Topic mining of tourist attractions based on a seasonal context aware LDA model

Topic mining of tourist attractions based on a seasonal context aware LDA model

Chao Huang[a,*], Qing Wang[a], Donghui Yang[a], Feifei Xu[b]

[a]*Department of Management Science and Engineering, School of Economics and Management, Southeast University, Jiangsu, Nanjing 210096, China*
[b]*Tourism Department, School of Humanities, Southeast University, Jiangsu, Nanjing 210096, China*

Abstract

With the rise of personalized travel recommendation in recent years, automatic analysis and summary of the tourist attraction is of great importance in decision making for both tourists and tour operators. To this end, many probabilistic topic models have been proposed for feature extraction of tourist attraction. However, existing state-of-the-art probabilistic topic models overlook the fact that tourist attractions tend to have distinct characteristics with respect to specific seasonal context. In this article, we contribute the innovative idea of using seasonal contextual information to refine the characteristics of tourist attractions. Along this line, we first propose STLDA, a season topic model based on latent Dirichlet allocation which can capture meaningful topics corresponding to various seasonal contexts for each attraction from a collection of attraction description documents. Then, an inference algorithm using Gibbs sampling is put forward to learn the posterior distributions and model parameters of our proposed model. In order to verify the effectiveness of STLDA model, we present a detailed experimental study using collected real-world textual data of tourist attractions. The experimental analysis results show that the superiority of STLDA over the basic LDA model in detecting the season-dependent topics and giving a representative and comprehensive summarization related to each tourist attraction. More importantly, it has great significance for improving the level of personalized attraction recommendation service.
Keywords: Probabilistic generative model, topic detection, contextual information, attraction recommendation

1. Introduction

With the rapid development of tourism market, the demand for intelligent travel services has been expected to increase remarkably. The prevalence of the Internet enables everyone to easily access travel related information from various websites. However, the sustained growth of travel data on the web may be overwhelming for tourists when selecting tourist attractions that specific to their personalized requirements. Meanwhile, tour op-

*Corresponding author. Tel: +86 138 1406 9012
Email address: huangchao@seu.edu.cn (Chao Huang)

erators need to present customized tourist attractions for potential tourists so as to survive in competitive market and make more profit. As an effective tool to achieve precision marketing for tour operators and assist decision marking for tourists, the personalized recommendation technique has attracted a great deal of attention over the past few years. Personalized attraction recommendation focuses on identifying the most relevant attractions to recommend to tourists, where the content-based method is popularly used in this case since this method cater well to tourists' needs. The content-based attraction recommendation approach aims to maximize the relevance between the tourists' preferences and attractions' features. A critical challenge along this line is to get a comprehensive understanding of the characteristics of tourist attractions. Therefore, it is highly desirable to produce a precise analysis and summary of online attraction information, with the objective of providing decision support for both tourists and tour operators.

Topic detection and extraction is a well-studied research [1–4] that aims at identifying a group of words that form topics from a collection of documents. Thematic analysis has been actively investigated in feature extraction of tourist attraction and gradually become an important attraction profiling technique in recent years. Topic-based feature analysis for a given attraction facilitate users and tour operators to capture the high-level concepts that reveal representative and comprehensive attributes of a tourist attraction, which is beneficial for further attraction selection or tourism planning. For instance, Pang et al.[5] conducted a topic segmentation for popular attractions in the United States by employing topics extracted from the user-generated travelogues on the web. In Yeh and Cheng's study [6], the popular tourist attractions in Taiwan were segmented into nine subject categories including natural, museum, heritage, park, animal, religious site, shopping, nightlife and visitor center on the basis of properties of attractions. In Hao et al.'s study [7], tourist destinations mentioned in travelogues on the travel websites were characterized by topics such as desert, museum, seaside and mountain, which are mined from these travelogues. Another related work is Hao et al.[8], in which the authors proposed to generate overviews for locations by mining representative topic tags from travelogues. Topic detection is also applied in Shen et al.'s study [9], where the topic features of tourist attractions were mined from user comments on travel websites and then matched with tourists' preferences to generate personalized attraction recommendation for them.

Probabilistic topic models have been proposed for topic extraction from textual data and successfully applied to a series of text mining tasks in different research fields over the past decade, owing to their powerful capability of discovering meaningful latent topics from large collection of documents automatically and simultaneously representing documents with these discovered topics. Topic models are usually based upon the

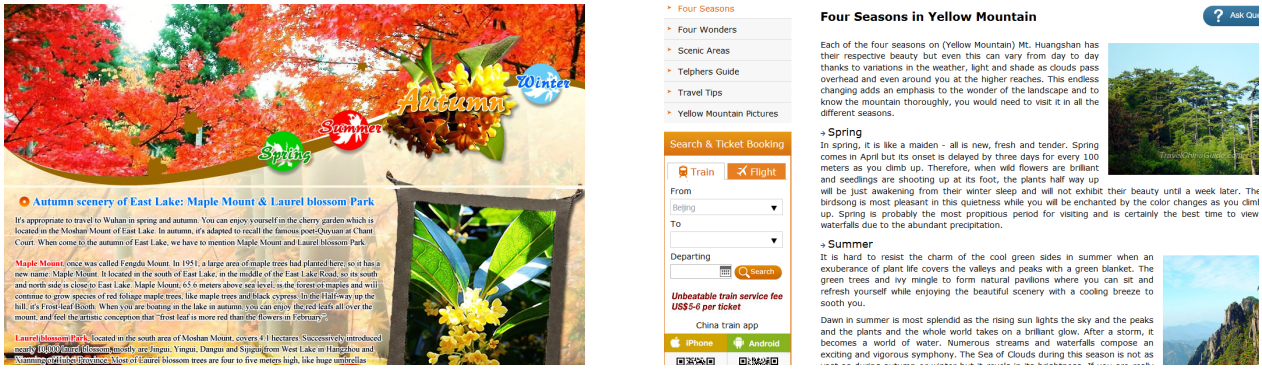
assumption that documents are mixture of topics, where each topic is a probability distribution over words. Early Explorations of topic modeling technique include latent semantic analysis (LSA) model [10], probabilistic latent semantic analysis (PLSA) model [11] and their varieties, where PLSA model is a useful step toward probabilistic modeling of text. Latent Dirichlet Allocation (LDA) model [12] was first proposed by David Blei and is considered to be one of the most popular topic models for its better probability statistical foundation. LDA is a well-defined generative probabilistic model that generalizes easily to new documents and improve PLSA by introducing Dirichlet priors on the model parameters, which overcomes the overfitting problem suffered in PLSA. Since LDA model can accurately extract tourism topic preferences of users as well as topic features of attractions from travel related information, it has attracted extensive attention from researchers in the field of personalized travel recommendation over the past few years. For example, Arbelaitz et al.[13] employed LDA to extract topics with respect to interests of tourists from user generated content on the travel websites, which aimd to promote a destination for tourists. Hao et al.[7] proposed a location-topic model based on LDA to mine local topics that characterize locations from a large collection of travel logs, and further to recommend the travel destinations on the basis of tourists' travel intentions. In Jiang et al.'s research [14], the topics about user preference were extracted from the textual description of photos on social media to model users by leveraging an expanded model of LDA, then personalized attraction recommendation was performed accordingly. In Shen et al.'s study [9], LDA was introduced to obtain topic and topic probability distribution of each attraction on the basis of a collection of user comments crawled from travel websites, then the similarities between attractions were measured for further attraction recommendation.

Recently, a promising research direction in topic modeling is to include contextual information with the aim of detecting latent topics that can reflect the effect of varying contexts. Incorporating additional contextual information into topic models in the field of personalized travel recommendation can better identify the topic features regarding user preferences and attraction characteristics, which can be used in decision support tasks that are context dependent. In terms of personalized travel recommendation, time is an essential factor of contextual information. Tourists' preferences and requirements may vary over time, leading to the changes in travel behavior [15–17]. Meanwhile, tourist attractions tend to have distinct characteristics with respect to specific time context [18]. To this end, several studies have attempted to link time information to topic models. For example, Wang and McCallum [19] presented a probabilistic topic model with consideration of the document's timestamp that explicitly models time jointly with word co-occurrence patterns, which aimed at extracting a probability distribution over continuous time for each topic. Blei and Lafferty [20] proposed a

dynamic topic model based on LDA to capture the evolution of topics in a long period from a large document collections that sequentially organized. In Lu's study [21], Probit-Dirichlet hybrid allocation topic model was developed by including temporal features of documents to detect the cyclical topic dynamics that reflect users' habits in the user generated content, which can be further used to recommend products for users exposed at specific contexts. Liu et al. [22] developed a probabilistic topic model by incorporating location and time information, which can extract the topics of each travel package corresponding to its suitable travel time for following personalized travel package recommendation.

