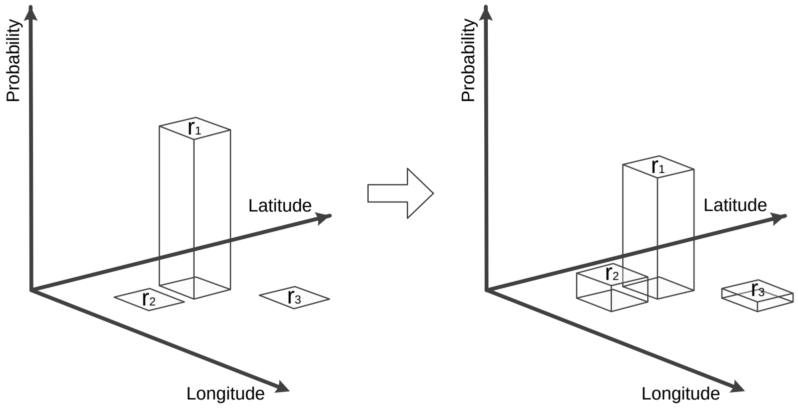


Fig. 4 Example of Smoothing

travel topics with an incremental mixture probability framework, where the quality of the travel topics highly depends on the photo points of tour trajectories. In practice, those points (photos) are sparse comparing with the footprint of travelers. In other words, they probably visit some areas that are close to their photo locations. In order to take these areas into consideration, we perform a smoothing technique which propagates the probability of a region to its nearby regions.

Figure 4 demonstrates an example to smooth a travel topic over 3 ROAs. In the original travel topic $\overrightarrow{\theta}$, travelers visit only r_1; however, it may overlook surrounding areas (e.g., r_2 and r_3). After smoothing, it propagates some probability to surrounding ROAs, r_2 and r_3. Note that r_1 propagates more probability to r_2 than to r_3 since r_2 is closer to r_1 than r_3.

In this work, we use an 2D Gaussian smoothing technique. We assume latitude and longitude are independent, with the same variance (i.e., the covariance of latitude and longitude is zero). In particular, the smoothing function for a travel topic $\overrightarrow{\theta}_s$ is given as follows.

$$p(r_i|\overrightarrow{\theta}_s) \propto \sum_{j=1}^{|\mathbb{R}|} \frac{1}{2\pi\sigma^2} e^{-\frac{D_E(r_i,r_j)}{2\sigma^2}} p(r_j|\overrightarrow{\theta}) \tag{10}$$

where $D_E(r_i, r_j)$ is the minimum Euclidean distance of r_i and r_j and it is defined as

$$D_E(r_i, r_j) = \min\{D_E(t_{m,n}, t_{m',n'})|\forall_{t_{m,n},t_{m',n'}} \mathbb{R}(t_{m,n}) = r_i \wedge \mathbb{R}(t_{m',n'}) = r_j\} \tag{11}$$

Construction algorithm. Algorithm 1 gives the pseudo code of the incremental construction framework. For each trajectory t_i, we first find a set of similar trajectories $\mathbb{T}^{t_i}_{sim}$ with Equation (6) and construct the travel topic $\overrightarrow{\theta}_{t_i}$ with Equation (8). To improve the quality of $\overrightarrow{\theta}_{t_i}$, it is then refined by the smoothing technique. After constructing all independent travel topics, we pick a travel topic of t_i, whose originality value $p(t_i|\overrightarrow{\theta}_{t_i})$ is the largest, into the result set for avoiding cold start. Subsequently, we evaluate each travel topic by their *originality*. A travel topic $\overrightarrow{\theta}_{t_i}$ is included in the result set $\overrightarrow{\theta}_{1:K}$ only if its originality of $\overrightarrow{\theta}_{t_i}$ is larger than a given threshold λ_{ori} (line 9). The construction is terminated when there is no more original travel topic in $\mathbb{T}$.

Algorithm 1 Incremental travel topic construction

```
    Input: Tour trajectory set $\mathbb{T}$ and a set of ROAs $\mathbb{R}$
    Output: A set of travel topics $\overrightarrow{\theta}_{1:K}$
    Algorithm $constructTopics(\mathbb{T}, \mathbb{R})$
 1: for $t_i \in \mathbb{T}$ do                                    ▷ construct independent $\overrightarrow{\theta}_i$
 2:     form similar trajectories $\mathbb{T}^{t_i}_{sim}$ by Equation (6)
 3:     construct $\overrightarrow{\theta}_{t_i}$ of $\mathbb{T}^{t_i}_{sim}$ by Equation (8)
 4:     smooth $\overrightarrow{\theta}_{t_i}$ by Equation (10)
 5: $\overrightarrow{\theta}_{1:K} := \{\overrightarrow{\theta}_{t_i}\}$; $K := 1$;         ▷ Pick $t_i$ who has the largest $p(t_i|\overrightarrow{\theta}_{t_i})$
 6: sort $\mathbb{T}$ by $ori(t_i, \overrightarrow{\theta}_{t_i}, \overrightarrow{\theta}_1, ..., \overrightarrow{\theta}_K)$ by Equation (9)
 7: for $t_i \in \mathbb{T}$ do                  ▷ descending order to $ori(t_i, \overrightarrow{\theta}_{t_i}, \overrightarrow{\theta}_1, ..., \overrightarrow{\theta}_K)$
 8:     for all $t_j \in \mathbb{T}^{t_i}_{sim}$ do
 9:         if $ori(t_j, \overrightarrow{\theta}_{t_i}, \overrightarrow{\theta}_1, ..., \overrightarrow{\theta}_K) < \lambda_{ori}$ then          ▷ originality checking
10:             break
11:     if the loop is not broken then
12:         $\overrightarrow{\theta}_{1:K} := \overrightarrow{\theta}_{1:K} \cup \{\overrightarrow{\theta}_{t_i}\}$; $K := K + 1$;        ▷ $\overrightarrow{\theta}_{t_i}$ is a result
13:         goto line 6
14: return $\overrightarrow{\theta}_{1:K}$
```

3.2 Regions of attraction refinement

Instead of manually defined ROAs, we recommend to identify ROAs by clustering techniques [40,44,45] as the clustering result should better reflect the footprint of travelers. However, these clustering techniques only take into account of distance and density information in their processing. In this work we introduce a refinement approach which polishes ROAs by additionally considering other closeness metrics.

Intuitively, two neighbor ROAs should be merged if they attract similar set of tourists. The similarity can be assessed by the divergence closeness of their travel topic associations. By pairing up with other closeness measures, we propose a novel refinement framework to polish ROAs for travel topic analysis. To the best of our knowledge, we are the first work to identify ROAs based on the result from a mixture model (i.e., travel topics). In the following, we first introduce three closeness measures of ROAs and then introduce our refinement algorithm.

Travel topic distribution closeness. A travel topic is described as the probability distribution over different ROAs (as shown in Figure 3(b)). By using Bayes' formula, we have:

$$p(\overrightarrow{\theta}_j|r_i) = \frac{p(r_i|\overrightarrow{\theta}_j)p(\overrightarrow{\theta}_j)}{\sum\limits_{k=1}^{K} p(r_i, \overrightarrow{\theta}_k)} = \frac{\pi_j p(r_i|\overrightarrow{\theta}_j)}{\sum\limits_{k=1}^{K} \pi_k p(r_i|\overrightarrow{\theta}_k)} \tag{12}$$

where $\pi_{1:k}$ are the weights of travel topics in the mixture model which can be derived from Equation (5) when $\overrightarrow{\theta}_{1:K}$ is given.

