Equip Tourists with Knowledge Mined from Travelogues

Qiang Hao[†*], Rui Cai[‡], Changhu Wang[‡], Rong Xiao[‡], Jiang-Ming Yang[‡], Yanwei Pang[†], Lei Zhang[‡]

[†]Tianjin University, Tianjin 300072, P.R. China

[‡]Microsoft Research Asia, Beijing 100190, P.R. China

[†]{qhao, pyw}@tju.edu.cn; [‡]{ruicai, chw, rxiao, jmyang, leizhang}@microsoft.com

ABSTRACT

With the prosperity of tourism and Web 2.0 technologies, more and more people have willingness to share their travel experiences on the Web (e.g., weblogs, forums, or Web 2.0 communities). These so-called travelogues contain rich information, particularly including location-representative knowledge such as attractions (e.g., *Golden Gate Bridge*), styles (e.g., *beach*, *history*), and activities (e.g., *diving, surfing*). The location-representative information in travelogues can greatly facilitate other tourists' trip planning, if it can be correctly extracted and summarized. However, since most travelogues are unstructured and contain much noise, it is difficult for common users to utilize such knowledge effectively. In this paper, to mine location-representative knowledge from a large collection of travelogues, we propose a probabilistic topic model, named as Location-Topic model. This model has the advantages of (1) differentiability between two kinds of topics, i.e., local topics which characterize locations and global topics which represent other common themes shared by various locations, and (2) representation of locations in the local topic space to encode both location-representative knowledge and similarities between locations. Some novel applications are developed based on the proposed model, including (1) destination recommendation for on flexible queries, (2) characteristic summarization for a given destination with representative tags and snippets, and (3) identification of informative parts of a travelogue and enriching such highlights with related images. Based on a large collection of travelogues, the proposed framework is evaluated using both objective and subjective evaluation methods and shows promising results.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models – *statistical*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering*

General Terms

Algorithms, Experimentation.

Keywords

Travelogue mining, probabilistic topic model, recommendation.

1. INTRODUCTION

Travel, as an integral part of human history, has become more and more popular in people's everyday lives in recent years, partly owing to the increasing amount of travel-related information and services on the Web, which provide people with efficient ways to online plan and prepare for their trips. Meanwhile, Web 2.0 technologies facilitate and encourage people to contribute rather than just obtain information, leading to a huge amount of user-generated content (UGC) on the Web. In the tourism domain, more and more people have willingness to record and share their travel experiences on weblogs, forums or travel communities, in the form of textual travelogues and photos taken during the trips.

Since travel-related UGC not only underlies the communities and social network among travelers but also provides other web users with rich information related to travel, how to leverage it has attracted extensive attention in the literature. For instance, a lot of work has been proposed to mine knowledge from user-contributed photos on *Flickr* [6] to support various applications such as landmark discovery and recognition [22], landmark image selection [11][18], location explorer [1][5], and image tag suggestion [15]. By contrast, fewer research efforts have been dedicated to knowledge mining from travelogues. One related work is [8], in which the authors proposed to generate overviews for locations by mining representative tags from travelogues. However, to the best of our knowledge, the complete framework of travelogue mining and its applications has not been specially investigated.

We claim that travelogues can serve as a promising resource of travel-related knowledge, which is complementary to user-generated photos because travelogues cover various travel-related aspects, including not only landmarks and natural things which correspond to specific visual descriptions in photos, but also abstract aspects (e.g., history, culture, *genius loci*) which are informative to tourists but difficult to visualize using photos. With such rich information, travelogues could support more comprehensive descriptions of locations and comparisons between locations than user-generated photos, and thus could be leveraged to recommend locations according to various queries. In addition, travelogues contain rich textual contexts of locations to meet various information needs. For example, representative snippets can be extracted to describe a location's characteristics and linked to original travelogues as detailed context.

Furthermore, in spite of the access to a great deal of structured travel-related information (e.g., vacation packages, flights, hotels) offered by travel websites and travel agents, many people who are planning a trip prefer to learn experience and guidance from other travelers. Travelogues supplement this structured information with unstructured but personal descriptions of tourist destinations and services. Although the information in a single travelogue is possibly noisy or biased, numerous travelogues as a whole could reflect people's overall preference and understanding of travel resources, and thus can serve as a reliable knowledge source.

However, acquiring the knowledge in travelogues is a non-trivial task, especially for common users. Actually, there is a gap between raw travelogues and the information needs of tourists due to the data's intrinsic limitations listed as follows:

* This work was performed at Microsoft Research Asia.

WWW 2010, April 26-30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

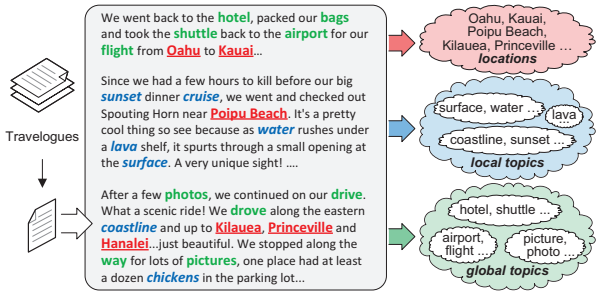


Figure 1. Different topics in travelogues, where *local topics* are shown in italic and blue; *global topics* are shown in green; and *locations* are shown in underline and red.

- **Noisy topics**: As other UGC, most travelogues are unstructured and contain much noise. For instance, the depictions of destinations and attractions, in which the common tourists are most interested, are usually intertwined with topics common in travelogues related to various locations.
- **Multiple viewpoints**: For each destination, there are usually various descriptions coming from many previous travelers. When trying to comprehensively know about a destination, users usually confront a dilemma that the viewpoint in each single travelogue may be biased, while it is time-consuming to read and summarize a number of related travelogues to outline an overview of the destination's characteristics.
- **Lack of destination recommendation**: Although a large collection of travelogues can cover most of popular destinations in the world, the depictions in a single travelogue usually focus on only one or a few destinations. Hence, for tourists who have particular travel intentions (e.g., go to a *beach*, go *hiking*) and need to determine where to go, there is no straightforward and effective way to obtain recommended destinations, except for surveying a lot of travelogues.
- **Lack of destination comparison**: In travelogues, besides the explicit comparison made by the authors, there is little information about the similarity between destinations, which is helpful for tourists who need suggestions about destinations similar (or dissimilar) to the ones that they are familiar with.

To overcome these limitations of raw travelogue data and bridge the gap to real information needs, several kinds of information processing techniques need to be leveraged. (1) For the issue of *noisy topics*, we need to discover topics from travelogues and further distinguish location-related topics with other noisy ones. (2) For the issue of *multiple viewpoints*, we need to find a representation of locations that summarizes all the useful descriptions of a location to capture its location-representative knowledge (i.e., local characteristics such as attractions, activities, styles). (3) To provide *destination recommendation*, a metric of relevance is necessary to suggest locations most relevant to tourists' travel intentions. (4) For *destination comparison*, a location similarity metric is necessary to compare locations from the perspective of travel. We believe that the first two points should be given primary importance because the location-representative knowledge mined from location-related topics underlies the ranking and similarity measurement of locations.

In this paper, we consider the above issues and investigate the problem of mining location-representative knowledge[1] from a large amount of travelogues to facilitate tourists to fully utilize such knowledge.

[1] Knowledge useful for tourists (e.g., accommodation, expense) but not specific to locations is beyond our objective in this paper.