Despite recent progressions, these time-dependent topic models are mainly focus on the long-term evolution of topics in a whole corpus, while the topics of each document in the corpus remain constant. Specifically, their research usually based upon the assumption that each document in the corpus is associated with one timestape and all documents are collected over time. Then these topic models are applied to the document collections that sequentially organized to discover time sensitive topics. However, the hypothesis is oversimplified because one document may exhibit the feature of more than one time period. In the case of topic extraction for tourist attractions, the attraction textual data is significantly different from other common documents since the content of an attraction description text often reveal a strong seasonal pattern, which is an intrinsic feature of the attraction and should be considered as important contextual information with respect to this attraction. In order to clearly illustrate the seasonal characteristics existing in attractions description documents, Fig. 1 shows snapshots of two famous tourist attractions in China. Fig. 1(a) is the description text of East Lake Scenic Area from its official website (http://www.whdonghu.gov.cn/english.htm) and Fig. 1(b) is the description document of Yellow Mountains from TravelChinaGuide (www.travelchinaguide.com/attraction/anhui/huangshan/seasons.htm). From these figures, it can be observed that both tourist attraction descriptions have distinct seasonal features. Besides, the description texts corresponding to different seasons for the same attractions show remarkable difference. It's apparently that none of the above mentioned topic models are applicable to deal with such unique attraction textual data because they may confound topics with respect to different time contexts in one document. Hence, it is necessary to develop a suitable approach to address the unique characteristics of the attraction textual data and precisely extract the topic features of tourist attractions with consideration of seasonal contextual information. However, to the best of our knowledge, so far no research has focused on this topic.

To fill this gap, we present a novel probabilistic topic model to detect meaningful topics corresponding to various seasonal contexts for each attraction from a collection of attraction description documents. The proposed Season Topic model based on LDA (STLDA) is a generative probability model, which can capture the potential



(a) Official website of East Lake Scenic Area (b) TravelChinaGuide website of Yellow Mountain

Fig. 1. Two snapshots that illustrate the seasonal characteristics in the description documents of attractions.

season-dependent topic clusters that naturally occurring in attractions documents. As a generative model, our learned topic model is substantially the joint probability distribution of seasonal contextual information as well as textual data, which specifies a probabilistic process to describe how words in attractions documents might be generated in particular when the seasonal feature in each attraction document is taken into account. By including seasonal contextual information, STLDA can model the variations of topic occurrence that reveal the changing seasonal contexts, which is unable to capture using other probabilistic topic models. As a result, our proposed model can detect the representative and comprehensive attributes corresponding to various seasonal contexts for each attraction and well represent the content of each attraction description document.

The rest of this paper is organized as follows. Section 2 is devoted to the methods including the basic LDA model and the proposed STLDA model. In Section 3, an inference algorithm using Gibbs sampling for the parameter estimation of our proposed model is discussed in detail. Section 4 illustrates the experimental results and analysis. Finally, Section 5 includes our conclusions.

2. Methodology

2.1. LDA Model

Latent Dirichlet Allocation (LDA) [12] is a generative probabilistic model that tries to capture the implicit topic structure from a collection of documents. It specifies a probabilistic procedure that depicts how the words in documents are generated. The basic idea is that each document is represented by a specific topic distribution and each topic is characterized by a probability distribution over words. The LDA model is a three-level hierarchical Bayesian model, where topics are associated with documents and words are associated with topics. There is a clear hierarchy followed by the document layer, topic layer and word layer.

1.Word layer: A word is the basic unit of discrete data, defined to be an item from a vocabulary of size V denoted by $V = \{w_1, w_2, \ldots, w_V\}$.

2.Topic layer: A topic $z_k, k \in \{1, 2, \cdots, K\}$ is associated with a multinomial φ_k over the V-word vocabulary and can be denoted by $\varphi_k = \langle p_{k,1}, p_{k,2}, \ldots, p_{k,V} \rangle$, where $p_{k,j}$ refers to the probability that word w_j is generated from topic z_k.

3.Document layer: A document is a sequence of N_m words denoted by $d_m = \{w_1, w_2, \ldots, w_{N_m}\}$. Likewise, each document is associated with a multinomial θ_m over K topics and can be represented as $\theta_m = \langle p_{m,1}, p_{m,2}, \ldots, p_{m,K} \rangle$, where $p_{m,z}$ refers to the probability that topic z is generated from document d_m.

Fig. 2(a) shows the graphical model representation of the LDA. In this graphical notation, nodes are random variables and arrows indicate conditional dependencies between two variables. The shaded and unshaded circles represent observed and latent variables respectively, while boxes refer to repeated sampling with the number of samples in the lower right corner of the boxes. It is well known that the Dirichlet distribution is the conjugate prior of the multinomial distribution. Therefore, a Dirichlet prior with parameter α for document-topic multinomial distribution θ_m and a Dirichlet prior with parameter β for topic-word multinomial distribution φ_k are chosen respectively. Given a corpus consisting of M documents, LDA makes assumptions that each word w is connected with a latent topic z. Each of topic $z_k, k \in \{1, 2, \cdots, K\}$ is related to a multinomial distribution φ_k defined on the V-word vocabulary, and each φ_k is chosen from a Dirichlet prior distribution with parameter β. Similarly, each document d_m is defined as a multinomial distribution θ_m over topics, drawn from a Dirichlet prior distribution with parameter α. The full generative process for each document d_m in a corpus is defined as follows:

1. For each topic $z_k, k \in \{1, 2, \cdots, K\}$
   a. Draw a topic-word multinomial distribution $\varphi_k \sim Dirichlet(\beta)$
2. For each document $d_m, m \in \{1, 2, \cdots, M\}$
   a. Draw a document-topic multinomial distribution $\theta_m \sim Dirichlet(\alpha)$
   b. For each word $w_{m,n}, n \in \{1, 2, \cdots, N_m\}$ in document d_m
      i. Draw a topic $z_{m,n} \sim Multinomial(\theta_m)$
      ii. Draw a word $w_{m,n} \sim Multinomial(\varphi_{z_{m,n}})$

Given the parameters α and β, the joint distribution over the random variables $(\mathbf{w_m}, \mathbf{z_m}, \varphi_k, \theta_m)$ then can be

derived from Fig. 2(a), which is given by:

$$p(\mathbf{w_m}, \mathbf{z_m}, \theta_m, \varphi_k | \alpha, \beta) = p(\theta_m|\alpha) p(\varphi_k|\beta) \prod_{n=1}^{N_m} p(z_{m,n}|\theta_m) p(w_{m,n}|\varphi_{z_{m,n}}) \tag{1}$$

Integrating over θ_m and φ_k, summing over $z_{m,n}$, the marginal distribution of a document can be obtained:

$$p(\mathbf{w_m}|\alpha, \beta) = \int_{\theta_m} \int_{\varphi_k} p(\theta_m|\alpha) p(\varphi_k|\beta) \Big(\prod_{n=1}^{N_m} \sum_{z_{m,n}} p(z_{m,n}|\theta_m) p(w_{m,n}|\varphi_{z_{m,n}}) \Big) d\theta_m \, d\varphi_k \tag{2}$$

Finally, taking the product of the marginal probability of every document in the corpus, the generative probability of a corpus is defined as follows:

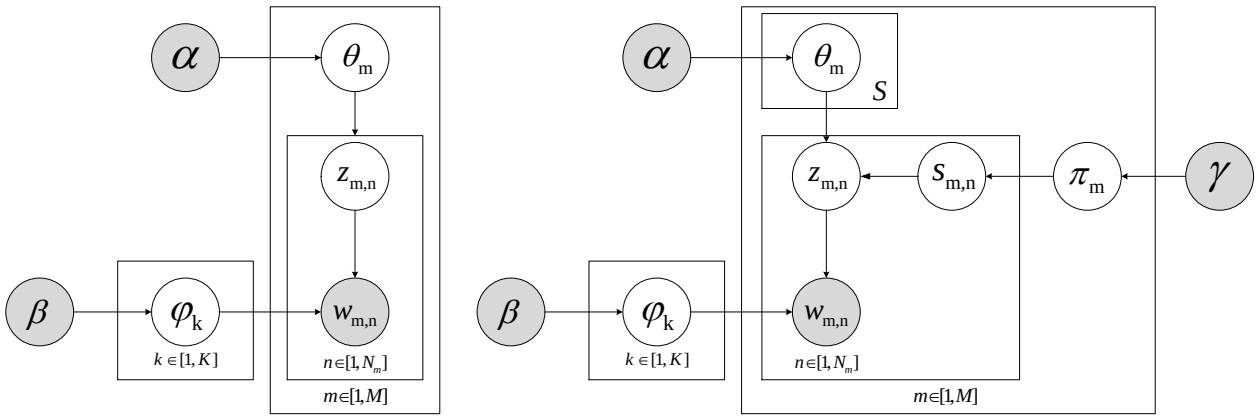
$$p(D|\alpha, \beta) = \prod_{m=1}^{M} \int_{\theta_m} \int_{\varphi_k} p(\theta_m|\alpha) p(\varphi_k|\beta) \Big(\prod_{n=1}^{N_m} \sum_{z_m} p(z_{m,n}|\theta_m) p(w_{m,n}|\varphi_{z_{m,n}}) \Big) d\theta_m \, d\varphi_k \tag{3}$$