Accordingly, we can measure the closeness of two ROAs by computing the divergence of their travel topic associations. In this work, the divergence is calculated by JS-divergence (Jensen-Shannon divergence) [29]. For the sake of presentation, we simply set $\alpha_{i,j} = p(\overrightarrow{\theta}_j|r_i)$ and $\overrightarrow{\alpha_i} = \{\alpha_{i,1}, \ldots, \alpha_{i,K}\}$ where $\overrightarrow{\alpha_i}$ can be regarded as the association of r_i to the entire topic set $\overrightarrow{\theta}_{1:K}$. The JS-divergence of two associations, $\overrightarrow{\alpha_i}$ and $\overrightarrow{\alpha_j}$, is defined as:

$$D_{JS}(r_i, r_j) = D_{JS}(\overrightarrow{\alpha_i}||\overrightarrow{\alpha_j}) = \frac{1}{2}D_{KL}(\overrightarrow{\alpha_i}||\overrightarrow{\alpha_m}) + \frac{1}{2}D_{KL}(\overrightarrow{\alpha_j}||\overrightarrow{\alpha_m}) \tag{13}$$

where D_{KL} represents a KL-divergence (Kullback–Leibler divergence) [25] and $\overrightarrow{\alpha_m} = \frac{1}{2}(\overrightarrow{\alpha_i} + \overrightarrow{\alpha_j})$. For discrete variable, $D_{KL}(\overrightarrow{\alpha_i}||\overrightarrow{\alpha_j})$ is defined as

$$D_{KL}(\overrightarrow{\alpha_i}||\overrightarrow{\alpha_j}) = \sum_{k=1}^{K} \alpha_{i,k} log \frac{\alpha_{i,k}}{\alpha_{j,k}} \tag{14}$$

According to Equation (14), an ROA is divergence close to another ROA if they are co-visited by a large set of tour trajectories. For instance, in Figure 3(a), r_1 and r_2 are divergence close since their distribution over different topics are resembling in $\overrightarrow{\theta}_1$ and $\overrightarrow{\theta}_2$. Obviously, r_1 and r_2 should be merged into one ROA since they are adjacent in spatial and co-visited by many tour trajectories.

Spatial closeness. In some cases, we should not merge two divergence close ROAs since they are not close in spatial space. For instance, `Empire State Building` and `Statue of Liberty` are co-visited by a large amount of travelers but we should not merge them into one ROA because they are geographically apart.

In this work, we assess the spatial closeness of two ROAs using Voronoi diagram and Euclidean distance. We call an ROA r_i spatially close to another ROA r_j if and only if r_i is a neighbor of r_j in Voronoi diagram and their minimum Euclidean distance (see Equation 11) is smaller than a threshold λ_E. The set of ROAs fulfilling the above constraints with r_i is denoted as $V(r_i, \lambda_E)$. Figure 5 gives an example to explain the spatial closeness, where

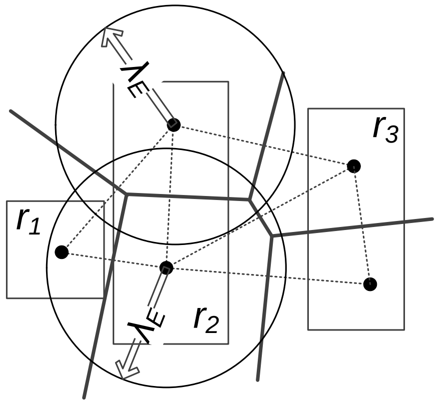


Fig. 5 Spatial closeness

the Voronoi diagram is constructed based on the objects of each ROA. We say that r_3 is a Voronoi neighbor of r_2 since there is at least one Voronoi cell of r_3 being a neighbor of r_2. However, they are not spatially close since their minimum distance $D_E(r_2, r_3)$ is larger than the threshold λ_E.

Sequence closeness. In some situations (e.g., two ROAs on the different sides of river without a bridge nearby), we may merge two ROAs incorrectly based on only divergence and spatial closeness. Therefore we use sequence closeness to address such problems. We call r_1 and r_2 sequentially close, $D_S(r_1, r_2) = \mathit{true}$, if r_1 and r_2 are consecutively visited in at least one of the trajectories. Formally, $D_S(r_i, r_j) = \mathit{true}$ if and only if $\exists t_{m,n} \in \mathbb{T}$, $\mathbb{R}(t_{m,n}) = r_i \wedge \mathbb{R}(t_{m,n+1}) = r_j$.

Refinement algorithm. We use a depth first approach to polish ROAs based on the closeness metrics. At the beginning of an iteration, we pick the most popular ROA r_i from $\mathbb{R}$, where the popularity is based on the number of tours passing it. Then, we assign r_i into a candidate set $\mathbb{C}$. For every ROA $r_j \in \mathbb{R} - \mathbb{C}$, we add r_j into $\mathbb{C}$ if r_j is close to some ROA $r_i \in \mathbb{C}$ in terms of their divergence, spatial, and sequence closeness. We terminate a running iteration if there is no more r_j fulfilling the closeness constraints. Finally, we construct a new ROA by merging all $r_i \in \mathbb{C}$ and remove the elements in $\mathbb{C}$ from $\mathbb{R}$. We iteratively execute the above procedures until $\mathbb{R}$ becomes empty. We list the pseudocode for the regions refinement in Algorithm 2.

Figure 6 shows an example of the merging. In this example, the solid line between two ROAs r_i and r_j indicates that r_i and r_j fulfill the spatial and sequence constraints. The value on the lines represents the JS-divergence of r_i and r_j. Suppose that λ_{JS} is set to 0.3 and r_1 is the first selected ROA. We add r_2 and r_4 into $\mathbb{C}$ since they are close to r_1 in terms of all closeness measures. Subsequently, r_3 is added into $\mathbb{C}$ since r_3 is close to r_2. At the end

Algorithm 2 Regions of Attraction Refinement

```
Input: Travel topics $\overrightarrow{\theta}_{1:K}$, ROAs $\mathbb{R}^{(i)}$
Output: Refined ROAs $\mathbb{R}^{(i+1)}$
Algorithm $refineRegions(\overrightarrow{\theta}_{1:K}, \mathbb{R}^{(i)})$
for $i := 1$ to $|\mathbb{R}|$ do                                  ▷ travel topic associations
    for $j := 1$ to $K$ do
        $\alpha_{i,j} := \pi_j p(r_i|\overrightarrow{\theta}_j)$
        $Sum := Sum + \alpha_{i,j}$
    for $j := 1$ to $K$ do
        $\alpha_{i,j} := \alpha_{i,j}/Sum$
while $\mathbb{R}^{(i)}$ is not empty do                            ▷ refinement
    pick $r_j \in \mathbb{R}^{(i)}$ where $r_j$ has the best popularity in $\mathbb{R}^{(i)}$
    $\mathbb{C} := \emptyset$; $\mathbb{Q} := \{r_j\}$
    while $\mathbb{Q}$ is not empty do
        pop $r_k$ from $\mathbb{Q}$
        for all $r_l \in V(r_k, \lambda_E)$ do                      ▷ spatial close
            if $D_{JS}(\overrightarrow{\alpha_k}||\overrightarrow{\alpha_l}) < \lambda_{JS}$ then      ▷ divergence close
                if $D_S(r_k, r_l) = true$ then                      ▷ sequence close
                    $\mathbb{R}^{(i)} := \mathbb{R}^{(i)} \setminus \{r_l\}$; push $r_l$ into $\mathbb{Q}$
        $\mathbb{C} := \mathbb{C} \cup \{r_k\}$
    Insert $\mathbb{C}$ into $\mathbb{R}^{(i+1)}$
return $\mathbb{R}^{(i+1)}$
```

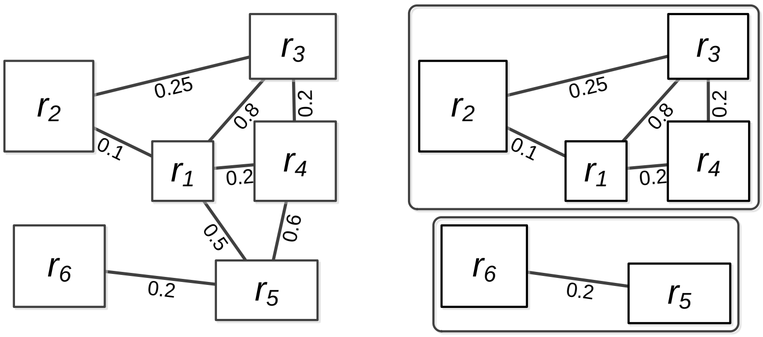
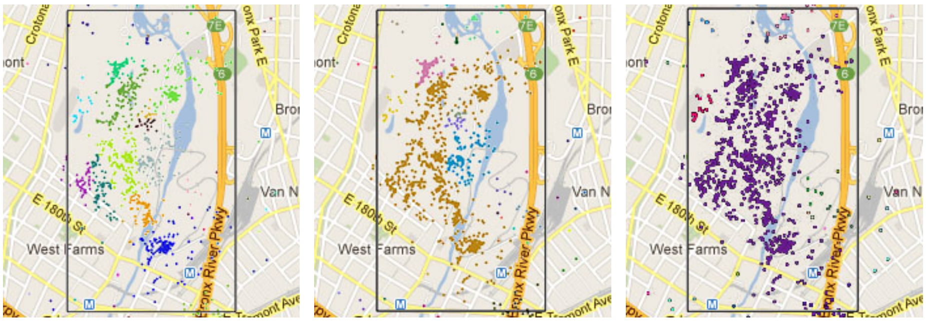


Fig. 6 An example of ROA refinement

of this iteration, we construct a refined ROA by merging r_1, r_2, r_3, and r_4. In the next iteration, we construct another refined ROA by merging r_5 and r_6.