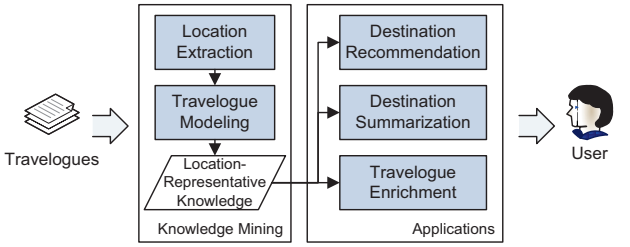


Figure 2. The overview of the proposed framework, in which location-representative knowledge is first mined from a travelogue corpus, and then used to support three applications.

In recent years, probabilistic topic models, such as latent Dirichlet allocation (LDA) [2], have been successfully applied to a variety of text mining tasks [3, 4, 13, 14, 16, 17, 19, 20]. This kind of models are suitable for the task of travelogue mining owing to its powerful capability of discovering latent topics from text and representing documents with such topics. However, to the best of our knowledge, the existing models are not applicable for our objective because none of them can be used to address the limitations of travelogue data. Specifically, although documents under these models are represented as mixtures of the discovered latent topics, the entities appearing in the documents (e.g., locations mentioned in travelogues) either lack of representation in the topic space, or are represented as mixtures of all the topics, rather than the topics appropriate to characterize these entities. Considering the noisy topics in travelogues, the representation of locations using all the topics would be contaminated by the noise and thus is unreliable for further relevance and similarity metrics.

Therefore, we propose a probabilistic topic model, i.e., Location-Topic (LT) model, to discover topics from travelogues and simultaneously represent locations with appropriate topics. Specifically, we define two different types of topics (as illustrated in Figure 1), i.e., *local topics* which characterize specific locations from the perspective of travel (e.g., *lava*, *coastline*), and *global topics* (e.g., *hotel*, *airport*) which do not characterize certain locations but rather extensively co-occur with various locations in travelogues. Since each local topic corresponds to some specific locations and travel-related characteristics, a location's overall characteristics can be generally represented in the local topic space, as a mixture of (i.e., a multinomial distribution over) local topics.

Based on the LT model, the aforementioned limitations of travelogue data are handled to some extent because:

- By decomposing travelogues into local and global topics, we can obtain location-representative knowledge from local topics, with other semantics captured by global topics filtered out.
- By representing each location using local topics mined from the entire travelogue collection, multiple viewpoints of each location can be naturally summarized.
- Based on the representation of locations in the local topic space, both the relevance of a location to a given travel intention and the similarity between locations can be measured.

As shown in Figure 2, given a collection of travelogues (in the implementation, either of two data sets: 100K English travelogues, or 94K Chinese ones), we first extract the locations mentioned in the text. Then a LT model is trained on the collection to learn local and global topics, as well as the representation of locations

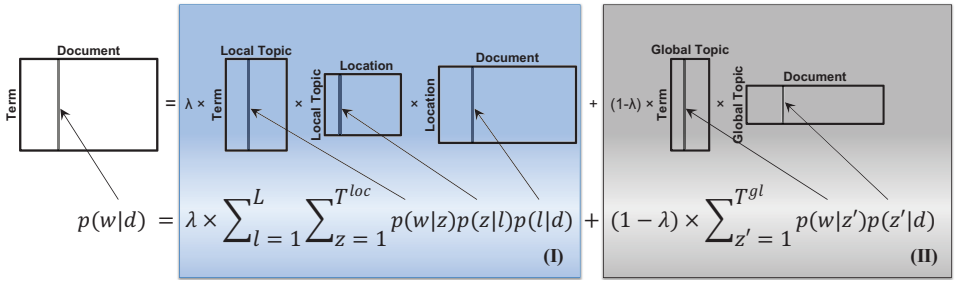


Figure 3. An illustration of the travelogue decomposition with (I) T^{loc} local topics and (II) T^{gl} global topics. It should be noted that this figure mainly serves as an illustrative interpretation of the ideas, but does not exactly accord with the model details.

in the local topic space. Based on the learnt knowledge, we can fulfill different application tasks. Specifically, we consider a scenario where a user learns online knowledge to plan a trip in three steps: 1) selecting a destination from some recommended ones, 2) browsing the characteristics of the selected destination to get an overview, and 3) browsing some travelogues to figure out detailed travel route. To facilitate these three steps, the following three applications are implemented, respectively:

- **Destination Recommendation**: We recommend destinations to users, in terms of either similarity to a given destination or relevance to a given travel intention.
- **Destination Summarization**: Each destination is presented as an overview by summarizing its representative aspects with textual tags. Representative snippets are also offered as further descriptions to verify and interpret the relation between each tag and the destination.
- **Travelogue Enrichment**: To help a user better browse and understand travelogues, we identify the informative parts of a travelogue and highlight them with related images.

The paper is organized as follows. In Section 2, we introduce the proposed Location-Topic model. Then we describe three applications of the LT model in Section 3. Experimental and evaluation results are shown in Section 4. Section 5 presents the related work. In Section 6, we give the conclusion and future work.

2. TRAVELOGUE MODELING

In this section, we present the Location-Topic (abbreviated as LT) model and its usage for further applications. By modeling the generative process of travelogues, the model could discover topics from travelogues and represent locations with the learnt topics.

2.1 Basic Idea

Following the existing work on probabilistic topic models, we treat each travelogue document as a mixture of topics, where each topic is a multinomial distribution over terms in vocabulary and corresponds to some specific semantics. As discussed in Section 1, we further assume that travelogues are composed of local and global topics, and each location is represented by a multinomial distribution over local topics. Thus, the proposed LT model aims at discovering local and global topics, as well as each location's distribution over local topics, from a travelogue collection.

We use Figure 3 to provide an illustrative and intuitive explanation how we decompose travelogue documents into local topics and global topics. A travelogue collection can be represented by a *Term-Document* matrix where the j^{th} column encodes the j^{th} document's distribution over terms, as illustrated at the top left of Figure 3. Based on this representation, our goal is equivalent to decomposing a given *Term-Document* matrix into multiple matrices, including *Term-LocalTopic*, *Term-GlobalTopic*, and *LocalTopic-Location* matrices. In Figure 3, there are another two matrices that we should learn, i.e., *GlobalTopic-Document* matrix and *Location-Document* matrix. The former is the same as that of common topic models, whereas the latter is specific and important to our objective, and thus need particular discussion.

For *Location-Document* matrix, we have some observed information, namely the user-provided location labels associated with each travelogue. However, such document-level labels are not fit for our scenario because they are usually too coarse and incomplete to support knowledge mining for all the locations described in travelogues, and sometimes they are even labeled incorrectly. Hence, we rely on the locations extracted from text instead of these labels. There are several methods for location extraction, e.g., looking up a gazetteer, or applying a Web service like *Yahoo Placemaker*[2]. As such pre-processing is not our focus, we employ a beforehand implemented location extractor based on a gazetteer and location disambiguation algorithms handling geographic hierarchy and context of locations, which can achieve high accuracy by considering all the candidate locations in a document simultaneously. We will detail this location extractor elsewhere.

Intuitively, the extracted locations can serve as strong indications of locations described in travelogues. However, these extracted locations are improper to be taken as the real *Location-Document* matrix, due to an observed gap between them and the locations actually described. For instance, a series of locations may be mentioned only as a trip summary, but without (or with quite unequal) descriptions in the contextual text. Besides, we also observe that in a typical travelogue, the author usually concentrates on depicting some locations in consecutive sentences. That is, consecutive words tend to correspond to the same locations. Considering these observations, we assume that all the words in a text segment (e.g., a document, paragraph, or sentence) share a multinomial distribution over locations, which is affected by a Dirichlet prior derived from the extracted locations in the segment. In this way, the *Location-Document* matrix is kept variable to better model the data, while also benefiting from the extracted locations as priors.