In LDA, there are two sets of parameters that need to be estimated from a collection of documents, one is the topic distribution in each document and the other is the word distribution in each topic. In reality, only the documents can be observed, while the topic structure including topics and topic probability proportions is hidden. The key issue of LDA model is to use the observed documents to infer the latent topic structure. Therefore, some statistical approaches have been fully utilized for inferring the latent variables that can generate the observed collection of documents best. The exact inference for posterior estimation is intractable in general, thus a wide variety of approximate inference algorithms are considered for LDA, including Expectation-Maximization [23], Gibbs Sampling [24, 25] and Variational approximation [26].

2.2. STLDA: a new probabilistic model for attractions

STLDA is a novel probabilistic topic model with the aim to extract topics from a collection of attraction documents by taking advantage of information of documents as well as the intrinsic seasonal characteristic in each document. STLDA is an expanded model of LDA by adding an additional season layer between the document layer and the topic layer. Therefore, STLDA is a four-level hierarchical Bayesian model, where seasonal features are correlated with documents, under which topics are associated with seasonal characteristics and words are related to topics.

While the generative process of STLDA has the similarity to a certain extent with some topic models in the text modeling domain, such as Topic-Aspect model [27], Topic-Link LDA model [28] and Author-Topic model [29], the logical structures of these models are totally different. For example, the Author-Topic model introduces



(a) LDA (b) STLDA

Fig. 2. The graphical model for the LDA and STLDA.

two hyper-parameters that try to model the content of documents and the interests of authors, thus it only have two sets of latent variables that need to be estimated and it is still a three-level hierarchical Bayesian model in nature. In Topic-Aspect model, the authors decompose the generative process of words into background model and aspect model, then use a binary switching variable to determine if the word is the common background word that appear independently of a document's topical content or topical word that associated with a topic. Similarly, the Topic-Link LDA model introduces a binary variable to model a link between two documents with the aim to identify a set of high-level topics covered by the documents in the collection as well as the social network of the authors of the documents. The STLDA model has a crucial enhancement that can clearly identify the meaningful topics corresponding to various seasonal contexts for each tourist attraction. As a result, the tourist attractions are described more comprehensively and precisely on the season level of fine-grained, which can benefit the further analysis. In this paper, by using the intrinsic seasonal characteristic in each tourist attraction, we assume that the words in attraction documents have distinct seasonal tendencies. The STLDA model is represented as a probabilistic graphical model in Fig. 2(b).

Assume that we have a corpus with a collection of M documents denoted by $\{d_1, d_2, \ldots, d_M\}$, each document in the corpus is a sequence of N_m words represented by $d_m = (w_1, w_2, \ldots, w_{N_m})$, and each word in the document is an entity from a vocabulary with V distinct words denoted by $\{w_1, w_2, \ldots, w_V\}$. The number of season segments is S and the total number of topics is K. In our probability generative model, we assume that each word w is related to one of latent topics z, just like the model of LDA does. Each of topic $z_k, k \in \{1, 2, \cdots, K\}$ is defined as a multinomial distribution φ_k over the V-word vocabulary, and each φ_k is chosen from a Dirichlet prior distribution with parameter β. Each document d_m is modeled by S different multinomial distribution $\theta_{m,s}$

over the K topics with respect to different season labels $s, s \in \{1,2,\cdots,S\}$, all drawn from a Dirichlet prior distribution with parameter α, which significantly distinguish STLDA from the original LDA model that each document d_m is defined as just one multinomial distribution θ_m over the K topics. Besides, another distribution π_m is defined for each document d_m, $m \in \{1,2,\cdots,M\}$ over the S season segments, drawn from a Dirichlet prior distribution with parameter γ. The process for generating a word $w_{m,n}$ in document d_m under STLDA has three steps. Frist, a season label s is choosen from the document's specific season distribution π_m. Then a topic is sampled from the topic distribution $\theta_{m,s}$ conditioned on both the document and the intrinsic seasonal feature of the attraction document. Finally, a word is drawn from distribution over words defined by the topic. The notations of STLDA model to be used throughout the paper are summarized with brief descriptions in Table 1. The full generative process of STLDA model for each document d_m in a corpus is defined as follows:

1. For each topic z_k, $k \in \{1,2,\cdots,K\}$
   a. Draw a topic-word multinomial distribution $\varphi_k \sim Dirichlet(\beta)$
2. For each document d_m, $m \in \{1,2,\cdots,M\}$
   a. Draw a document-season multinomial distribution $\pi_m \sim Dirichlet(\gamma)$
   b. For each season label s, $s \in \{1,2,\cdots,S\}$ under document d_m
      i. Draw a document-season-topic multinomial distribution $\theta_{m,s} \sim Dirichlet(\alpha)$
3. For each word $w_{m,n}$, $n \in \{1,2,\cdots,N_m\}$ in document d_m
   a. Draw a season label $s_{m,n} \sim Multinomial(\pi_m)$
   b. Draw a topic $z_{m,n} \sim Multinomial(\theta_{m,s_{m,n}})$
   c. Draw a word $w_{m,n} \sim Multinomial(\varphi_{z_{m,n}})$

As previously mentioned, STLDA considers both general description of attraction and seasonal features exisiting in attraction document in a unified manner and can detect the meaningful topics with respect to different seasons for each attraction. Fig. 3 is a running example of STLDA model. As can be seen from this figure, there is a clear hierarchy followed by the attraction document layer, season layer, topic layer and word layer. The words constitute a number of topics and the tourist attraction corresponds to various topics in different seasons, where the weights labeled in the corresponding edges indicate the topic occurrence probabilities. For example for the attraction in spring, the detected topics are T7 with probability value 0.635 and T20 with probability value 0.208, while in winter the attraction corresponds to topics T16, T4 and T26 and the probability values are 0.613, 0.184 and 0.136 respectively.

Now, the likelihood function for the observed tourism attractions textual data can be formulated according

Table 1
Notations used in this paper

Concept	Notation	Description
Data	M	the number of attraction documents in the corpus D
	V	the number of distinct words that appear at least once in the corpus D
	d_m	the bag-of-words in document m
	D	the set of d_m for all $m \in \{1,2,\cdots,M\}$
	$w_{m,n}$	the nth word in document m, which is an item from the vocabulary
STLDA	K	the number of topics
	S	the number of season labels
	$z_{m,n}$	the topic index of nth word in document m
	$s_{m,n}$	the season label of nth word in document m
	$\theta_{m,s}$	a multinomial distribution over topics specific to document m and season label s
	Θ	the set of $\theta_{m,s}$ for all $m \in \{1,2,\cdots,M\}$ $s \in \{1,2,\cdots,S\}$
	φ_k	a multinomial distribution that represents the relevance of words in V for the kth topic
	Φ	the set of φ_k for all $k \in \{1,2,\cdots,K\}$
	π_m	a multinomial distribution over season labels for document m
	Π	the set of π_m for all $m \in \{1,2,\cdots,M\}$
	α	Dirichlet prior for Θ, where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$
	β	Dirichlet prior for Φ, where $\beta = (\beta_1, \beta_2, \ldots, \beta_V)$
	γ	Dirichlet prior for Π, where $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_S)$
Model inference	$\mathbf{w}_{-\mathbf{i}}$	vector of all words that appear in corpus excluding word w_i
	$\mathbf{z}_{-\mathbf{i}}$	vector of topic assignments for all words in corpus except word w_i
	$\mathbf{s}_{-\mathbf{i}}$	season labels vector for all words in corpus excluding word w_i
	$n_k^{(t)}$	the count of word t assigned to topic k in corpus
	$n_{m,j}^{(k)}$	the count of words assigned to topic k and season label j in document m
	$n_m^{(j)}$	the count of words assigned to season label j in document m
	Γ	gamma function

to our proposed probability generative model. Given the hyperparameters α, β and γ, the joint distribution over the random variables $(\mathbf{w_m}, \mathbf{z_m}, \mathbf{s_m}, \varphi_k, \theta_{m,s}, \pi_m)$ can be derived from Fig. 2(b), which is given by:

$$
\begin{aligned}
&p(\mathbf{w_m}, \mathbf{z_m}, \mathbf{s_m}, \varphi_k, \theta_{m,s}, \pi_m | \alpha, \beta, \gamma) \\
&= \overbrace{\prod_{n=1}^{N_m} \underbrace{p(w_{m,n}|\varphi_{z_{m,n}}) p(z_{m,n}|\theta_{m,s_{m,n}}) p(s_{m,n}|\pi_m)}_{word\,plate} \underbrace{p(\pi_m|\gamma)}_{season plate} p(\theta_{m,s}|\alpha) \underbrace{p(\varphi_k|\beta)}_{topic plate}}^{document\,plate}
\end{aligned}
\tag{4}
$$

where $\theta_{m,s}$, φ_k and π_m denote model parameters to be estimated, $\mathbf{z_m}$ and $\mathbf{s_m}$ are hidden variables, $\mathbf{w_m}$ refer to known variable. α, β and γ are the Dirichlet priors for $\theta_{m,s}$, φ_k and π_m respectively. Furthermore, the clear hierarchy struture of the model is also maked in the above formula.