3.3 Mutual Reinforcement

As shown in Section 3.1 and 3.2, the travel topics are constructed based on a set of ROAs while the ROA refinement is done by giving a set of travel topics. It is obvious that we can manage these two processes into an iterative framework. We first obtain a set of small ROAs, $\mathbb{R}^{(0)}$, by a standard clustering approach. Then, we perform Algorithm *constructTopics* to compute a set of travel topics $\overrightarrow{\theta}^{(1)}_{1:K}$ based on $\mathbb{R}^{(0)}$. Subsequently, we can pass $\overrightarrow{\theta}^{(1)}_{1:K}$ to Algo-



(a) Initial $\mathbb{R}^{(0)}$ (b) Intermediate $\mathbb{R}^{(i)}$ (c) Final $\mathbb{R}^{(n)}$

Fig. 7 ROA refinement by mutual reinforcement in Bronx Zoo, NY

rithm *refineRegions* and it returns a set of refined ROAs $\mathbb{R}^{(1)}$. We iteratively execute this mechanism until only a little (less than 1% in this work) ROAs are refined in this iteration. Figure 7 shows three different stages of ROAs in `Bronx Zoo, NY`. The pseudocode of the complete framework is shown in Algorithm 3.

Algorithm 3 Mutual reinforcement

```
   Input: Tour trajectory set $\mathbb{T}$
   Output: Travel Topics $\overrightarrow{\theta}_{1:K}$, ROAs $\mathbb{R}$
   Algorithm $mutualreinforcement(\mathbb{T})$
1: compute $\mathbb{R}^{(0)}$ by standard clustering on the data in $\mathbb{T}$
2: $i := 0$
3: while $true$ do
4:     $i := i + 1$
5:     $\overrightarrow{\theta}^{(i)}_{1:K} := constructTopics(\mathbb{T}, \mathbb{R}^{(i-1)})$
6:     $\mathbb{R}^{(i)} = refineRegions(\overrightarrow{\theta}^{(i)}_{1:K}, \mathbb{R}^{(i-1)})$
7:     if $\mathbb{R}^{(i)} = \mathbb{R}^{(i-1)}$ then go to line 8
8: return $\overrightarrow{\theta}^{(i)}_{i:K}$ and $\mathbb{R}^{(i)}$
```

4 Experiments

4.1 Datasets

We evaluate our solutions using eight photo datasets collected from Paris, New York, Rome, London, Tokyo, Hong Kong, Berlin and Barcelona, respectively, where the photos and their metadata are came from 2 sources (4 of the datasets

Table 1 Statistic Information of Data

City	**Photos**	**Tours**	**Init. ROAs**	**Source**
Paris, France	1,511.531	15,540	4,995	Flickr API
New York City, US	1,067,964	3,416	5,458	Flickr API
Roma, Italy	343,917	1,981	2,407	Flickr API
London, UK	1,305,977	8,134	7,362	Flickr API
Tokyo, Japan	147,268	795	1,404	Yahoo! [37]
Hong Kong, China	251,403	1,216	4,710	Yahoo! [37]
Berlin, Germany	310,661	1982	3.847	Yahoo! [37]
Barcelona, Spain	310,904	1,328	2,404	Yahoo! [37]

are crawled from Flickr$^{\text{TM}}$and 4 of them are provided by Yahoo! Webscope$^{\textbf{TM}}$ Program [37]. Table 1 shows the detail statistic of our datasets.

For the sake of identification quality, we keep a tour trajectory in the datasets only if it contains 10 or more photos at different locations. We generate a set of initial ROAs $\mathbb{R}$ by applying mean-shift algorithm [7] on the raw datasets. The ROAs who are visited by at least 2 users are used as the input of Algorithm 3.

In this work, we empirically determined the parameter values by investigating the first few travel topics. As a note, the performances are not very sensitive to the similarity ratio λ_{sim} (default 0.6), JS-divergence threshold λ_{JS} (default 0.02), and Euclidean distance threshold λ_E (default 50m). The effectiveness of the originality threshold λ_{ori} (default 10) is thoroughly evaluated in [19]. The effect of background coefficient b is used to leverage the effect of those popular ROAs, which is set to a reasonable value (i.e., 0.4).

4.2 Identification of Regions of Attraction

Though the ROA identification is not the main concern in this work, the quality of ROAs, however, is important to the travel topic extraction. In this section, we demonstrate the effectiveness of our refinement and reinforcement process (cf. Section 3.2 and 3.3) on different clustering techniques [26,40].

Metric functions. The evaluation is based on a set of *ground truth* ROAs $\mathbb{M}$, that are extracted from the meta-data of OpenStreetMap [2]. In this work, we use a *score* function, $Score(\mathbb{R}, \mathbb{M})$, to assess the identification quality, which aims to assess the uniqueness of the identification. We claim that an ROA r_i accurately identifies a ground truth region m_j if all photos in m_j are contained in r_i. Given a ground truth region m_j, the most accurate ROA r_j^* that identifies m_j can be defined as follows.

$$r_j^* = \underset{r_j^* \in \mathbb{R}}{\operatorname{argmax}} \frac{|r_j^* \cap m_j|}{|m_j|} \tag{15}$$

[2] To simplify our evaluation, we use minimum bounded rectangle to assign the photos into the ROA of $\mathbb{M}$.

where $|r_j^*|$ and $|m_j|$ are the number of photos in r_j^* and m_j, respectively, $|r_j^* \cap m_j|$ is the number of photos in the intersection of r_i^* and m_j.

Based on the definition of r_j^*, we define a function ρ to measure the uniqueness degree as follows.

$$\rho(r_j^*, m_j) = \frac{|r_j^* \cap m_j|}{|m_j|} \cdot \frac{1}{f_P(r_j^*)} \tag{16}$$

where $f_P(r_j^*)$ is a penalty factor which indicates the number of the ground truth ROAs treating r_j^* as their representation. Accordingly, the score of $\mathbb{R}$ is defined by the average uniqueness degree of the ground truth ROAs in $\mathbb{M}$.

$$Score(\mathbb{R}, \mathbb{M}) = \sum_{m_j \in \mathbb{M}} \rho(r_j^*, m_j)/|\mathbb{M}| \tag{17}$$

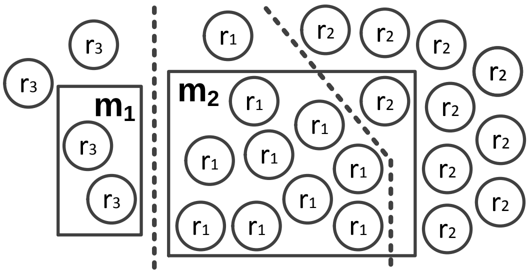


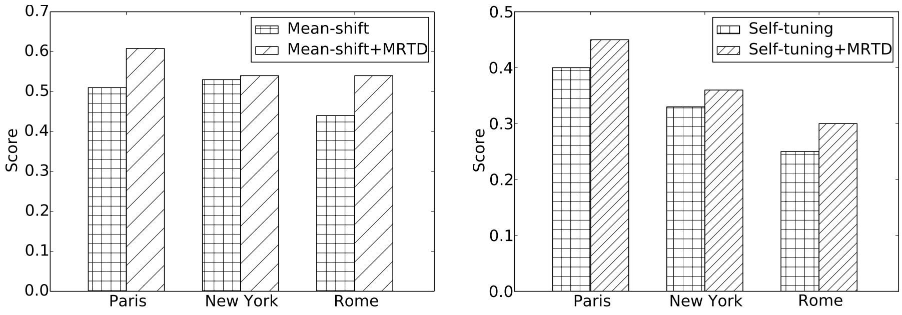
Fig. 8 Example of evaluation

We use an example in Figure 8 to demonstrate $Score(\mathbb{R}, \mathbb{M})$. In this example, there are three ROAs $\mathbb{R}$ identified by our algorithm and two real ROAs $\mathbb{M}$ extracted from OpenStreetMap. Suppose $Score(\{r_1, r_2, r_3\}, \mathbb{M})$ is 0.95. If we merge r_1 and r_2, then $Score(\{r_1 \cup r_2, r_3\}, \mathbb{M})$ becomes 1 since m_2 is uniquely identified by the refined ROA $r_1 \cup r_2$ (i.e., $f_P(r_1 \cup r_2)$=1).