As the decomposition of likelihood $p(w|d)$ shown in Figure 3, each word in a document is assumed to be "written" by making a binary decision between two paths, i.e., (1) selecting a location, a local topic, and a term in sequence, and (2) selecting a global topic and a term in sequence. Once decomposed as above, a travelogue collection preserves its location-representative knowledge in *LocalTopic-Location* matrix, and topics in *Term-LocalTopic* and *Term-GlobalTopic* matrices.

[2] http://developer.yahoo.com/geo/placemaker/

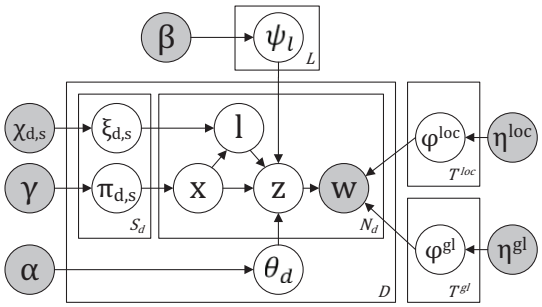


Figure 4. Graphical representation of the proposed LT model.

2.2 Generative Process of Travelogues

In the LT model, each location l is represented by ψ_l, a multinomial distribution over local topics, with symmetric Dirichlet prior β; each document d is associated with θ_d, a multinomial distribution over global topics, with symmetric Dirichlet prior α.

We extend the extensively used bag-of-words assumption to treat each document d as a set of S_d non-overlapping segments, where a segment could be a sentence, a paragraph, or a sliding window in the document. Each segment s is associated with (1) a bag-of-words, (2) a binomial distribution over global topics versus local topics $\pi_{d,s}$, with Beta prior $\gamma = \{\gamma^{gl}, \gamma^{loc}\}$, and (3) a multinomial distribution $\xi_{d,s}$ over segment s's corresponding location set

$$\mathcal{L}_{d,s} \stackrel{\text{def}}{=} \{l | l \text{ appears in segment } s \text{ in } d\},$$

with Dirichlet prior parameterized by $\chi_{d,s}$ defined as

$$\chi_{d,s} \stackrel{\text{def}}{=} \left\{\delta_{d,s,l} = \mu \cdot \#(l \text{ appears in segment } s \text{ in } d)\right\}_{l \in \mathcal{L}_{d,s}},$$

where "$\#(\cdot)$" is short for "the number of times" and coefficient μ denotes the precision of the prior. In the implementation, each paragraph in a travelogue is treated as a raw segment, with further merging to ensure that each segment contains at least one location.

The graphical representation of the LT model is shown in Figure 4. Accordingly, the generative process of a travelogue tion C, which consists of D documents covering L unique locations and W unique terms, is defined as follows:

- For each local topic $z \in \{1, \dots, T^{loc}\}$, draw a multinomial distribution over terms, $\varphi_z^{loc} \sim Dir(\eta^{loc})$.
- For each global topic $z \in \{1, \dots, T^{gl}\}$, draw a multinomial distribution over terms, $\varphi_z^{gl} \sim Dir(\eta^{gl})$.
- For each location $l \in \{1, \dots, L\}$, draw a multinomial distribution over local topics, $\psi_l \sim Dir(\beta)$.
- For each document $d \in \{1, \dots, D\}$:
  - Draw a multinomial distribution over global topics, $\theta_d \sim Dir(\alpha)$.
  - For each segment s of document d:
    - draw a binomial distribution over global topics versus local topics, $\pi_{d,s} \sim Beta(\gamma)$;
    - draw a multinomial distribution over locations in s, $\xi_{d,s} \sim Dir(\chi_{d,s})$.
  - For each word $w_{d,n}$ in segment s of document d:
    - draw a binary switch $x_{d,n} \sim Binomial(\pi_{d,s})$;
    - if $x_{d,n} = loc$, draw a location $l_{d,n} \sim Multinomial(\xi_{d,s})$, and then draw a local topic $z_{d,n} \sim Multinomial(\psi_{l_{d,n}})$;
    - if $x_{d,n} = gl$, draw a global topic $z_{d,n} \sim Multinomial(\theta_d)$;
    - draw word $w_{d,n} \sim Multinomial\left(\varphi_{z_{d,n}}^{x_{d,n}}\right)$.

2.3 Parameter Estimation

To estimate the parameters of the LT model, we need to estimate the latent variables conditioned on the observed variables, namely $p(\boldsymbol{x}, \boldsymbol{l}, \boldsymbol{z} | \boldsymbol{w}, \boldsymbol{\delta}, \alpha, \beta, \boldsymbol{\gamma}, \eta)$, where $\boldsymbol{x}, \boldsymbol{l}, \boldsymbol{z}$ are vectors of assignments of global/local binary switches, locations, and topics for all the words in travelogue collection C, respectively. We use the collapsed Gibbs sampling [7] with the following updating formulas.

For global topic $z \in \{1, \dots, T^{gl}\}$,

$$p\left(x_i = gl, z_i = z \middle| w_i = w, \boldsymbol{x}_{\setminus i}, \boldsymbol{z}_{\setminus i}, \boldsymbol{w}_{\setminus i}, \alpha, \boldsymbol{\gamma}, \eta^{gl}\right) \propto \frac{n_{w,\setminus i}^{gl,z} + \eta^{gl}}{\sum_{w'} n_{w',\setminus i}^{gl,z} + W\eta^{gl}} \cdot \frac{n_{d,\setminus i}^{gl,z} + \alpha}{n_{d,\setminus i}^{gl} + T^{gl}\alpha} \cdot \left(n_{d,s,\setminus i}^{gl} + \gamma^{gl}\right).$$

For local topic $z \in \{1, \dots, T^{loc}\}$ and location $l \in \mathcal{L}_{d,s}$,

$$p\left(x_i = loc, l_i = l, z_i = z \middle| w_i = w, \boldsymbol{x}_{\setminus i}, \boldsymbol{l}_{\setminus i}, \boldsymbol{z}_{\setminus i}, \boldsymbol{w}_{\setminus i}, \beta, \boldsymbol{\gamma}, \eta^{loc}\right) \propto \frac{n_{w,\setminus i}^{loc,z} + \eta^{loc}}{\sum_{w'} n_{w',\setminus i}^{loc,z} + W\eta^{loc}} \cdot \frac{n_{l,\setminus i}^{loc,z} + \beta}{n_{l,\setminus i} + T^{loc}\beta} \cdot \frac{n_{d,s,\setminus i}^{l} + \chi_{d,s,l}}{n_{d,s,\setminus i}^{loc} + \chi_{d,s}} \cdot \left(n_{d,s,\setminus i}^{loc} + \gamma^{loc}\right),$$

where $n_{w,\setminus i}^{gl,z}$ is the number of times term w is assigned to global topic z, and similarly $n_{w,\setminus i}^{loc,z}$ is that for local topic z. $n_{d,\setminus i}^{gl,z}$ is the number of times a word in document d is assigned to global topic z, while $n_{d,\setminus i}^{gl}$ is the number of times a word in document d is assigned to a global topic. $n_{l,\setminus i}^{loc,z}$ is the number of times a word assigned to location l is assigned to local topic z, out of $n_{l,\setminus i}$ words assigned to location l in total. $n_{d,s,\setminus i}^{l}$ is the number of times a word in segment s of document d is assigned to location l, and consequently to a local topic. $n_{d,s,\setminus i}^{gl}$ and $n_{d,s,\setminus i}^{loc}$ denote the number of times a word in segment s of document d is assigned to global and to local topics, respectively. For all the counts, subscript $\setminus i$ indicates that the i-th word is excluded from the computation.