By integrating out the distributions of $\theta_{m,s}$, φ_k and π_m and summing over $z_{m,n}$ and $s_{m,n}$, the likelihood of a

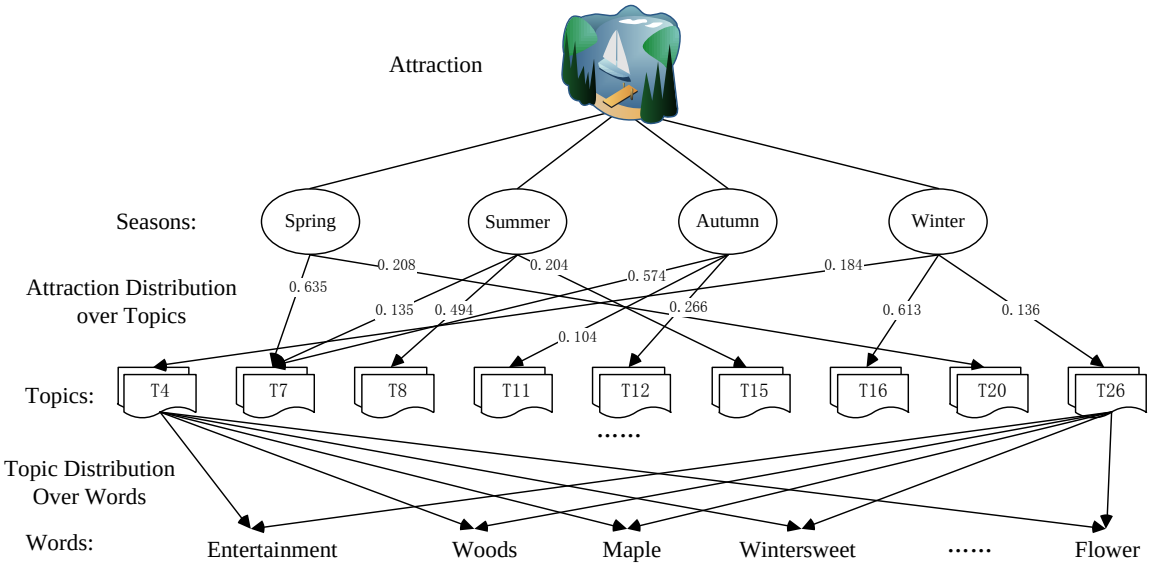


Fig. 3. Example of STLDA model.

document $\mathbf{w_m}$ can be obtained:

$$
\begin{aligned}
p(\mathbf{w_m}|\alpha,\beta,\gamma) = & \int_{\Theta}\int_{\Phi}\int_{\Pi} p(\theta_{m,s}|\alpha)p(\varphi_k|\beta)p(\pi_m|\gamma) \\
& \cdot \prod_{n=1}^{N_m}\sum_{z_{m,n}}\sum_{s_{m,n}} p(w_{m,n}|\varphi_{z_{m,n}})p(z_{m,n}|\theta_{m,s_{m,n}})p(s_{m,n}|\pi_m)d\Theta\, d\Phi\, d\Pi
\end{aligned}
\tag{5}
$$

where Θ, Φ and Π denote the set of $\theta_{m,s}$ for all $m \in \{1,2,\cdots,M\}\, s \in \{1,2,\cdots,S\}$, the set of φ_k for all $k \in \{1,2,\cdots,K\}$ and the set of π_m for all $m \in \{1,2,\cdots,M\}$ respectively.

Finally, the likelihood of the complete corpus D is determined by the product of the likelihoods of the independent documents:

$$
\begin{aligned}
p(D|\alpha,\beta,\gamma) = & \prod_{m=1}^{M}\int_{\Theta}\int_{\Phi}\int_{\Pi} p(\theta_{m,s}|\alpha)p(\varphi_k|\beta)p(\pi_m|\gamma) \\
& \cdot \prod_{n=1}^{N_m}\sum_{z_{m,n}}\sum_{s_{m,n}} p(w_{m,n}|\varphi_{z_{m,n}})p(z_{m,n}|\theta_{m,s_{m,n}})p(s_{m,n}|\pi_m)d\Theta\, d\Phi\, d\Pi
\end{aligned}
\tag{6}
$$

There are three sets of latent variables that need to be estimated in our model, including: the topic distribution of the corresponding per document-season pair $\theta_{m,s}$, the per topic-word multinomial distribution φ_k and the per document-season multinomial distribution π_m.

3. Model inference and parameter estimation

3.1. Model inference based on Gibbs sampling

Given a collection of documents, the posterior distribution of the latent variables including $\theta_{m,s}$, φ_k and π_m can be computed when the likelihood function of the whole corpus is maximized. The procedure of variable inference is to invert the document generative process and infer the latent variables from the observed corpus. While the exact inference for posterior estimation is intractable in general, a wide variety of approximate inference algorithms are considered to estimate these parameters. In this paper, we adopt the Gibbs sampling algorithm to perform approximate inference. Gibbs sampling is a simple and widely applicable Markov chain Monte Carlo algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult [30]. In Gibbs sampling, each parameter is sequentially sampled according to the full conditional distribution of the parameter conditioned on the observed data and all the other parameters. Since this method can get global optimum solution in theory and can be easily implemented while other inference algorithms such as variational EM algorithm is only guaranteed to find a local optimum [31], it is commonly used in Bayesian inference domain, especially for topic modeling.

In the case of our model, the target of inference is the posterior distribution of the hidden variables $\mathbf{z}$ and $\mathbf{s}$, which is defined as follows:

$$p(\mathbf{z},\mathbf{s}|\mathbf{w}) = \frac{p(\mathbf{z},\mathbf{s},\mathbf{w})}{p(\mathbf{w})} = \frac{\prod_{i=1}^{W} p(z_i,s_i,w_i)}{\prod_{i=1}^{W}\sum_{k=1}^{K}\sum_{j=1}^{S} p(z_i=k,s_i=j,w_i)} \tag{7}$$

where W is the number of all the words that appear in corpus, $\mathbf{z}$ and $\mathbf{s}$ are vector of topics that appear with the words in corpus and season labels that assigned to the words of corpus respectively. Moreover, the hyperparameters are omitted for simplicity. This posterior distribution covers a large space of discrete random variables, causing intractable in exact computation. At this point, the Gibbs sampling procedure comes into play. In our setting, the desired Gibbs sampler runs a Markov Chain that exploits the full conditional distribution $p(z_i,s_i|\mathbf{z}_{-\mathbf{i}},\mathbf{s}_{-\mathbf{i}},\mathbf{w})$ so as to simulate the posterior distribution $p(\mathbf{z},\mathbf{s}|\mathbf{w})$, where the subscript $-i$ refers to a quantity that excludes word w_i from the observed data.

Specifically, in our model, the full conditional probability distribution for a word w_i based on the topics and season labels of all other variables as well as the observed data is defined as follows:

$$p(z_i,s_i|\mathbf{w},\mathbf{z}_{-\mathbf{i}},\mathbf{s}_{-\mathbf{i}},\alpha,\beta,\gamma) = \frac{p(w_i,z_i,s_i|\mathbf{w}_{-\mathbf{i}},\mathbf{z}_{-\mathbf{i}},\mathbf{s}_{-\mathbf{i}},\alpha,\beta,\gamma)}{p(w_i|\mathbf{w}_{-\mathbf{i}},\mathbf{z}_{-\mathbf{i}},\mathbf{s}_{-\mathbf{i}},\alpha,\beta,\gamma)} \tag{8}$$

where $\mathbf{w}_{-\mathbf{i}}$, $\mathbf{z}_{-\mathbf{i}}$ and $\mathbf{s}_{-\mathbf{i}}$ represent vector of all words that appear in corpus excluding word w_i, vector of topic assignments for all words except word w_i and season labels vector for all words in corpus excluding word w_i respectively.