Evaluation. Our reinforcement framework, Mutual Reinforcing Travel Topic Discovery (MRTD), requires a set of pre-defined ROAs as input. As reported by [40], mean-shift [26] and self-tuning [40] are the best two methods to identify ROAs from geo-tagged data. Thus we demonstrate the effect of our ROA refinement process using the result of these methods as input. For fairness, we tune the best parameters for [26] [3].

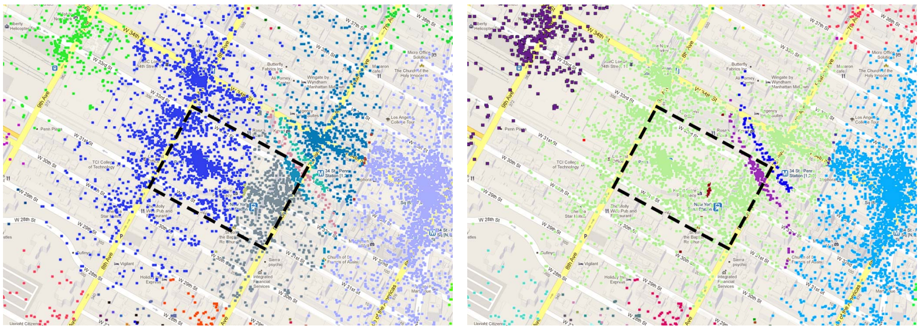
As shown in Figure 9, our refinement process improves the ROA quality in terms of their uniqueness (i.e., *Score* function). This demonstrates merging small ROAs based on their closeness (i.e., travel topics, spatial, and sequence) is effective.

[3] [40] is a parameter free technique.



(a) Mean-shift (b) Self-tuning

Fig. 9 Evaluation of ROA identification



(a) Mean-shift (b) MRTD

Fig. 10 Identification of Madison Square Garden

We illustrate a real ROA identification result in Figure 10. In this example, `Madison Square Garden` is divided into several small ROAs by Mean-shift, that obviously affects the travel topic analysis. The ROAs returned by our iterative framework are more preferred as our framework successfully refines a set of small ROAs into a large ROA (i.e., the coverage area is very close to the ground truth ROA). The quality of ROAs is important to our problem as each travel topic is extracted based on the ROAs (cf. Algorithm 1).

4.3 Travel Topic Analysis

In this section, we compare our topic discovery result with two recent works, RouteSimilarity [45] and Topic-Markov Model [26], and two baseline approaches, Frequent Pattern tree-based pattern fragment Growth mining method (FP-Growth) [14] and Latent Dirichlet Allocation (LDA) [4]. RouteSimilarity [45] uses longest common sub-sequence as the similarity metric and applies hierarchical agglomerative clustering to group similar travel trajectories as a travel topic. Topic-Markov Model [26] recommends ROAs based on probabilistic latent semantic analysis (PLSA), where the result of PLSA is analogous to the result of our mixture probability model. FP-Growth is an efficient method for mining frequent patterns and LDA is another widely accepted method in topic discovery problems.

We randomly select 10% trajectories as the testing data. Except the testing data, all remaining trajectories are treated as the training data. Note that RouteSimilarity [45] does not summarize the result where each travel topic simply consists of a group of trajectories instead of a summarization (cf., Definition 3). To fairly assess the performance of RouteSimilarity, we calculate the importance of an ROA r_j of a topic $\overrightarrow{\theta}_i$ by the number of trajectories (in $\overrightarrow{\theta}_i$) passing r_j. Moreover, every method uses the same set of ROAs (generated by MRTD) as input since the quality is better than other clustering approaches (cf., Figure 8).

Given a list of k travel topics $\overrightarrow{\theta}_{1:k}$ (i.e., a subset of $\overrightarrow{\theta}_{1:K}$), we claim that a user would satisfy if the suggested travel topic(s) cover lots of her travel interests (i.e., ROAs). In this work, we only use the k most similar topics $\overrightarrow{\theta}_{1:k}$ to assess the quality of a topic set $\overrightarrow{\theta}_{1:K}$ since users are not interested in screening everything but would instead like a more informative and manageable result (i.e., k topics). First, we define the similarity between a topic $\overrightarrow{\theta}_i$ and a trajectory t as follows.

$$similarity(t, \overrightarrow{\theta}_i) = \sum_{r_j \in R(t)} hit(r_j, \overrightarrow{\theta}_i) \tag{18}$$

where $hit(r_j, \overrightarrow{\theta}_i) = 1$ if r_j is one of top-m ROAs (e.g., $m = 100$) in $\overrightarrow{\theta}_i$ and $hit(r_j, \overrightarrow{\theta}_i) = 0$ otherwise.

Given the k most similar topics $\overrightarrow{\theta}_{1:k}$ based on $similarity(\cdot)$, we define the *coverage ratio* metric that indicates the percentage of a user tour t being covered by $\overrightarrow{\theta}_{1:k}$ (where $\overrightarrow{\theta_1} \prec \overrightarrow{\theta_2} \prec \cdots \overrightarrow{\theta_k}$).

$$coverage(t, \overrightarrow{\theta}_{1:k}) = \frac{1}{|R(t)|} \sum_{r_j \in R(t)} \max_{\overrightarrow{\theta}_i \in \overrightarrow{\theta}_{1:k}} \{\frac{hit(r_j, \overrightarrow{\theta}_i)}{i}\} \tag{19}$$

where the denominator, i, can be viewed as a punishment based on the ranking order. In the ideal case, every ROA of t is covered by the first topic so that there is no punishment. $R(t)$ represents the set of ROAs visited by tour t, that indicates the ground truth travel interests of a user (i.e., their real travel tour). If the *coverage ratio* is high, the user should be satisfied with the suggested topics as her interests (i.e., trajectory t) are well covered by $\overrightarrow{\theta}_{1:k}$.