After such a Gibbs sampler reaches burn-in, we can harvest several samples and count the assignments to estimate the parameters:

$$\varphi_{z,w}^{x} \propto n_w^{x,z} + \eta^x, x \in \{gl, loc\},\ z = 1, \dots, T^x,$$
$$\psi_{l,z} \propto n_l^{loc,z} + \beta,\ z = 1, \dots, T^{loc}.$$

2.4 Utilizing the Model

Once estimated, the parameters of the LT model can support several applications by providing the data representations and similarity metrics for both locations and terms.

2.4.1 Location Representation and Similarity Metric

Each location l can be represented in either the T^{loc}-dimensional local topic space or the W-dimensional term space. For the former, location l is simply represented by ψ_l namely its corresponding multinomial distribution over local topics. For the latter, we derive a probability distribution over terms conditioned on location l directly from the raw Gibbs samples, by counting the words assigned to location l, as

$$p(w|l) \propto n_l^w,\ w = 1, \dots, W,$$

where n_l^w is the number of times term w is assigned to location l.

According to the location representation in the local topic space, the symmetric similarity between two locations l_1 and l_2 is measured by the distance between their corresponding multinomial distributions over local topics ψ_{l_1} and ψ_{l_2}, as

$$LocSim(l_1, l_2) = exp\{-\tau D_{JS}(\psi_{l_1} \| \psi_{l_2})\},$$

where $D_{JS}(\cdot||\cdot)$ denotes the Jensen-Shannon (JS) divergence defined as $D_{JS}(p\|q)=\frac{1}{2}D_{KL}\left(p\left\|\frac{p+q}{2}\right.\right)+\frac{1}{2}D_{KL}\left(q\left\|\frac{p+q}{2}\right.\right)$, while $D_{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler (KL) divergence; coefficient $\tau>0$ is used to normalize different numbers of local topics.

2.4.2 Term Representation and Similarity Metric

In addition to that of locations, we also need a representation and corresponding similarity metric of terms, so as to measure the relevance of a location (or a snippet) to a given query term in the application of destination recommendation (or summarization). Hence, we expand each term w in the vocabulary into a probability distribution over the learnt T^{loc} local topics, denoted by δ_w, as

$$\begin{cases}\delta_w=\{p(z|w)\}_{z=1}^{T^{loc}}\\ p(z|w)\propto p(w|z)p(z)\propto\varphi_{z,w}^{loc}n_z^{loc}\end{cases},$$

where n_z^{loc} is the total number of words assigned to local topic z. Accordingly, the symmetric similarity between two terms w_1 and w_2 is measured based on their distributions over local topics, as

$$TermSim(w_1,w_2)=exp\{-\tau D_{JS}(\delta_{w_1}\|\delta_{w_2})\}.$$

2.4.3 Inference

Given the estimated parameters Ω, we can infer hidden variables for unseen travelogues. Specifically, a Gibbs sampler is run on the unseen document d using the following updating formulas:

$$p(x_i=gl,z_i=z|w_i=w,\boldsymbol{x}_{\backslash i},\boldsymbol{z}_{\backslash i};\Omega)$$
$$\propto\varphi_{z,w}^{gl}\cdot\frac{n_{d,\backslash i}^{gl,z}+\alpha}{n_{d,\backslash i}^{gl}+T^{gl}\alpha}\cdot\left(n_{d,s,\backslash i}^{gl}+\gamma^{gl}\right),z=1,\dots,T^{gl},$$
$$p(x_i=loc,l_i=l,z_i=z|w_i=w,\boldsymbol{x}_{\backslash i},\boldsymbol{l}_{\backslash i};\Omega)$$
$$\propto\varphi_{z,w}^{loc}\cdot\psi_{l,z}\cdot\frac{n_{d,s,\backslash i}^{l}+\chi_{d,s,l}}{n_{d,s,\backslash i}^{loc}+\chi_{d,s}}\cdot\left(n_{d,s,\backslash i}^{loc}+\gamma^{loc}\right),z=1,\dots,T^{loc}.$$

After collecting a number of samples, we can infer a distribution over locations for each term w appearing in document d by counting the number of times term w is assigned to each location l as

$$p(l|w)=\frac{\#(w\text{ appears in }d\text{ and is assigned to }l)}{\#(w\text{ appears in }d)}.$$

3. APPLICATIONS

In this section, we introduce how to leverage the learnt LT model to enable three interesting applications: *destination recommendation*, *destination summarization*, and *travelogue enrichment*.

3.1 Destination Recommendation

The first question raised by a tourist is: *where should I go*? Meanwhile, a tourist has some preferences about the travel destinations, which are usually expressed in terms of two criteria:

- Being similar to a given location
  - "*I quite enjoyed the trip to Honolulu last year. Is there any other destination with similar style?*"
- Being relevant to a given travel intention
  - "*I plan to go hiking next month. Could you recommend some destinations good for hiking?*"

3.1.1 Similarity-Oriented Recommendation

Given a query location l_q and a candidate destination set $\mathcal{L}$, each destination $l\in\mathcal{L}$ has a similarity to l_q in the local topic space, defined as $LocSim$ in Section 2.4.1. Besides, each destination has a query-independent popularity which should also be considered. The ranking score for recommendation is computed as

$$Score_{l_q}(l)=\log LocSim(l_q,l)+\mu\log Pop(l),\ l\in\mathcal{L},\mu\geq 0,$$

where coefficient μ controls the influence of the static popularity $Pop(l)$ in ranking. Here, $Pop(l)$ is simply defined as the occurrence frequency of location l in the travelogue collection C, as

$$Pop(l)=\frac{\#(l\text{ appears in }C)}{\sum_{l'\in\mathcal{L}}\#(l'\text{ appears in }C)}.$$

3.1.2 Relevance-Oriented Recommendation

Given a travel intention described by a term w_q (e.g., "*hiking*"), we rank destinations in terms of relevance to w_q. Since a travel intention usually contains more comprehensive semantics than a single term, we expand w_q in the local topic space as δ_{w_q} (a distribution over the local topics, as introduced in Section 2.4.2). In this way, the relevance of each location l to the w_q can be measured using KL-divergence. The ranking score is thus computed as

$$Score_{w_q}(l)=-D_{KL}\left(\delta_{w_q}\|\psi_l\right)+\nu\log Pop(l),\ l\in\mathcal{L},\nu\geq 0,$$

where ψ_l is location l's distribution over the local topics. Actually, with the above query expansion strategy, it is straightforward to support multi-word queries for more complex travel intentions.

3.2 Destination Summarization

Once a destination has been determined, a tourist would like to know more details of the destination, like

— "*What are the most representative things in San Francisco? Can you tell me with a few words or sentences?*"

To summarize the representative aspects of a destination, we first generate a few representative tags, and then identify related snippets for each tag to further describe and interpret the relation between the tag and the destination. For a given location l_q, we can obtain its probability distribution over terms $\{p(w|l_q)\}_{w=1:W}$ as described in Section 2.4.1, and simply select those terms with highest probabilities in this distribution as the representative tags. Then, given a representative tag w_q, we generate its corresponding snippets by ranking all the sentences $\{s\}$ in the travelogue collection according to the query "l_q+w_q". Specifically, a sentence s consisting of a (mentioned) location set $\mathcal{L}_s$ and a term set $\mathcal{W}_s$ is rated in terms of the *geographic relevance* to location l_q and the *semantic relevance* to tag w_q, as

$$Score_{l_q,w_q}(s)=GeoRele_{l_q}(s)\times SemRele_{w_q}(s),\text{ where}$$
$$GeoRele_{l_q}(s)=\#(l_q\text{ appears in }\mathcal{L}_s)/|\mathcal{L}_s|,\text{ and}$$
$$SemRele_{w_q}(s)=\sum_{w\in\mathcal{W}_s}TermSim(w_q,w)\ /\log(1+|\mathcal{W}_s|),$$

where $|\cdot|$ denotes the cardinality of a set, and $TermSim$ is the pair-wise term similarity defined in Section 2.4.2. Note that all the terms in sentence s contribute to the semantic relevance more or less, according to their similarities to the query tag.