To apply a Gibbs sampling algorithm, the joint probability distribution of the observed words, topics and season labels assignments of the whole corpus is first derived by dividing this joint distribution into three parts, which is given by:

$$p(\mathbf{w},\mathbf{z},\mathbf{s}|\alpha,\beta,\gamma) = p(\mathbf{w}|\mathbf{z},\beta)p(\mathbf{z}|\mathbf{s},\alpha)p(\mathbf{s}|\gamma) \tag{9}$$

since the first part $p(\mathbf{w}|\mathbf{z},\beta)$ in the above equation is independent of α and γ, the second part $p(\mathbf{z}|\mathbf{s},\alpha)$ is irrelevant to β and γ, and the third part $p(\mathbf{s}|\gamma)$ is unrelated to α and β, these elements of the joint probability distribution can now be handled separately.

The first part $p(\mathbf{w}|\mathbf{z},\beta)$ can be derived from a multinomial on the observed word counts given the associated topics:

$$p(\mathbf{w}|\mathbf{z},\Phi) = \prod_{i=1}^{W} p(w_i|z_i) = \prod_{i=1}^{W} \varphi_{z_i,w_i} \tag{10}$$

where Φ indicates the set of φ_k for all $k \in \{1,2,\cdots,K\}$ and φ_{z_i,w_i} refers to the probability that word w_i is generated from topic z_i. This formula shows that the W words of the corpus are observed according to independent multinomial trials with parameters conditioned on the corresponding topic indice z_i. By splitting the product over words into one product over topics and one over the vocabulary, we obtain the following equation:

$$p(\mathbf{w}|\mathbf{z},\Phi) = \prod_{k=1}^{K}\prod_{t=1}^{V} p(w_i = t|z_i = k) = \prod_{k=1}^{K}\prod_{t=1}^{V} \varphi_{k,t}^{n_k^{(t)}} \tag{11}$$

where $n_k^{(t)}$ denotes the count of word t assigned to topic k in corpus.

Then, the target distribution $p(\mathbf{w}|\mathbf{z},\beta)$ can be obtained by integrating over Φ:

$$\begin{aligned} p(\mathbf{w}|\mathbf{z},\beta) &= \int_{\Phi} p(\mathbf{w}|\mathbf{z},\Phi)p(\Phi|\beta)d\Phi \\ &= \Big(\frac{\Gamma(\sum_{t=1}^{V}\beta_t)}{\prod_{t=1}^{V}\Gamma(\beta_t)}\Big)^K \prod_{k=1}^{K} \frac{\prod_{t=1}^{V}\Gamma(\beta_t + n_k^{(t)})}{\Gamma(\sum_{t=1}^{V}\beta_t + n_k^{(t)})} \end{aligned} \tag{12}$$

where $\beta = (\beta_1,\beta_2,\ldots,\beta_V)$ denotes the hyperparameters for topic-word multinomial distribution φ_k and Γ refers to the gamma function.

Analogous to $p(\mathbf{w}|\mathbf{z},\beta)$, the second part of the joint probability distribution $p(\mathbf{z}|\mathbf{s},\alpha)$ can be derived from

the following formula:

$$p(\mathbf{z}|\mathbf{s},\Theta)=\prod_{i=1}^{W}p(z_i|s_i,d_i)=\prod_{m=1}^{M}\prod_{k=1}^{K}\prod_{j=1}^{S}p(z_i=k|s_i=j,d_i=m)=\prod_{m=1}^{M}\prod_{k=1}^{K}\prod_{j=1}^{S}\theta_{m,j,k}^{n_{m,j}^{(k)}} \tag{13}$$

where the notation d_i refers to a document in the collection of $\{d_1,d_2,\ldots,d_M\}$ that the word w_i belongs to, $\theta_{m,j,k}$ denotes the probability of topic k appeared with season label j in document m and $n_{m,j}^{(k)}$ is the count of words assigned to topic k and season label j in document m. By integrating out Θ, the second element $p(\mathbf{z}|\mathbf{s},\alpha)$ can be formulated as follows:

$$\begin{aligned}p(\mathbf{z}|\mathbf{s},\alpha)&=\int_{\Theta}p(\mathbf{z}|\mathbf{s},\Theta)p(\Theta|\alpha)d\Theta\\&=\Big(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\Big)^{M*S}\prod_{m=1}^{M}\prod_{j=1}^{S}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+n_{m,j}^{(k)})}{\Gamma(\sum_{k=1}^{K}\alpha_k+n_{m,j}^{(k)})}\end{aligned} \tag{14}$$

where $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_K)$ is the hyperparameters for document-season-topic multinomial distribution $\theta_{m,j}$.

Likewise, the third part $p(\mathbf{s}|\gamma)$ can also be obtained by integrating out Π:

$$\begin{aligned}p(\mathbf{s}|\gamma)&=\int_{\Pi}p(\mathbf{s}|\pi)p(\pi|\gamma)d\Pi\\&=\Big(\frac{\Gamma(\sum_{j=1}^{S}\gamma_j)}{\prod_{j=1}^{S}\Gamma(\gamma_j)}\Big)^{M}\prod_{m=1}^{M}\frac{\prod_{j=1}^{S}\Gamma(\gamma_j+n_m^{(j)})}{\Gamma(\sum_{j=1}^{S}\gamma_j+n_m^{(j)})}\end{aligned} \tag{15}$$

where $\gamma=(\gamma_1,\gamma_2,\ldots,\gamma_S)$ refers to the hyperparameters for document-season multinomial distribution π_m, $\pi_{m,j}$ and $n_m^{(j)}$ indicate the probability of season label j associated with document m and the count of words assigned to season label j in document m respectively. Therefore, the joint probability distribution of the observed words, topics and season labels assignments of the whole corpus will be obtained by multiplying the results of equation (12), (14) and (15) according to formula (9).

After deriving the joint probability distribution, the topic assignment z_i and season label assignment s_i for each word w_i in the corpus can be sampled with the following full conditional probability distribution:

$$\begin{aligned}p(z_i=k,s_i=j|\mathbf{w},\mathbf{z}_{-\mathbf{i}},\mathbf{s}_{-\mathbf{i}},\alpha,\beta,\gamma)&=\frac{p(\mathbf{w},\mathbf{z},\mathbf{s},\alpha,\beta,\gamma)}{p(\mathbf{w}_{-\mathbf{i}},w_i,\mathbf{z}_{-\mathbf{i}},\mathbf{s}_{-\mathbf{i}},\alpha,\beta,\gamma)}\\&=\frac{p(\mathbf{w},\mathbf{z},\mathbf{s},\alpha,\beta,\gamma)}{p(\mathbf{w}_{-\mathbf{i}},\mathbf{z}_{-\mathbf{i}},\mathbf{s}_{-\mathbf{i}},\alpha,\beta,\gamma)p(w_i|\mathbf{w}_{-\mathbf{i}},\mathbf{z}_{-\mathbf{i}},\mathbf{s}_{-\mathbf{i}},\alpha,\beta,\gamma)}\\&\propto\frac{p(\mathbf{w},\mathbf{z},\mathbf{s}|\alpha,\beta,\gamma)}{p(\mathbf{w}_{-\mathbf{i}},\mathbf{z}_{-\mathbf{i}},\mathbf{s}_{-\mathbf{i}}|\alpha,\beta,\gamma)}\end{aligned} \tag{16}$$