Figure 11 is a concrete example to demonstrate the coverage metric. Our objective is to assess the quality of 3 most similar travel topics $\overrightarrow{\theta}_1, \overrightarrow{\theta}_2, and \overrightarrow{\theta}_3$

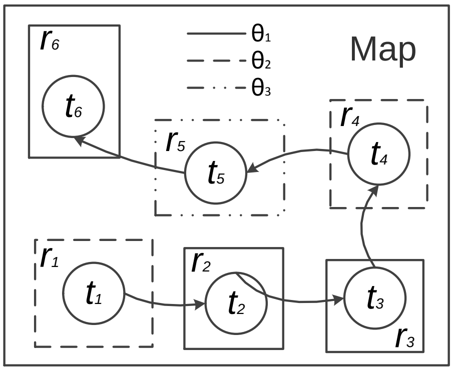


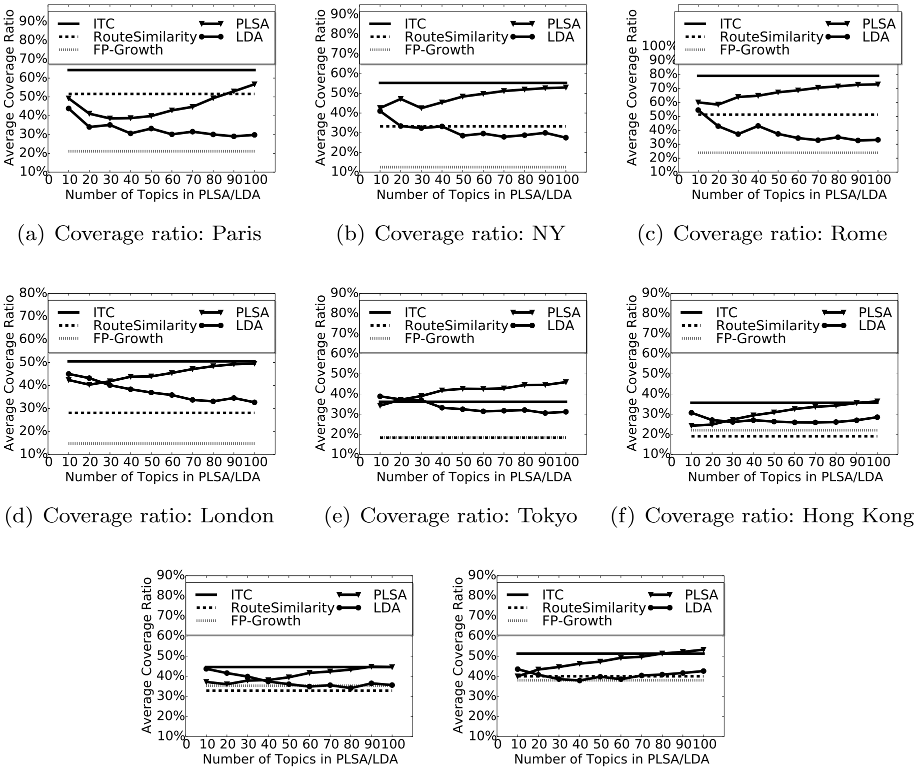
Fig. 11 Example of coverage

to a tour t (illustrated by the arcs in the figure). There are 3 ROAs covered by $\overrightarrow{\theta}_1$, 2 ROAs covered by $\overrightarrow{\theta}_2$ and 1 ROA covered by $\overrightarrow{\theta}_3$ (cf. function $hit(\cdot)$). Thereby, $coverage(t, \overrightarrow{\theta}_{1:k}) = (3/1 + 2/2 + 1/3)/6 = 13/18 = 0.72$.

Figure 12 and Figure 13 show the coverage ratio and construction time as a function of number of topics on all eight datasets, after setting $k = 1$ (i.e., assessing the best travel topic). It should be noted that only LDA and PLSA are sensitive to the number of topics. In our internal tuning, the performance of RouteSimilarity is steadily worse than our method since RouteSimilarity groups topics by visit order similarity but not by user interests. Thereby, we only report the best performance of RouteSimilarity in Figure 12. Interestingly, FP-Growth has the lowest construction time but its coverage ratio is also the lowest since the extracted topics contain too many popular ROAs (i.e., over-popular landmarks). Moreover, PLSA is superior to LDA for travel topic discovery as PLSA can generate a trajectory with certain accuracy based on fewer topics [4]. Our approach, ITC (cf. Algorithm 1), offers the best coverage ratio in the majority cases since ITC aims at generating trajectories by limited travel topics (based on the incremental construction and the originality assessment). PLSA and LDA are better than ITC on the Tokyo dataset when their number of topics are properly set; however, it is time consuming to tune the best number of topics as shown in Figure 13.

Figure 14 shows the coverage ratio as a function of k (i.e., the number of k most similar topics) on all datasets. Based on the tradeoff between the quality and the construction cost, we set the number of topics in PLSA and LDA as 50. ITC consistently outperforms other methods in terms of the coverage ratio (except Tokyo). This means that ITC offers more informative travel topics to users when there are only k most similar topics available.

[4] This is because the travel topics discovered by PLSA do not necessarily follow the Dirichlet distribution.



(a) Coverage ratio: Paris (b) Coverage ratio: NY (c) Coverage ratio: Rome

(d) Coverage ratio: London (e) Coverage ratio: Tokyo (f) Coverage ratio: Hong Kong

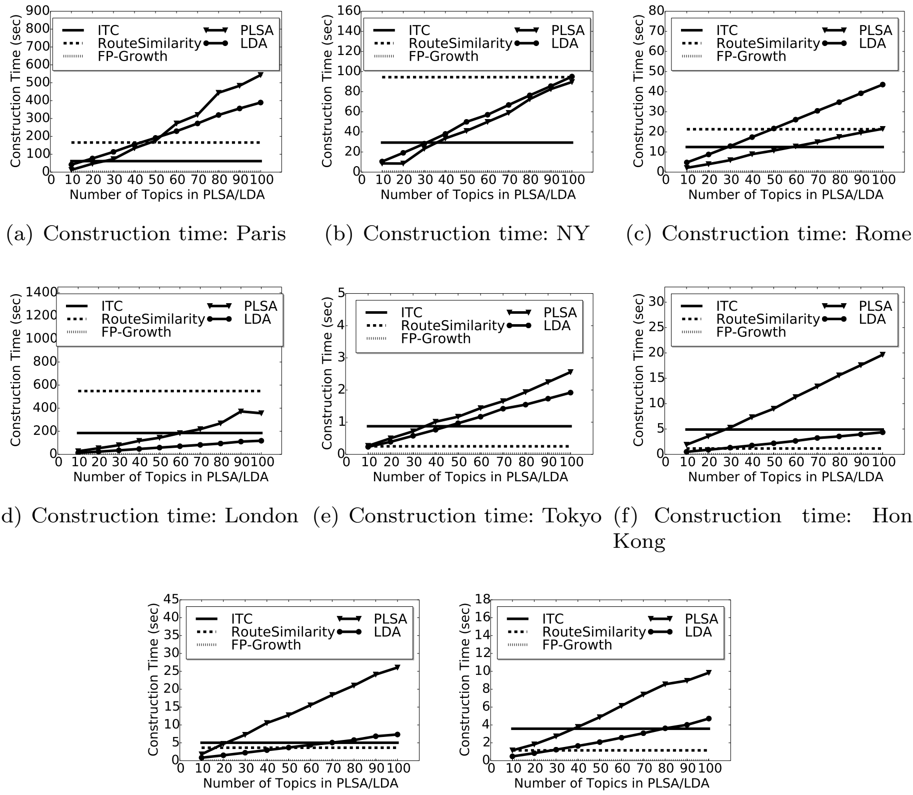
(g) Coverage ratio: Berlin (h) Coverage ratio: Barcelona

Fig. 12 Travel topic coverage ratio

4.4 Travel Topic Illustrations

In this section, we demonstrate our result by plotting the first two travel topics (in terms of their coefficients π_i) and the background (most popular) topic in Pairs, New York and Rome. In each travel topic, we report the best 5 ROAs based on their probability $p(r_j|\overrightarrow{\theta}_i)$. Background topic demonstrates the most common travel interests while a travel topic shows a special travel interest. To illustrate the result more clearly, we highlight the areas in the map of the cities and show the representative building of each area in tables.