3.3 Travelogue Enrichment

Besides the brief summarization, a tourist would also like to browse through some travelogues written by other tourists. Given a travelogue, a reader is usually interested in the places visited by the author and how these places look like.

— *"Where did Jack visit when he was in New York City? And how do those places look like?"*

To facilitate browsing, we extract the highlights of a travelogue and enrich them with images to provide additional visual descriptions. Given a travelogue d which refers to a set of locations $\mathcal{L}_d$,

we treat the informative depictions of locations in $\mathcal{L}_d$ as the highlights. As described in Section 2.4.3, each term w in travelogue d has a probability $p(l|w)$ to be assigned to location $l \in \mathcal{L}_d$. Hence, the highlight corresponding to location l is represented as a W-dimensional term-vector $\boldsymbol{u}_l = (u_{l,1}, \ldots, u_{l,W})$, where

$$u_{l,w} = \#(w \text{ appears in } d) \times p(l|w),\ w = 1, \ldots, W.$$

To visually enrich every identified highlight $\boldsymbol{u}_l$, we select images from a candidate image set $\mathcal{R}_l$ that is geographically relevant to location l. Each image $r \in \mathcal{R}_l$ is annotated with a set of tags $\mathcal{T}_r$, and is also represented as a W-dimensional vector $\boldsymbol{v}_r = (v_{r,1}, \ldots, v_{r,W})$, where

$$v_{r,w} = \sum_{t \in \mathcal{T}_r} TermSim(t, w),\ w = 1, \ldots, W.$$

Then, the relevance of image r to highlight $\boldsymbol{u}_l$ is computed as

$$Score_{\boldsymbol{u}_l}(r) = \langle \boldsymbol{u}_l, \boldsymbol{v}_r \rangle \cdot \frac{1}{\log(1+|\mathcal{T}_r|)},\ r \in \mathcal{R}_l,$$

where $\langle \cdot, \cdot \rangle$ denotes inner product, and the second term is used to normalize images with different numbers of tags. Moreover, to diversify the resulting images, we select images one by one. Once the k^{th} image r_k is chosen, we update $\boldsymbol{u}_l^{(k)} = (u_{l,1}^{(k)}, \ldots, u_{l,W}^{(k)})$ to decay the semantics already illustrated by the selected images, as

$$u_{l,w}^{(k)} = \begin{cases} u_{l,w}^{(k-1)} \times exp(-\tau \cdot v_{r_k,w}), & k \geq 1 \\ u_{l,w} & k = 0 \end{cases},\ w = 1, \ldots, W,$$

where $\tau > 0$ is a coefficient to control the strength of decay.

4. EXPERIMENTAL RESULTS

In this section, we present experimental results of the LT model and its applications. Both objective and subjective evaluation methods are used to evaluate the effectiveness of the framework.

4.1 Data Set

There are many sources of travelogues on the Web, either from Weblogs such as *Windows Live Spaces*, or dedicated travel websites like *TravelPod*[3], *IgoUgo*[4], and *TravelBlog*[5]. We collected approximately 100,000 travelogues written in English and related to tourist destinations in the United States, to form an English corpus. A location extractor was applied to extract locations mentioned in these travelogues, yielding 18,000 unique locations. As some subjective evaluations require participators' knowledge, we also built a Chinese corpus from *Ctrip*[6], consisting of 94,000 Chinese travelogues related to around 32,000 locations in China.

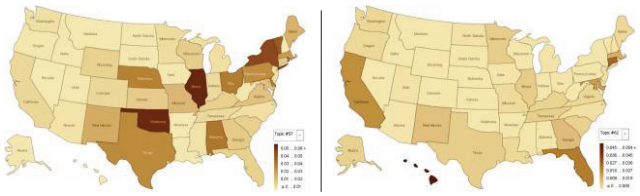
4.2 Travelogue Modeling

After pre-processing including stemming and stop-word removal, we trained a LT model on each corpus to learn a number of local topics and global topics. The numbers of local and global topics were set empirically to 100 and 50, respectively. The training procedure for each corpus included 2,000 iterations of Gibbs sampling and lasted for approximately 40 hours on a server with an AMD Opteron quad-core 2.4GHz processor.

To illustrate the topics learnt by the LT model, we show the top terms (i.e., terms with the highest probabilities) of some topics in Table 1. We can see that local topics characterize some tourism styles and corresponding locations, including both natural styles like *seaside* (local #62) and cultural styles like *museum* (local #57); whereas global topics correspond to common themes such as *accommodation* (global #26) and *opinion* (global #19), which tend to appear in travelogues related to almost any destination.

Table 1. Top terms of example local and global topics.

local #23	local #57	local #62	local #66	local #69
desert	museum	dive	casino	mountain
cactus	art	snorkel	gamble	peak
canyon	collect	fish	play	rocky
valley	gallery	aquarium	slot	snow
hot	exhibit	sea	table	high
west	paint	boat	machine	feet
heat	work	whale	game	lake
spring	sculpture	reef	card	summit
global #8	**global #19**	**global #22**	**global #26**	**global #37**
flight	great	kid	room	rain
airport	best	family	hotel	weather
fly	fun	old	bed	wind
plane	beautiful	children	inn	cold
check	enjoy	fun	breakfast	temperature
bag	wonderful	love	bathroom	storm
air	love	young	night	sun
travel	amaze	age	door	warm



(a) Local topic #57 (*museum*, …) (b) Local topic #62 (*dive*, …)

Figure 5. Geographic distributions of two local topics. The darker a region is, the higher correlation with the topic it has.

To exemplify the relationships between local topics and locations, we utilize, following [13], the *Many Eye* visualization service[7] to visualize the spatial distribution of some local topics. Based on the LT model, the correlation between a local topic z and a location l is measured by the conditional probability $p(z|l)$, which is equal to $\psi_{l,z}$ in the LT model. In Figure 5, we plot two local topics (#57 *museum* and #62 *seaside* in Table 1) on the U.S. state map respectively, where a darker state indicates a higher $p(z|l)$ it has. Both maps show uneven geographic distributions of local topics, indicating high dependence between local topics and locations. From Figure 5 (a) we see that *New York*, *Illinois*, and *Oklahoma* are famous for {*museum*, *art*, …}; while in Figure 5 (b) *Hawaii* shows the highest correlation with {*dive*, *snorkel*, …}. This demonstrates the learnt relationships between local topics and locations are reasonable and consistent with prior knowledge.

4.3 Destination Recommendation

4.3.1 Similarity-Oriented Recommendation

Since the effectiveness of similarity-oriented destination recommendation highly relies on the pair-wise similarity metric of locations, we directly evaluate this metric's capability of discovering similar locations from a given set.

[3] http://www.travelpod.com/
[4] http://www.igougo.com/
[5] http://www.travelblog.org/
[6] http://www.ctrip.com/
[7] http://manyeyes.alphaworks.ibm.com/manyeyes/

(a) LT model (b) Baseline method

Figure 6. Location similarity graphs generated by the LT model and the baseline method, where different colors and shapes stand for different location categories.