The above formula can then be factored, which is shown as follows:

$$\begin{aligned}\frac{p(\mathbf{w},\mathbf{z},\mathbf{s}|\alpha,\beta,\gamma)}{p(\mathbf{w}_{-\mathbf{i}},\mathbf{z}_{-\mathbf{i}},\mathbf{s}_{-\mathbf{i}}|\alpha,\beta,\gamma)} &= \frac{\left(\frac{\Gamma(\sum_{t=1}^{V}\beta_t)}{\prod_{t=1}^{V}\Gamma(\beta_t)}\right)^{K}\prod_{k=1}^{K}\frac{\prod_{t=1}^{V}\Gamma(\beta_t+n_k^{(t)})}{\Gamma(\sum_{t=1}^{V}\beta_t+n_k^{(t)})}}{\left(\frac{\Gamma(\sum_{t=1}^{V}\beta_t)}{\prod_{t=1}^{V}\Gamma(\beta_t)}\right)^{K}\prod_{k=1}^{K}\frac{\prod_{t=1}^{V}\Gamma(\beta_t+n_{k,-i}^{(t)})}{\Gamma(\sum_{t=1}^{V}\beta_t+n_{k,-i}^{(t)})}}\cdot\frac{\left(\frac{\Gamma(\sum_{j=1}^{S}\gamma_j)}{\prod_{j=1}^{S}\Gamma(\gamma_j)}\right)^{M}\prod_{m=1}^{M}\frac{\prod_{j=1}^{S}\Gamma(\gamma_j+n_m^{(j)})}{\Gamma(\sum_{j=1}^{S}\gamma_j+n_m^{(j)})}}{\left(\frac{\Gamma(\sum_{j=1}^{S}\gamma_j)}{\prod_{j=1}^{S}\Gamma(\gamma_j)}\right)^{M}\prod_{m=1}^{M}\frac{\prod_{j=1}^{S}\Gamma(\gamma_j+n_{m,-i}^{(j)})}{\Gamma(\sum_{j=1}^{S}\gamma_j+n_{m,-i}^{(j)})}} \\ &\cdot\frac{\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)^{M*S}\prod_{m=1}^{M}\prod_{j=1}^{S}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+n_{m,j}^{(k)})}{\Gamma(\sum_{k=1}^{K}\alpha_k+n_{m,j}^{(k)})}}{\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)^{M*S}\prod_{m=1}^{M}\prod_{j=1}^{S}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+n_{m,j,-i}^{(k)})}{\Gamma(\sum_{k=1}^{K}\alpha_k+n_{m,j,-i}^{(k)})}} \\ &= \frac{\alpha_k+n_{m,j}^{(k)}-1}{\sum_{k=1}^{K}\alpha_k+n_{m,j,-i}^{(k)}}\frac{\beta_t+n_k^{(t)}-1}{\sum_{t=1}^{V}\beta_t+n_{k,-i}^{(t)}}\frac{\gamma_j+n_m^{(j)}-1}{\sum_{j=1}^{S}\gamma_j+n_{m,-i}^{(j)}}\end{aligned} \tag{17}$$

Noting that, the above derivations exploited one important property of Gamma function $\Gamma(n+1)=n\Gamma(n)$.

Finally, the full conditional probability distribution for each variable z_i and s_i can be equally separated from the joint sampling distribution of topic assignment and sesaon label indicator for word w_i, which are given by:

$$p(z_i=k|\mathbf{w},\mathbf{z}_{-\mathbf{i}},\mathbf{s},\alpha,\beta,\gamma)\propto\left(\alpha_k+n_{m,j}^{(k)}-1\right)\frac{\beta_t+n_k^{(t)}-1}{\sum_{t=1}^{V}\beta_t+n_{k,-i}^{(t)}} \tag{18}$$

and

$$p(s_i=j|\mathbf{w},\mathbf{s}_{-\mathbf{i}},\mathbf{z},\alpha,\beta,\gamma)\propto\left(\gamma_j+n_m^{(j)}-1\right)\frac{\alpha_k+n_{m,j}^{(k)}-1}{\sum_{k=1}^{K}\alpha_k+n_{m,j,-i}^{(k)}} \tag{19}$$

3.2. parameter estimation

Gibbs sampling will serially draw each variable of z_i and s_i from the full conditional distribution of each variable conditioned on remaining variables and the observations according to equation (18) and equation (19). The algorithm iterates repeatedly until the Markov chain has reached a stationary state, where the stationary distribution of the Markov chain is our desired posterior distribution. According to Gibbs sampling rule, a bunch of draws that are approximately from our desired posterior distribution are obtained once the Markov chain has converged. Then, the model parameters Θ, Φ and Π corresponding to the stationary state of the Markov chain can be estimated based on these samples. The pseudocode of Gibbs sampling procedure for our model is presented in Algorithm 1, which runs over the three periods: initialisation period, burn-in period and sampling period.

By applying Bayes' rule, the multinomial distributions $\theta_{m,j}$, φ_k and π_m with their Dirichlet prior α, β and γ

can be redefined as follows:

$$p(\theta_{m,j}|D,\alpha) = Dirichlet(\theta_{m,j}|\mathbf{n_{m,j}} + \alpha) \tag{20}$$

$$p(\varphi_k|D,\beta) = Dirichlet(\varphi_k|\mathbf{n_k} + \beta) \tag{21}$$

$$p(\pi_m|D,\gamma) = Dirichlet(\pi_m|\mathbf{n_m} + \gamma) \tag{22}$$

where D denotes the observed corpus, $\mathbf{n_{m,j}}$, $\mathbf{n_k}$ and $\mathbf{n_m}$ refer to vector of topic observation counts for document m and season label j, term observation counts for topic k and season label observation counts for document m respectively. Specifically, $\mathbf{n_{m,j}}$, $\mathbf{n_k}$ and $\mathbf{n_m}$ are denoted by $(n_{m,j}^{(1)}, n_{m,j}^{(2)}, \ldots, n_{m,j}^{(K)})$, $(n_k^{(1)}, n_k^{(2)}, \ldots, n_k^{(V)})$ and $(n_m^{(1)}, n_m^{(2)}, \ldots, n_m^{(S)})$ respectively.

The probability of topic k asscociated with season label j in document m now can be approximated by using the expectation of Dirichlet distribution, which is given by:

$$\theta_{m,j,k} = \frac{n_{m,j}^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_{m,j}^{(k)} + \alpha_k} \tag{23}$$

For the remaining parameters φ_k and π_m using the same method, the approximate probability of word t in a topic k is formulated as follows:

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t} \tag{24}$$

The approximate probability of season label j in a document m is presented as follows:

$$\pi_{m,j} = \frac{n_m^{(j)} + \gamma_j}{\sum_{j=1}^{S} n_m^{(j)} + \gamma_j} \tag{25}$$

Note that the model parameter Θ is a $M \times S \times K$ matrice, where each document season-topic distribution corresponds to a $S \times K$ matrice such that the value in the jth row and the kth column of each matrice is $\theta_{m,j,k}$. Similarly, Φ is a $K \times V$ matrice such that the value in the kth row and the tth column of Φ is $\varphi_{k,t}$ and Π is a $M \times S$ matrice such that the value in the mth row and the jth column of Π is $\pi_{m,j}$.

Algorithm 1 Gibbs sampling procedure of STLDA.

Input: Corpus of attraction documents, α, β, γ, K, S
Output: topic assignment and season label assignment for all words in the corpus

```
Initialize M × S × K matrice Θ, K × V matrice Φ, M × S matrice Π
Zero all count variables: n^(k)_{m,j}, n_{m,j}, n^(t)_k, n_k, n^(j)_m, n_m
for each document d_m, m ∈ {1,2,···,M} do
  for each word w_{m,n}, n ∈ {1,2,...,N_m} in document d_m do
    sample topic index z_{m,n} = k ~ Multinomial(1/K)
    sample season label indice s_{m,n} = j ~ Multinomial(1/S)
    increment counts and sums: n^(k)_{m,j} + 1, n_{m,j} + 1, n^(t)_k + 1, n_k + 1, n^(j)_m + 1, n_m + 1
  end for
end for
while not finished do
  for each document d_m, m ∈ {1,2,···,M} do
    for each word w_{m,n}, n ∈ {1,2,...,N_m} in document d_m do
      decrement counts and sums: n^(k)_{m,j} − 1, n_{m,j} − 1, n^(t)_k − 1, n_k − 1, n^(j)_m − 1, n_m − 1
      sample a new topic index k̃ ~ p(z_i = k|w, z_{−i}, s) according to equation (18)
      sample a new season label j̃ ~ p(s_i = j|w, s_{−i}, z) using equation (19)
      increment counts and sums: n^(k̃)_{m,j̃} + 1, n_{m,j̃} + 1, n^(t)_k̃ + 1, n_k̃ + 1, n^(j̃)_m + 1, n_m + 1
    end for
  end for
  if the Markov chain has converged then
    for every 100 iterations do
      update matrices Θ, Φ and Π with new sampling results
    end for
    output matrices Θ, Φ and Π according to equation (23), (24) and (25)
  end if
end while
```

4. Experimental results

In this section, we evaluate the performances of the proposed STLDA model on real-world travel data, and compare the model with the basic LDA model both qualitatively and quantitatively. It should be pointed out that as far as we know, no literature has conducted the similar research on seasonal topic features of tourist attractions. Therefore, all experimental results of our proposed model are compared with the original LDA model in this study. Specifically, we present the data collection and pre-processing in Section 4.1. The predictive power of the STLDA model measured by the perplexity value and the running time comparisons are presented in Section 4.2. In Section 4.3, we illustrate how STLDA can accurately capture season-dependent topics and improve the topic representation of tourist attractions. In the following experiments, Gibbs sampling algorithm is used both for STLDA and LDA model. We run Markov chains for 1000 iterations to produce samples of latent variables in each of the experiments. Previous studies [32–34] have shown that topic models are not sensitive

to hyperparameters and can produce reasonable results with a simple symmetric Dirichlet prior. During the Gibbs sampling, we use empirical values for the smoothing parameters $\alpha = 50/K$, $\beta = 0.01$ and $\gamma = 0.01$. All experiments are conducted on a PC with an Intel i5 CPU and 4GB of RAM.