Travel Topic of Paris. According to Table 2, we can found that the background topic contains the world-renowned landmarks, such as `Eiffel Tower`, `Louvre Palace`, and `Arc de Triomphe`. Besides, the best 5 ROAs in these two travel topics of Paris are completely different which reveals the effective-



(a) Construction time: Paris (b) Construction time: NY (c) Construction time: Rome

(d) Construction time: London (e) Construction time: Tokyo (f) Construction time: Hong Kong

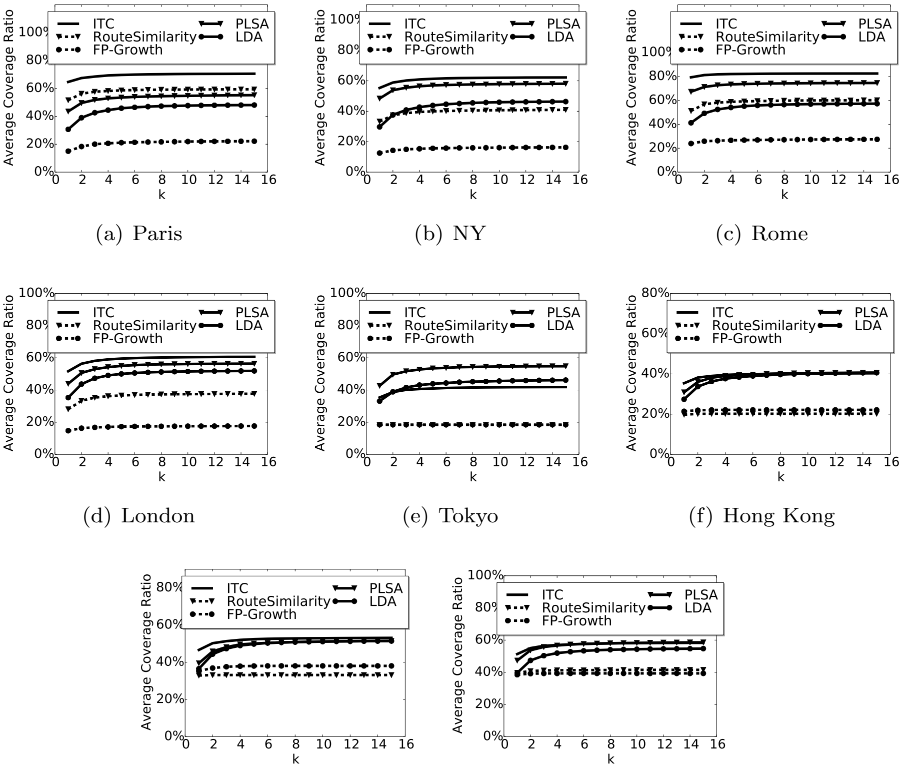
(g) Construction time: Berlin (h) Construction time: Barcelona

Fig. 13 Travel topic construction time

Table 2 Travel Topic of Paris

Rank	1^{st} **Travel Topic,** $\overrightarrow{\theta}_1^{Paris}$	2^{nd} **Travel Topic,** $\overrightarrow{\theta}_2^{Paris}$	**Background Topic,** $\overrightarrow{\beta}^{Paris}$
1	Louvre Palace	Place Saint-Michel	Eiffel Tower
2	Great Palace	Pont des Arts	Louvre Palace
3	Gare du Nord	Jardins du Trocadéro	Arc de Triomphe
4	Gare Saint-Lazare	Place de l'Opéra	basilique du sacré-cœur
5	Tuileries Palace	Pont Neuf	Cathédrale Notre Dame de Paris

ness of our incremental topic construction algorithm. More specifically, the first travel topic visits the famous museums (`Louvre Palace`, `Great Palace` and `Tuileries Palace`) and 2 historical stations (`Gare du Nord` and `Gare Saint-Lazare`). The tourists, who are interested in ancient wisdom, will find it is very useful if we recommend this travel topic to them. The second travel topic mainly recommends the public squares (`Place Saint-Michel`, `Jardins du Trocadéro`, `Place de l'Opéra`, `Invalides Gardens`) and the bridges (`Pont`



(a) Paris (b) NY (c) Rome

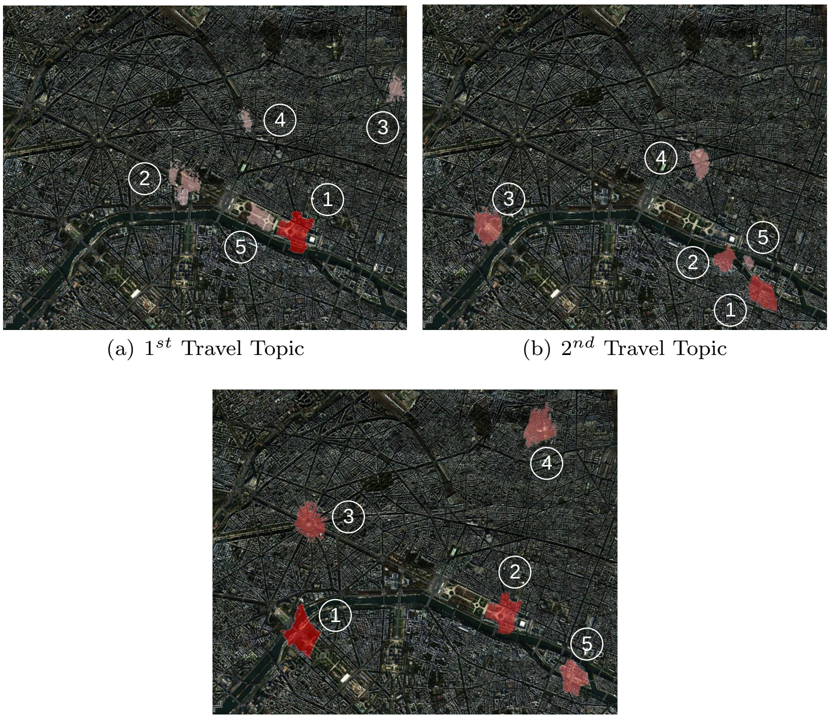
(d) London (e) Tokyo (f) Hong Kong

(g) Berlin (h) Barcelona

Fig. 14 Effect of k in $\theta_{1:k}$

`des Arts` and `Pont Neuf`). The 2nd topic lists the most awesome open places for taking photos in Paris, which could be a very good reference to those tourists who would like to have a recreation trip in Paris. We plot their geographical locations in Figure 15 for clarity.

Travel Topic of New York. As shown in Figure 16 and Table 3, the back-



(a) 1^{st} Travel Topic

(b) 2^{nd} Travel Topic

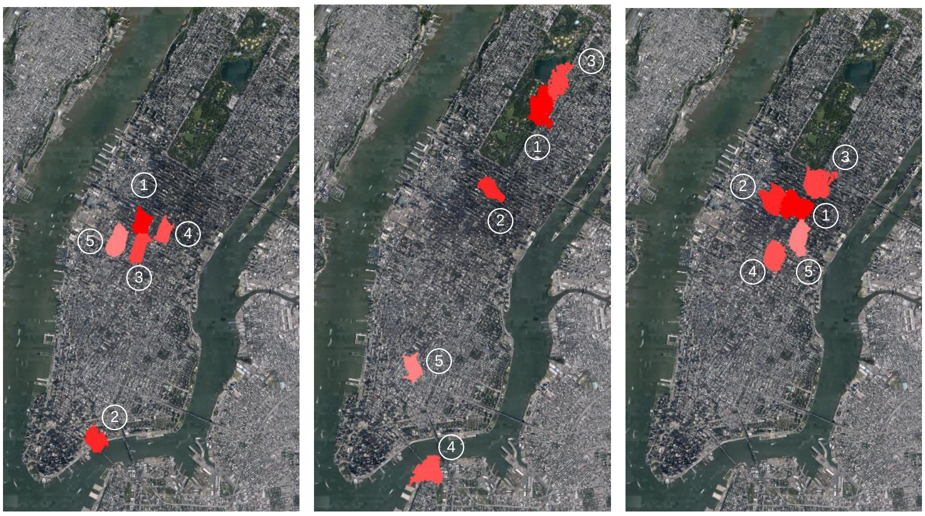
(c) Background Topic

Fig. 15 Travel topic illustration of Paris

Table 3 Places in New York travel topics

Rank	1^{st} **Travel Topic,** $\overrightarrow{\theta}_1^{NY}$	2^{nd} **Travel Topic,** $\overrightarrow{\theta}_2^{NY}$	**Background Topic,** $\overrightarrow{\beta}^{NY}$
1	Broadway theatre	Metropolitan Museum of Art	Rockefeller Center
2	Brooklyn Bridge	American Folk Art Museum, Museum of Modern Art	Times Square
3	Herald Square	National Academy Museum, Solomon R. Guggenheim Museum, Cooper-Hewitt Design Museum	Apple Store
4	New York Public Library	Fulton Ferry, Brooklyn	Empire State Building
5	Madison Square Garden	SoHo	Grand Central Terminal

ground topic of New York City consists of modern buildings and skyscrapers, such as `Rockefeller Center`, `Times Square`, `Apple Store`, `Empire State Building` and `Grand Central Terminal`. The first travel topic, $\overrightarrow{\theta}_1^{NY}$, includes some popular places, such as `Broadway theatre` and `Herald Square` (i.e., venues for theatrical performances), `Brooklyn Bridge` (i.e., the oldest suspension bridges), and `the New York Public Library` (i.e., 3rd largest public library in the world). The second topic is related to art, where it covers a lot of museums (`Metropolitan Museum of Art`, `Museum of Modern Art`, etc.) and `SoHo` (`SoHo-Cast Iron Historic District`) which used to be the