We first collected the top destinations recommended by *TripAdvisor*[8] for four travel intentions including *Beaches & Sun*, *Casinos*, *History & Culture*, and *Skiing*. After filtering out locations not appearing in our corpus, we built a location set consisting of 36 locations, based on which pair-wise location similarities were computed (as describe in Section 2.4.1) to form a location similarity graph. To demonstrate how well the graph is consistent with the ground-truth similarity/dissimilarity between four categories of locations, we use the *NetDraw*[9] software to visualize this graph where similar locations tend to be positioned close to each other, as shown in Figure 6 (a). As a comparison, we implemented a baseline method which formed a pseudo document for each location by concatenating all the travelogues referring to it, and then measured the pair-wise location similarity using the common TF-IDF-based cosine similarity. Comparing the two graphs in Figure 6 (a) and (b), we can see that different categories of locations are roughly differentiated by our similarity metric, while under the baseline metric some of them are coupled together. This is owing to one advantage of the LT model, namely preserving the information that characterizes and differentiates locations when projecting the travelogue data into a low-dimensional topic space.

4.3.2 Relevance-Oriented Recommendation

To evaluate the relevance-oriented recommendation, we collected the top destinations recommended by *TripAdvisor* for five travel intentions, i.e., *Beaches & Sun*, *Casinos*, *Family Fun*, *History & Culture* and *Skiing*, as the ground-truth for five queries, respectively. For the sake of uniformity, all the queries are truncated into unigrams. Besides the LT model-based method presented in Section 3.1.2, we also set up a baseline method, which ranks locations for a query term in decreasing number of travelogues containing both a location and the query term. The resulting location ranking lists of both methods are evaluated by the number of locations, within the top K ones, matching the ground-truth locations. The evaluation results are shown in Table 2, while Table 3 lists some top destinations recommended by our approach.

From Table 2 we observe that the locations recommended by the LT model generally match more ground-truth ones than the baseline; whereas the baseline exceeds our approach at the top 5 and top 10 results for the query *family*. This observation can be interpreted as the two sides of a coin. On one hand, our method measures each location's relevance to the query term in the local topic space to naturally expand the query with similar terms, and thus enable partial match and improve the relevance measurement for queries well captured by local topics (e.g., *beach*, *casino*). On the other hand, for queries mainly captured by global topics (e.g., *family*, a top term of the global topic #22 shown in Table 1), this query expansion mechanism is less reliable, due to the low confidence of these terms' distributions over local topics.

[8] http://www.tripadvisor.com/Inspiration/

[9] http://www.analytictech.com/Netdraw/netdraw.htm

Table 2. Comparison of the relevance-oriented destination recommendation results for five queries.

Query	#Ground-truth	Methods	#Matches at top K			
			K=5	K=10	K=15	K=20
beach	35	baseline	1	4	7	9
		LT model	4	9	12	13
casino	6	baseline	2	2	3	3
		LT model	4	5	5	5
family	38	baseline	4	6	8	11
		LT model	3	5	8	11
history	12	baseline	4	6	8	8
		LT model	5	8	9	10
skiing	20	baseline	2	4	4	6
		LT model	3	5	10	12

Table 3. Top destinations recommended by the LT model-based method, where those in the ground-truth shown in bold.

Query	Top 10 recommended destinations
beach	**Myrtle Beach**, **Maui**, Miami, **Santa Monica**, **Destin**, **Hilton Head Island**, **Virginia Beach**, **Daytona Beach**, **Key West**, **San Diego**
casino	**Las Vegas**, **Atlantic City**, Lake Tahoe**, Biloxi**, **Reno**, Deadwood, New Orleans, Detroit, **Tunica**, New York City
family	**Orlando**, Las Vegas, **New York City**, **Washington, D.C.**, New Orleans, Charleston, **Myrtle Beach**, **Chicago**, San Francisco, Walt Disney World
history	**New Orleans**, **Charleston**, **Williamsburg**, **Washington, D.C.**, **New York City**, Chicago, Las Vegas, **Philadelphia**, **San Francisco**, **San Antonio**
skiing	**Lake Tahoe**, **Park City**, South Lake Tahoe, Jackson Hole, **Vail**, **Breckenridge**, Winter Park, Salt Lake City, Beaver Creek, **Steamboat Springs**

4.4 Destination Summarization

4.4.1 Representative Tag Generation

To compare with the location-representative tag generation approach described in Section 3.2, we implemented three baseline methods. The first one ("TF") is to generate a pseudo document for each location by concatenating all the travelogue paragraphs referring to it, and then rank terms in decreasing frequency in the pseudo document. The second one ("TF-IDF") is to further multiply each term's frequency with the Inverse Document Frequency (IDF) to penalize common terms. The third baseline is similar to the LT model-based approach but disable the global topics.

As there is no existing ground-truth of location-representative tags, we built one by borrowing people's knowledge. For each location, we first formed a tag pool by merging the top tags generated by the proposed method and three baselines, and then asked 20 graduate students to select the top 10 most representative tags. Finally, each tag was rated according to the number of times it was selected, to generate a ranking list of tags as the ground-truth. Considering the participators' background knowledge, we used the Chinese corpus in this experiment and involved 20 popular tourist destinations in China to form the questionnaire.

Based on the ground-truth, the tag ranking list generated by each method is evaluated using the Normalized Discounted Cumulative

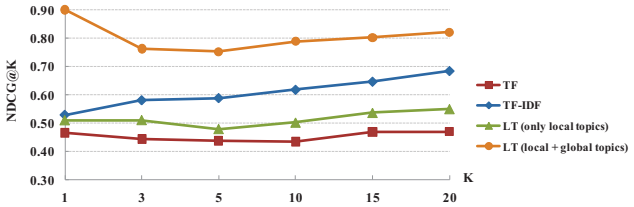


Figure 7. NDCG@K results of location-representative tags generated by (a) TF, (b) TF-IDF, (c) LT model with only local topics, and (d) LT model with both local and global topics.

Table 4. Representative tags generated by the LT model-based method for example destinations in the United States.

Destination	Top 10 representative tags
Anchorage	bear, moose, alaskan, glacier, fish, cruise, salmon, wildlife, trail, mountain
Boston	fenway, whale, historic, sox, cape, england, red, history, revere, church
Chicago	michigan, institute, field, lake, museum, cta, tower, loop, windy, cub
Las Vegas	strip, casino, show, hotel, bellagio, gamble, fountain, venetian, mgm, slot
Los Angeles	hollywood, star, studio, universal, movie, boulevard, theatre, china, getty, sunset
Maui	island, beach, snorkel, whale, ocean, luau, volcano, dive, fish, surf
New York City	subway, broadway, brooklyn, zero, avenue, island, yorker, manhattan, village, greenwich
Orlando	disney, park, universal, resort, world, theme, studio, kingdom, magic, epcot
San Francisco	bay, cable, alcatraz, chinatown, wharf, bridge, prison, bart, fisherman, pier
Washington, D.C.	museum, memorial, monument, national, metro, capitol, war, smithsonian, lincoln, president

Gain at top K (NDCG@K) [9], which is commonly used in the IR area to measure the accuracy of ranking results. The results averaged over all the 20 locations are shown in Figure 7. It can be seen that our method significantly outperforms the baselines consistently at top K ranking positions. Out of the baselines, the TF-IDF method outperforms the TF method consistently, owing to the penalty to noisy tags commonly co-occurring with various locations. However, this frequency-based penalty mechanism is too coarse to filter out all the noisy tags. Our approach properly filters out these tags using global topics. When global topics are disabled and all the information is modeled by local topics as in the third baseline, the performance is even worse than the TF-IDF method.

In addition to the above quantitative evaluation, we also generated representative tags for some U.S. destinations based on the English corpus. As exemplified in Table 4, the generated tags include not only landmarks (e.g., *bellagio*, *alcatraz*) but also styles (e.g., *historic*, *beach*) and activities (e.g., *gamble*, *dive*).