4.1. Data collection and pre-processing

We employ Wikipedia (http://www.wikipedia.org) as the primary source of the experimental data from which attraction description information is retrieved. Wikipedia, the collaboratively edited encyclopedia avaiable on the Web with over 30 million articles written in 293 languages and more than 5 million English Wikipedia articles, provides rich information on various aspects including plenty of travel-related knowledge. Our experiment uses the English database of Wikipedia to acquire the attraction description texts. Meanwhile, attraction information is also collected from official websites of the attractions. Since the acquired information regarding a specific attraction is not enough for topic detection, we make full use of abundant travel information from various travel-related websites such as Wikitravel (http://wikitravel.org) and TravelChinaGuide (https://www.travelchinaguide.com). Travelogues from Wikitravel and professional descriptions from TravelChinaGuide are searched by the name of the attractions. It's worth noting that travelogues can serve as a reliable resource of attraction textual information, which is complementary to professional description texts because travelogues cover various travel-related aspects, including not only general scenery description, but also variety of cultural activities that travelers participated in specific attraction, which may be representative characteristics of that attraction. Then a comprehensive attraction description document is generated by integrating all these related information, which contains abundant knowledge for topic detection.

We construct an attraction corpus that consists of attractions description documents written in English. Each document in the corpus is associated with a single famous tourist attraction in China, covering 160 unique attractions in total. The selected attractions including natural landscape and cultural landscape are mainly 5A or 4A tourist attractions evaluated by China National Tourism Administration, where 5A represents the highest level of tourist attraction in China. Table 2 shows the summary of our data collection.

Table 2
Summary of our data collection

Number of attractions	Number of distinct words	Number of total words	Average words in each attraction
93 (5A) 56 (4A) 11 (others) 160 (total)	12011	215995	1350

Since attraction textual information acquired from the Internet is unstructured and usually contains much

disturbance, it is necessary to perform preprocessing on the original attractions textual data before the subsequent experiments. Firstly, punctuations, numbers and other non-alphabet characters are removed. Secondly, All words are lowercased, stop words are removed based on a stop word list from natural language toolkit (NLTK) [35]. Thirdly, for the purpose of reducing the vocabulary size, the low frequency words that appear less than twice in corpus are also filtered out. After preprocessing of the textual information in each attraction, the word distribution of a document can be obtained. Finally, the corpus is further expressed with a data format that can be identifiable by STLDA and LDA model.

4.2. Performance evaluation using perplexity and running time

Perplexity, widely used in the natural language modeling fields, is an important indicator to demonstrate the predictive power of a model [36]. A lower perplexity value means that a higher likehood is achieved on a test dataset, thus indicates a better generalization performance of a model. Given a test dataset D of M documents, the perplexity value can be calculated as follows:

$$perplexity(D) = exp\Big\{-\frac{\sum_{m=1}^{M} logp(\mathbf{w_m})}{\sum_{m=1}^{M} N_m}\Big\} \tag{26}$$

where $p(\mathbf{w_m})$ denotes the generative probability of document m and N_m is the number of words in document m.

In our experiments, we use perplexity to measure the generalization performance of our proposed model. For the attraction corpus, we randomly allocate 75% of the attraction documents for training and the remaining for testing. Fig. 4 shows the results of the perplexity comparison of STLDA and LDA with different number of topics varying from 10 to 100. As shown in Fig. 4, STLDA presents lower perplexity value than LDA with different number of topics, which indicates STLDA owns a better predictive power for unseen documents than the original LDA model. Further analysis shows that the perplexity performance is improved about 28.68% on average. This is due to the ability of STLDA to detect meaningful topics corresponding to various seasonal contexts for attractions by taking the intrinsic seasonal features of attractions into consideration. Therefore, STLDA model can well represent the content of new attractions documents and this leads to its better perplexity performance. From Fig. 4, we can also obtain the optimal number of topics extracted from the attraction corpus for both STLDA and LDA model. The perplexity values of these two models decrease rapidly with the number of topics increasing from 10 to 30, while the performances of these two models become worse when further increasing the latent topic number from 30 to 100. The experimental results reveal that the optimum number of topics for the attraction corpus is 30.

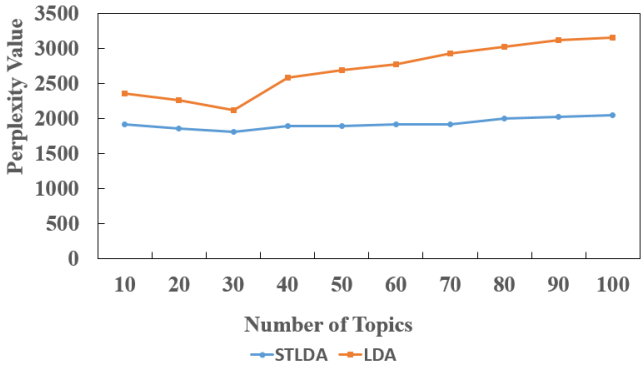


Fig. 4. Perplexity value comparison

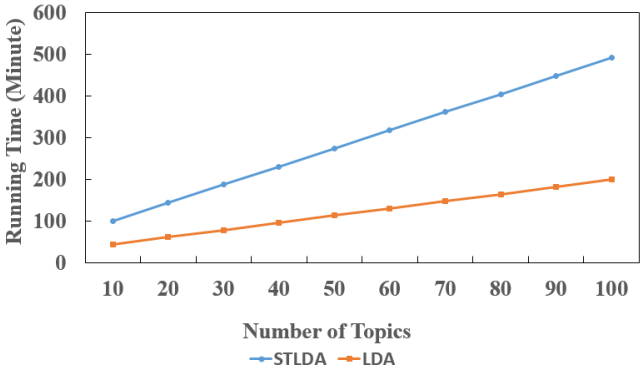


Fig. 5. Running time comparison

The statistical significance of the difference between the STLDA and LDA model regarding the perplexity performance is further assessed by using the Wilcoxon signed ranks test. The Wilcoxon test is a nonparametric test method that is used when overall distribution is unknown [37]. According to the test result, the value of Z statistics is -2.803 and the concomitant probability α is 0.005, less than the significance level of 0.05, which indicates the perplexity performance of STLDA is significantly better than that of LDA at the 95% confidence level. The statistical analysis demonstrates that seasonal contextual information contribute positively to the performance of topic modeling.

To evaluate the time complexity of our proposed model on attraction corpus, we summarize the running time of STLDA and LDA for different number of topics K in Fig. 5. From this figure, it can be easily observed that the running time of STLDA are all longer than LDA with different values of K. This is because STLDA adds an additional season layer on the basis of basic LDA model and this leads to its higher computational complexity. However, from Fig. 5 we can find that just like the LDA model, STLDA still has the linear time complexity and its running time grows linearly as the number of topics increases.

4.3. Topic representation for tourist attractions

To demonstrate the effectiveness of our proposed model, we train STLDA and LDA model on attraction corpus to learn topics for following analysis respectively. The number of topics is set empirically to 30 according to Section 4.2. By analyzing the learned topics of STLDA and LDA, we find that the representative topical words generated from these two models are very close. In order to fairly compare the results of STLDA and LDA model, we present the 22 topics that sharing the same meaning by these two models in Table 3, where some of redundant and meaningless topics in each of these two models are abandoned. The topic number j denotes the jth topic discovered by the model. For illustrating the topics learnt by our proposed model, we also show the representative ten words of each topic in Table 3. Note that here we use representative words in