(a) 1^{st} Travel Topic (b) 2^{nd} Travel Topic (c) Background Topic

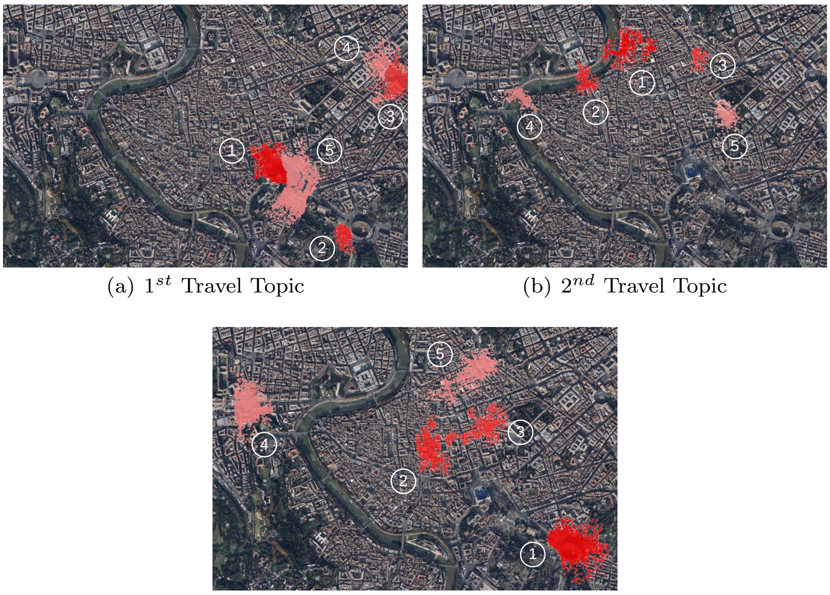
Fig. 16 Travel topic illustration of New York

district of many artists' lofts and art galleries located, and is now a place for outlets and shops.

Travel Topic of Rome. Figure 17 and Table 4 list the top-2 travel topics and the background topic of Rome. It is not surprising that the background topic covers `The Colosseum`, `Pantheon`, `Saint Peter's Square`, and `Spanish Steps` since these places are commonly visited by travelers. For instance, `The Colosseum` is the most reputable ancient building of Roman Empire, `Pantheon` is considered as one of the greatest works of Roman architecture and engineering which is still used for celebrating masses, `Saint Peter's Square` is the open space in front of `St. Peter's Basilica` which remains one of the largest churches of Renaissance architecture, and `Spanish Steps` is the widest stair case in Europe. The first travel topic, $\overrightarrow{\theta}_1^{Rome}$, is a set of his-

Table 4 Places in Rome travel topics

Rank	1^{st} **Travel Topic,** $\overrightarrow{\theta}_1^{Rome}$	2^{nd} **Travel Topic,** $\overrightarrow{\theta}_2^{Rome}$	**Background Topic,** $\overrightarrow{\beta}^{Rome}$
1	Altare della Patria	Ara Pacis	The Colosseum
2	Arch of Constantine	Palace of Justice	Pantheon
3	Piazza della Repubblica	Piazza near Spanish Steps	Fontana di Trevi, Accademia di San Luca, Istituto Nazionale per la Grafica
4	Acqua Felice, Santa Maria della Vittoria, Santa Susanna	Ponte Vittorio Emanuele II	Saint Peter's Square
5	Roman Forum	Quirinal Palace	Spanish Steps



(a) 1^{st} Travel Topic

(b) 2^{nd} Travel Topic

(c) Background Topic

Fig. 17 Travel topic illustration of Rome

toric building in Rome. It contains monuments(Altare della Patria, Arch of Constantine), churches(Santa Maria della Vittoria, Santa Susanna) and plazas(Piazza della Repubblica, Roman Forum). Travelers who like architectural aesthetics would not miss them. The second travel topic, $\overrightarrow{\theta}_2^{Rome}$, covers the official residence of the president of the Italian Republic(Quirinal Palace), the seat of Supreme Court of Cassation(Palace of Justice), and some places from the Tibre River to Spanish Steps (including Ara Pacis, Piazza near Spanish Steps, Ponte Vittorio Emanuele II). The travel interest of $\overrightarrow{\theta}_2^{Rome}$ is a fusion of multiple domains such as nature, history and politics.

5 Related Work

Our work is related to trajectory mining problems. Giannotti et al. [13] discover moving behaviors based on the transition time of trajectories, where their ROAs are identified by grid and point density. Choudhury et al. [9,10] proposed a method to construct trajectories from photo data automatically. Wei et al. [38] construct the top-k popular routes from uncertain trajectories using a mutual reinforcement approach. Cho et al. [8] observed that a long-distance travel correlates with user's social relationship such that they proposed a model to capture periodic day-to-day movements and social structures. Popescu and Greffenstette [33] studied how to construct high quality trajectories from noise geo-tagged photos. Jeung et al.[17] refine the trajectory clustering by decomposing large clusters which helps to detect hidden trajectory patterns by Hidden Markov Models (HMM). Lee et al. [27] proposed a problem that extracts similar sub-trajectories from a trajectory dataset. Monreale et al. [31] studied a framework which recommends the next moving location based on trajectory patterns. The development of trajectory mining problems is summarized in [18]. However, all these work focus on the visiting order of trajectories while our work ignores the visiting order and concentrates on mining the latent interests of tourism. A travel topic in this work is the interest of ROAs for a group of travelers instead of a single trajectory.

To the best of our knowledge, Zheng et al. [45] is the closest competitor to our work which extracts travel topics from a set of user trajectories. Their framework first identifies ROAs by DBSCAN algorithm based on a user trajectory database and then discover similar trajectories as a travel topic using hierarchical agglomerative clustering, where the the similarity of two trajectories is measured by their longest common subsequence (LCSS). The major weakness of their work is that LCSS treats two trajectories as similar when their visiting orders are similar. As explained in Section 1, the visiting order of tours is not the most critical factor to identify travel topics. Instead, our approach adopts a mixture model to extract travel topics based on the visiting ROAs of trajectories. This setting overcomes the weakness of LCSS such that our approach returns more meaningful result to travelers. In addition, our solution identifies better ROAs using a mutual reinforcing method instead of a simple clustering approach (e.g., DBSCAN or hierarchical clustering).

Our work is also related to other tour recommendation problems. Kodama et al. [24] take user locations and preferences into account such that they can recommend surrounding locations by a spatial skyline method. Park et al. [32] recommend the locations of the most preferred items based on user profiles and their context information (i.e., weather information, temperature, season, time of day, periods and user location). Shi et al. [36] propose category-regularized matrix factorization for personalized recommendation system based on user-to-landmark preferences. Zheng et al. [44] analyze the user-location relationship and recommend locations based on a HITS framework [23]. Popescu and Grefenstette [34] adopt collaborate filtering to recommend personalized tours. Hao et al. [15] propose a Location-Topic model to extract location-representative knowledge from text. Xia et al. [39] apply Markov chain to conduct travel analysis and provide recommendations. Levandoski et al. [28] develop a local-aware recommendation system based on location-based rating. Kurashima et al. [26] propose a Markov-Topic model that combines transition probability and a spatial topic model based on a *probabilistic latent semantic analysis* (PLSA). Their work can be viewed as a competitor to ours as PLSA returns analogous probabilities to our mixture probability model. To the best of our knowledge, these tour recommendation problems either recommend the next visiting location or offer a suggested tour route to their users. Instead, our work aims at revealing travel topics based on user common interests which helps travel agents, trip advisers and governments to predominate the entire picture of the tour interests in a city.