4.4.2 Representative Snippet Generation

As it is quite subjective to evaluate the extent to which a textual snippet is informative for *something at somewhere*, we resorted to user study to evaluate the generated representative snippets. Based on the Chinese corpus, we prepared 20 groups of data, each consisting of a query in the form of "location + term", 5 snippets generated by the proposed method, and another 5 snippets from a baseline snippet ranking method based on the number of occurrences of the query in a snippet.

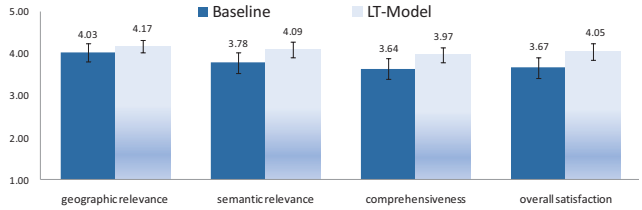


Figure 8. A subjective evaluation of representative snippets generated by the LT model-based method and the baseline.

Twenty graduate students were asked to assess the two snippet sets (presented in random order) in each group using 1 to 5 ratings, from four aspects namely (1) ***geographic relevance*** (i.e., to what extent the snippets are describing the query location), (2) ***semantic relevance*** (i.e., describing the query term), (3) ***comprehensiveness*** (i.e., providing rich information about the query), and (4) ***overall satisfaction***. Using these aspects we want to demonstrate whether the proposed method can suggest snippets not only relevant to the query but also informative and comprehensive. For each snippet set, we averaged all the users' evaluations as its ratings on the four aspects. The two methods are compared using pair-wise *t*-test on the 20 groups and exhibit significant differences ($p<0.01$) in all the four aspects. As depicted in Figure 8, although the difference in the geographic relevance is relatively small due to the straightforward measurement in both methods, our method shows significant advantages in other three aspects due to the query term expansion mechanism.

Besides, some examples generated based on the English corpus are illustrated in Table 5, where words relevant to the query term (shown in bold and italic) provide informative and comprehensive descriptions for the queries.

4.5 Travelogue Enrichment

For the evaluation of travelogue enrichment, we conducted a user study based on the Chinese corpus. The materials presented to users consist of 10 travelogue segments, each referring to at least one location and related characteristics. For each segment, there are two image sets (each with three images) generated by our method and a baseline method which simply uses the mentioned locations as queries to retrieve geo-tagged photos from *Flickr*.

We asked 20 graduate students to assess both image sets (presented in random order) of each segment from four aspects including (1) ***geographic relevance*** (i.e., to what extent the images are depicting the main locations in the segment), (2) ***semantic relevance*** (i.e., depicting the objects mentioned in the segment), (3) ***diversity*** (i.e., depicting different objects in the segment), and (4) ***overall satisfaction*** (well highlighting and enriching the text). The results are depicted in Figure 9, and pair-wise *t*-test on the 10 pairs of image sets exhibits significant differences ($p<0.01$) in all the four aspects. It indicates that the images selected by our method are more favorable compared with the baseline. Note that in the aspect of geographic relevance, the difference of two methods is small because the baseline is actually geo-based, while in other three aspects, our method exhibits larger advantages due to the learnt location-representative knowledge and query expansion enabled by the LT model.

Another two English examples are illustrated in Figure 10, where each travelogue segment is enriched by three images that depict

Table 5. Representative snippets generated by the LT model-based method for several "location + term" queries, where locations are shown in underline and words relevant to the query term are shown in bold and italic.

Query		Top snippets
Alaska + wildlife	1	This time, the bus stopped at the Alaska ***Wildlife*** Conservation Center, where we saw musk oxen, ***bison***, ***brown bears***, black ***bears***, ***eagles***, ***owls***, ***foxes***, reindeers, and a few other ***animals***.
	2	Along the way we stopped at the Alaska ***Wildlife*** Recreation Center to check on the musk ox, ***Moose***, ***Brown*** and Black ***Bears***, and many other Arctic ***animals***.
	3	The rest of my week in Alaska was filled with more ***wildlife*** sightings, including numerous ***brown bears*** and cubs, caribou, and a ***rare*** black ***wolf*** in Denali; and a ***moose*** with a calf and several ***wild*** trumpeter swans.
Las Vegas + casino	1	With the ringing of ***slot machines*** it finally hit me, I was ***rolling*** into a Las Vegas ***casino***! Sitting at the first row of ***slot machines*** we experienced one of the most amazing aspects of common ***casino*** etiquette.
	2	Being more inclined to stick my ***hard earned cash*** in a ***money*** market fund than a ***slot machine***, Las Vegas and ***casinos*** have never held much allure for me, but I have to admit that the Las Vegas Strip with all the big ***casinos*** seemed pretty glamorous and exciting.
	3	We have ***played*** in the ***Casinos*** of Las Vegas! They want you to stay in the ***Casinos*** so they have ***waitresses*** coming around giving you free ***drinks*** and all you have to do is tip her $1 or $2 dollars and she just keeps coming back.
Waikiki Beach + beach	1	The famous Waikiki Beach is very ***beautiful*** and is packed with ***swimmers*** from the early hours, it has lovely ***clean sand*** but the ***beach*** is getting smaller by the year and is already at the base of a couple of hotels! The dreaded global ***warming***!
	2	Since Waikiki Beach faces southwest, we were able to ***enjoy*** an extraordinary ***sunset*** right from the ***beach*** in front of our hotel, then we retired to our room where we spent the evening sitting out on the lanai, having some cocktails, and listening to the ***surf***.
	3	If you are near the Waikiki Beach area, ***enjoy*** the day at the ***beach***, ***relax***, and stay for the ***beautiful sunset***.

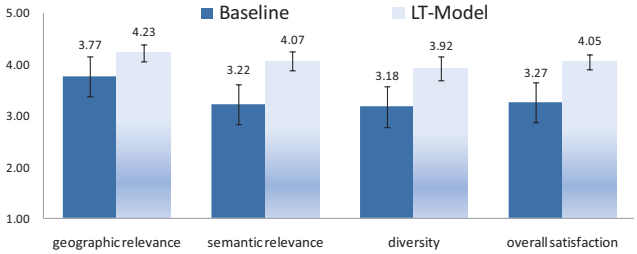


Figure 9. A subjective evaluation of travelogue enrichment by the LT model-based method and the baseline method.

its most informative parts. We also present each image's original tags and the words in text it corresponds to. For instance, the presented images in Figure 10 (a) depict representative and diverse semantics described in the text, i.e., *ocean*, *volcano*, and *beach*.

5. RELATED WORK

Some related work has been dedicated to organizing information on the Web to provide online travel assistant services. For instance, Jing et al. [10] proposed a travel plan assistant system which provided high-quality images relevant to given locations based on tourist sight extraction and image retrieval. Wu et al. [21] proposed a system to generate personalized tourism summary in the form of text, image, video, and news. In [12], a trip planning system was presented for place recommendation according to users' previous choices and tag-based place similarity.

Recently, leveraging user-contributed photo collections (e.g., *Flickr* [6]) has attracted lots of research efforts. Some of them [11][18] selected representative photos to visually summarize a given landmark or scene. Ahern et al. [1] analyzed the tags associated with photos to identify and visualize representative tags for arbitrary areas in the world; while Crandall et al. [5] utilized geo-tagged photos to discover worldwide popular places and their representative images. In [22], geo-tagged photos were leveraged to discover landmarks and build a world-scale landmark recognition engine. Moxley et al. [15] proposed an image tag suggestion tool based on mining of location tags from *Flickr* photos.