Our travel topic extraction is based on topic discovery problems of documents [16,19,42] while the corpus words (i.e., regions of attraction) are updated over each iteration in our work. The work [42] is a variant of [16], where [42] provides a study on the background distribution. Recently, the work [19] derives an idea from [3] that builds a dynamic model involving significance and novelty. Yin et al. [41] propose a geographical topic model for the documents with spatial information, which enriches the content of topic in GPS-associated information such as tweets with location in microblogging. All these models are adopted for text mining and discovering the topics of documents. In this work, we extend their ideas to support travel topic extraction from geo-tagged photoes.

6 Conclusion

In this manuscript, we studied a travel topic extraction problem in geo-tagged photo datasets. To the best of our knowledge, this is the first thorough study for this problem based on mixture models. We proposed a novel framework by mutually refining the travel topics and regions of attraction. We demonstrated that our approach can further improve the quality of ROAs that are generated by the state-of-the-art clustering techniques. Moreover, our travel topics are particularly helpful in trip recommendation especially when there is no prior personal information from users. We believe our travel topic discovery techniques are easily integrated into existing trip advisor systems. In the future, we plan to revisit the iterative refinement procedure where the ROAs are refined by both merging (i.e., generalizing tourist regions) and partitioning (i.e., specifying tourist spots) techniques.

References

1. Ayala, G., Sebastián, R., Díaz, M.E., Díaz, E., Zoncu, R., Toomre, D.: Analysis of spatially and temporally overlapping events with application to image sequences. IEEE TPAMI **28**(10), 1707–1712 (2006)
2. Bezahaf, M., Iannone, L., de Amorim, M.D., Fdida, S.: Lord: Tracking mobile clients in a real mesh. Ad Hoc Networks **9**(8), 1461–1475 (2011)
3. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: ICML, pp. 113–120 (2006)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
5. Chellappa, R.K., Sin, R.G.: Personalization versus privacy: An empirical examination of the online consumer's dilemma. Information Technology and Management **6**(2-3), 181–202 (2005)
6. Cheng, A.J., Chen, Y.Y., Huang, Y.T., Hsu, W.H., Liao, H.Y.M.: Personalized travel recommendation by mining people attributes from community-contributed photos. In: ACM Multimedia, pp. 83–92 (2011)
7. Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Mach. Intell. **17**(8), 790–799 (1995)

8. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: KDD, pp. 1082–1090 (2011)
9. Choudhury, M.D., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., Yu, C.: Constructing travel itineraries from tagged geo-temporal breadcrumbs. In: WWW, pp. 1083–1084 (2010)
10. Choudhury, M.D., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., Yu, C.: Constructing travel itineraries from tagged geo-temporal breadcrumbs. In: WWW, pp. 1083–1084 (2010)
11. Cranor, L.F., Reagle, J., Ackerman, M.S.: Beyond concern: Understanding net users' attitudes about online privacy. Cambridge, MA: MIT Press (2000)
12. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, Series B **39**(1), 1–38 (1977)
13. Giannotti, F., Nanni, M., Pinelli, F., Pedreschi, D.: Trajectory pattern mining. In: KDD, pp. 330–339 (2007)
14. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA., pp. 1–12 (2000). DOI 10.1145/342009.335372. URL http://doi.acm.org/10.1145/342009.335372
15. Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J.M., Pang, Y., Zhang, L.: Equip tourists with knowledge mined from travelogues. In: WWW, pp. 401–410 (2010)
16. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR, pp. 50–57 (1999)
17. Jeung, H., Shen, H.T., Zhou, X.: Mining trajectory patterns using hidden markov models. In: DaWaK, pp. 470–480 (2007)
18. Jeung, H., Yiu, M.L., Jensen, C.S.: Trajectory pattern mining. In: Computing with Spatial Trajectories, pp. 143–177. Springer (2011)
19. Jo, Y., Hopcroft, J.E., Lagoze, C.: The web of topics: discovering the topology of topic evolution in a corpus. In: WWW, pp. 257–266 (2011)
20. Kalnis, P., Mamoulis, N., Bakiras, S.: On discovering moving clusters in spatio-temporal data. In: SSTD, pp. 364–381 (2005)
21. Kennedy, L.S., Naaman, M.: Generating diverse and representative image search results for landmarks. In: WWW, pp. 297–306 (2008)
22. Kisilevich, S., Mansmann, F., Keim, D.A.: P-dbscan: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In: COM.Geo (2010)
23. Kleinberg, J.M.: Hubs, authorities, and communities. ACM Comput. Surv. **31**(4es), 5 (1999)
24. Kodama, K., Iijima, Y., Guo, X., Ishikawa, Y.: Skyline queries based on user locations and preferences for making location-based recommendations. In: GIS-LBSN, pp. 9–16 (2009)
25. Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Math. Statistics **22**, 49–86 (1951)
26. Kurashima, T., Iwata, T., Irie, G., Fujimura, K.: Travel route recommendation using geotags in photo sharing sites. In: CIKM, pp. 579–588 (2010)
27. Lee, J.G., Han, J., Whang, K.Y.: Trajectory clustering: a partition-and-group framework. In: SIGMOD Conference, pp. 593–604 (2007)
28. Levandoski, J.J., Sarwat, M., Eldawy, A., Mokbel, M.F.: Lars: A location-aware recommender system. In: ICDE, pp. 450–461 (2012)
29. Lin, J.: Divergence measures based on the shannon entropy. IEEE TIT **37**, 145–151 (1991)
30. Lo, E., Kao, B., Ho, W.S., Lee, S.D., Chui, C.K., Cheung, D.W.: Olap on sequence data. In: SIGMOD Conference, pp. 649–660 (2008)
31. Monreale, A., Pinelli, F., Trasarti, R., Giannotti, F.: Wherenext: a location predictor on trajectory pattern mining. In: KDD, pp. 637–646 (2009)
32. Park, M.H., Hong, J.H., Cho, S.B.: Location-based recommendation system using bayesian user's preference model in mobile devices. In: UIC, pp. 1130–1139 (2007)
33. Popescu, A., Grefenstette, G.: Deducing trip related information from flickr. In: WWW, pp. 1183–1184 (2009)

34. Popescu, A., Grefenstette, G.: Mining social media to create personalized recommendations for tourist visits. In: COM.Geo, p. 37 (2011)
35. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: SIGIR, pp. 103–110 (2007)
36. Shi, Y., Serdyukov, P., Hanjalic, A., Larson, M.: Personalized landmark recommendation based on geotags from photo sharing sites. In: ICWSM (2011)
37. Team, T.Y.L.W.: I3 - yahoo flickr creative commons 100m. Accessed: 2014-09-30
38. Wei, L.Y., Zheng, Y., Peng, W.C.: Constructing popular routes from uncertain trajectories. In: KDD, pp. 195–203 (2012)
39. Xia, J.C., Zeephongsekul, P., Arrowsmith, C.: Modelling spatio-temporal movement of tourists using finite markov chains. Math. Comput. Simul. **79**(5), 1544–1553 (2009). DOI 10.1016/j.matcom.2008.06.007. URL `http://dx.doi.org/10.1016/j.matcom.2008.06.007`
40. Yang, Y., Gong, Z., U, L.H.: Identifying points of interest by self-tuning clustering. In: SIGIR, pp. 883–892 (2011)
41. Yin, Z., Cao, L., Han, J., Zhai, C., Huang, T.S.: Geographical topic discovery and comparison. In: WWW, pp. 247–256 (2011)
42. Zhai, C., Velivelli, A., Yu, B.: A cross-collection mixture model for comparative text mining. In: KDD, pp. 743–748 (2004)
43. Zheng, Y., Zhang, L., Ma, Z., Xie, X., Ma, W.Y.: Recommending friends and locations based on individual location history. TWEB **5**(1), 5 (2011)
44. Zheng, Y., Zhang, L., Xie, X., Ma, W.Y.: Mining interesting locations and travel sequences from gps trajectories. In: WWW, pp. 791–800 (2009)
45. Zheng, Y.T., Zha, Z.J., Chua, T.S.: Mining travel patterns from geotagged photos. ACM TIST **3**(3), 56 (2012)