In [8], the authors proposed to generate overviews for locations by mining representative tags from travelogues and retrieving related images. Each travelogue is assumed to be related to only one location; neither similarity between locations nor the representation of locations in the learnt topic space is considered.

Probabilistic topic models, such as latent Dirichlet allocation (LDA) [2] and its extensions, have been successfully applied to many text mining tasks. Rosen-Zvi et al. [17] extended LDA by incorporating authors of documents as observed variables and representing authors with mixtures of topics. Some models [4][19] aimed to discover topics in different granularity levels other than document-level. In [16][20], locations (entities) appearing in documents were explicitly modeled as generated by topics, while in [13][14] locations served as labels associated with documents. In a very recent work [3], the model was sensitive to both entities and relationships between entities, given textual data segmented beforehand. In spite of the success in their respective scenarios, all the above models are not applicable to the travelogue mining scenario in this paper, as discussed in the section of Introduction.

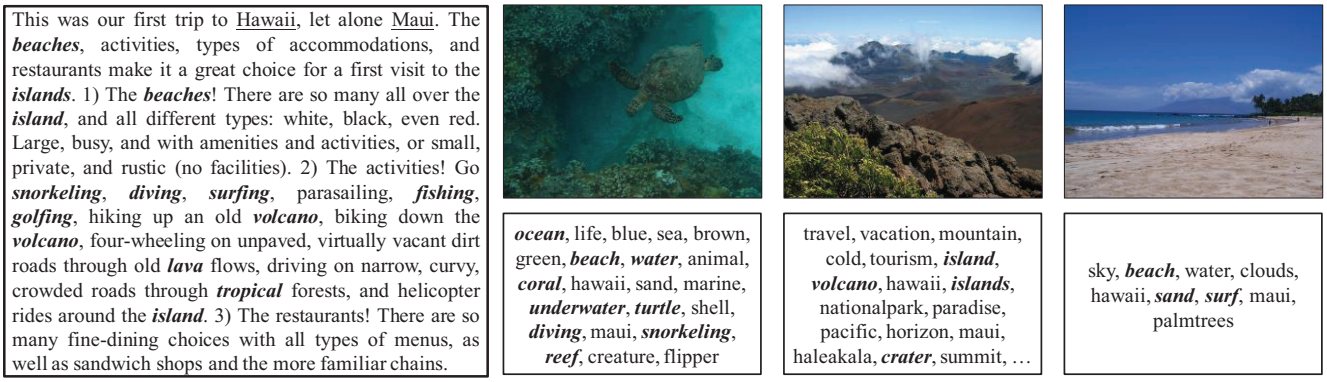
6. CONCLUSION AND FUTURE WORK

Travelogues contain abundant location-representative knowledge, which is informative for other tourists, but difficult to extract and summarize manually. In this paper, we have investigated the mining of location-representative knowledge from travelogues to facilitate tourists to utilize such knowledge. We proposed a probabilistic topic model, i.e., Location-Topic model, to discover local and global topics from travelogues and characterize locations using local topics. With this model, we could effectively (1) recommend destinations for flexible queries; (2) summarize destinations with representative tags and snippets; and (3) enrich the highlights of travelogues with images. The proposed framework was evaluated based on two large travelogue collections, showing promising results on the above tasks.

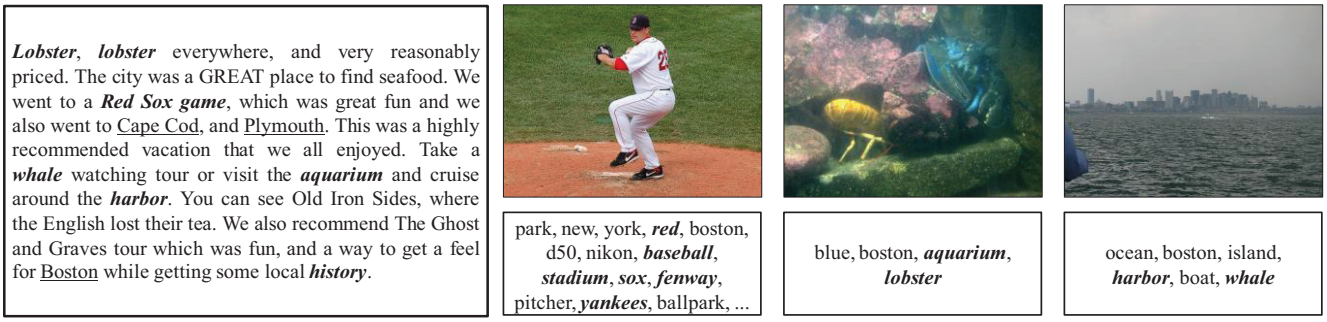
For the future work, we plan to incorporate prior knowledge of locations to improve the unsupervised knowledge mining. Another direction is to leverage more types of information in travelogues (e.g., opinions, travel routes, and temporal information) to meet more practical information needs such as itinerary planning.

7. REFERENCES

[1] S. Ahern, M. Naaman, R. Nair, and J. Yang. World Explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *Proc. JCDL*, 2007.



(a) A segment of a Maui travelogue titled as "Our Maiden Journey to Magical Maui"
(http://www.igougo.com/journal-j23321-Maui-Our_Maiden_Journey_to_Magical_Maui.html)



(b) A segment of a Boston travelogue titled as "The home of the Lobster"
(http://www.igougo.com/journal-j16396-Boston-The_home_of_the_Lobster.html)

Figure 10. Two example travelogue segments visually enriched by the proposed method, where each image's original tags are shown below the image; locations are shown in underline and informative words/tags are shown in bold and italic.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993-1022, 2003.
[3] J. Chang, J. Boyd-Graber, and D. M. Blei. Connections between the lines: augmenting social networks with text. In *Proc. KDD*, 2009.
[4] C. Chemudugunta, P. Smyth, and M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *Proc. NIPS*, 2006.
[5] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the World's Photos. In *Proc. WWW*, 2009.
[6] Flickr. http://www.flickr.com/
[7] T. Griffiths and M. Steyvers. Finding scientific topics. In *PNAS*, 101:5228–5235, 2004.
[8] Q. Hao, R. Cai, X.-J. Wang, J.-M. Yang, Y. Pang, and L. Zhang. Generating location overviews with images and tags by mining user-generated travelogues. In *Proc. ACM Multimedia*, 2009.
[9] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proc. SIGIR*, 2000.
[10] F. Jing, L. Zhang, and W.-Y. Ma. VirtualTour: an online travel assistant based on high quality images. In *Proc. ACM Multimedia*, 2006.
[11] L. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *Proc. WWW*, 2008.
[12] J. Kim, H. Kim, and J. Ryu. TripTip: a trip planning service with tag-based recommendation. In *Proc. CHI*, 2009.
[13] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *Proc. WWW*, 2008.
[14] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proc. WWW*, 2006.
[15] E. Moxley, J. Kleban, and B. S. Manjunath. SpiritTagger: a geo-aware tag suggestion tool mined from Flickr. In *Proc. MIR*, 2008.
[16] D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers. Statistical entity-topic models. In *Proceedings of KDD*, 2006.
[17] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proc. UAI*, 2004.
[18] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. In *Proc. ICCV*, 2007.
[19] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *Proc. WWW*, 2008.
[20] C. Wang, J. Wang, X. Xie, W.-Y. Ma. Mining geographic knowledge using location aware topic model. In *Proc. GIR*, 2007.
[21] X. Wu, J. Li, Y. Zhang, S. Tang, and S.-Y. Neo. Personalized multimedia web summarizer for tourist. In *Proc. WWW*, 2008.
[22] Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: building a web-scale landmark recognition engine. In *Proc. CVPR*, 2009.