Table 3
Topics extracted from STLDA for attraction corpus

Topic number of STLDA	Topic label	Representative words
T 1	Parks	willow, sun, parks, tranquility, orioles, falls, roadside, bud, creatures, pond
T 2	Architecture	relic, erected, construction, architectures, built, old, paintings, base, artificial, grounds
T 3	Folk art	sculptures, representative, treasures, burner, carved, create, paintings, color, art, custom
T 4	Entertainment	relaxing, entertainment, comfort, fun, artificial, vacation, resort, amusement, activity, sport
T 5	Village	village, folk, towns, woodcarvings, living, rock, pillar, fisherman, structures, originally
T 6	Hiking	hike, walk, climbing, natural, blossom, hills, species, perch, climate, path
T 7	Woods	ginkgo, luxuriant, woods, reserve, blossoms, resource, verdant, ecological, wetland, bud
T 8	Cultural activity	culture, festival, village, exhibition, listen, minorities, participate, activity, enthusiasm, folk
T 11	Harvest	harvest, field, woods, grow, rapeseed, ripe, mature, farm, golden, resource
T 12	Maple Leaves	colored, autumn, fields, fall, maple, leaves, maples, leaf, red, flower
T 13	Boating	willow, creeks, rowing, tour, boating, dragon, boat, water, streams, rushes
T 14	Mountain	dyke, mount, woods, climbing, buddhists, cableway, leaves, hill, peaks, original
T 15	Brook	brook, sparkling, river, waterfall, willow, streams, clean, flow, luxuriant, garden
T 16	Snowscape	rimed, snow, sliding, silver, freezing, wintersweet, ice, view, snowy, white
T 17	Classical garden	pagoda, hall, pavilion, garden, mountain, gate, courtyard, landscaped, traditional, phoenix
T 18	Seabeach	sunbath, coco, seaside, beach, sand, coast, fun, sunlight, enjoy, surf
T 20	Blossom	blossoming, haze, warm, views, wind, flowers, bloom, verdant, natural, enjoying
T 21	Playground	playground, parks, amazing, artificial, holiday, kingdom, magical, fantasy, cinema, fun
T 22	Temples	towers, cultural, history, burner, incense, buddhas, temple, monasteries, taoism, temples
T 23	Hotspring	hotspring, winter, warm, comfortable, leisure, pool, resort, holiday, healthy, temperature
T 25	Water sports	relaxing, fun, willow, cool, surf, volleyball, football, kingdom, participate, moisture
T 26	Ice sports	rimed, holiday, sliding, ski, iceboating, hockey, snowboarding, sled, ball, sports

each topic rather than words with the highest probabilities to represent the corresponding topic, where the later method is commonly used in topic analysis. The reason for this is that the topic features of attractions often reveal high-level concepts, which is highly correlated to travel-related words. The words irrelevant with travel are removed for better illustrating the extracted topics. We manually assign topic labels to the extracted topics to reflect our interpretation of their meaning.

As can be seen from Table 3, the extracted topics apparently characterize some features of attractions, including both natural styles like woods (topic 7), snowscape (topic 16) and cultural styles like playground (topic 21). The representative words in these different topics are quite informative and coherent. For example for topic 4, words such as relaxing, entertainment, comfort, fun, vacation and resort are related to each other and semantically coherent, conveying the meaning about leisure and entertainment, and thus we name the topic accordingly. The words in topic 7 such as luxuriant, woods, reserve and verdant are closely related to woods, while topic 8 is mainly about cultural activities of ethnic minorities. In addition to the meaning of these topical words, we also refer to classification criteria of Chinese national tourism resources [38] and the related study of paper [5–7, 39] to name all the topics that extracted from attraction corpus and list these topics for subsequent feature analysis of tourist attractions.

In the following, we illustrate how STLDA can accurately capture season-dependent topic clusters and im-

Table 4
Topics representation for selected tourist attractions using STLDA and LDA

Attraction	Model	Season	Detected topics and the corresponding probability value				
Nalati scenic spots	STLDA	Spring	Woods (0.635)	Blossom (0.208)			
		Summer	Cultural activity (0.494)	Brook (0.204)	Woods (0.135)		
		Autumn	Woods (0.574)	Maple leaves (0.266)	Harvest (0.104)		
		Winter	Snowscape (0.613)	Entertainment (0.184)	Ice sports (0.136)		
	LDA		Snowscape (0.232)	Woods (0.184)	Village (0.156)	Entertainment (0.102)	Blossom (0.102)
Yuntai Mountain	STLDA	Spring	Woods (0.648)	Hiking (0.175)	Blossom (0.121)		
		Summer	Brook (0.674)	Temples (0.193)	Mountain (0.129)		
		Autumn	Maple leaves (0.383)	Boating (0.269)	Woods (0.217)	Hiking (0.108)	
		Winter	Snowscape (0.645)	Entertainment (0.298)			
	LDA		Maple leaves (0.240)	Woods (0.220)	Blossom (0.176)	Snowscape (0.126)	Temples (0.107)
Zhangjiajie National Forest Park	STLDA	Spring	Blossom (0.678)	Woods (0.131)			
		Summer	Mountain (0.389)	Brook (0.351)	Cultural activity (0.189)		
		Autumn	Woods (0.604)	Harvest (0.159)	Maple leaves (0.102)		
		Winter	Snowscape (0.324)	Ice sports (0.228)	Woods (0.203)		
	LDA		Woods (0.224)	Blossom (0.179)	Maple leaves (0.131)	Snowscape (0.119)	Harvest (0.111)

prove the topic representation of tourist attractions. The topic probability distribution of each tourist attraction is deeply analyzed for obtaining the representative and comprehensive topic features of tourist attractions corresponding to various seasonal contexts. In order to compare STLDA and LDA model effectively, topics whose occurrence probability greater than 0.1 are selected for each tourist attraction in our experiments. Three tourist attractions, namely Nalati scenic spots, Yuntai Mountain and Zhangjiajie National Forest Park, are selected as typical examples to compare with LDA model. These three attractions are all Chinese national 5A tourist attractions rated by National Tourism Administration, but are located in Northwest of China, middle east of China and Central of China respectively. Table 4 summarizes the obtained topics and topic probability distributions using STLDA and LDA for these tourist attractions. The detected topics for each tourist attraction are arranged in descending order according to the probability values and the probability values of topics are shown in parentheses.

From Table 4, we can clearly see that the topics and topics' occurrence probability of all these three tourist

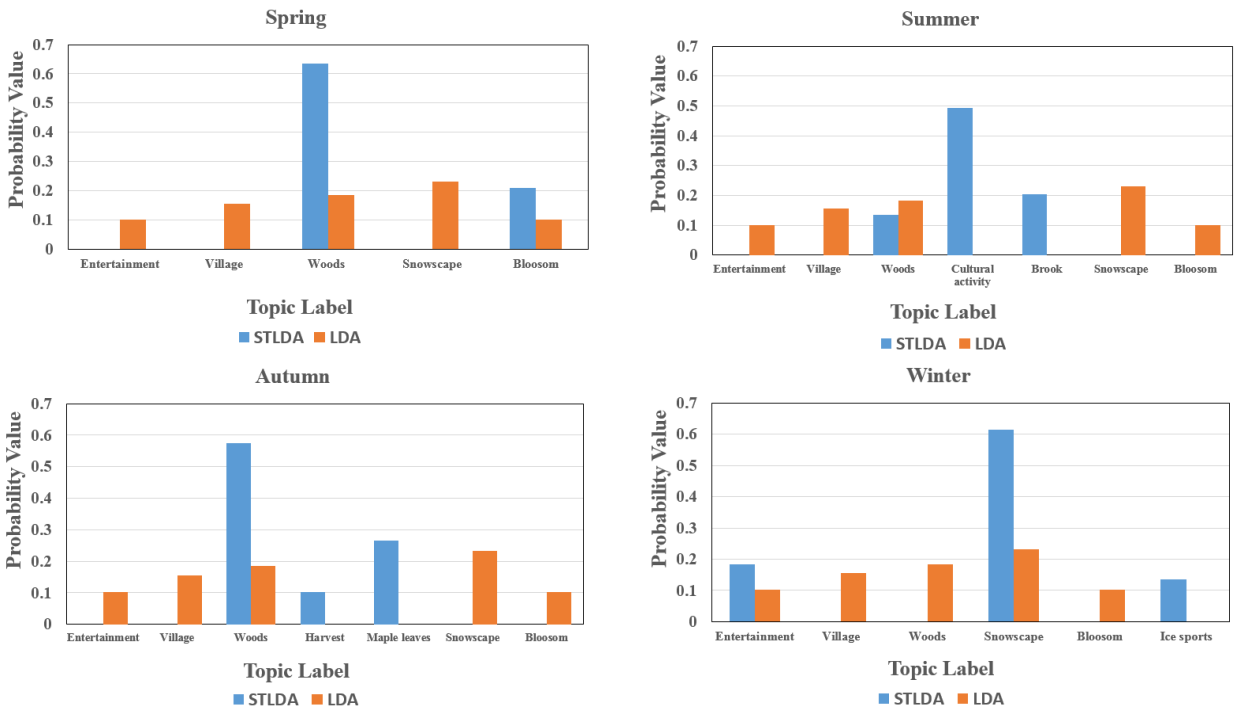


Fig. 6. The topic distribution for Nalati scenic spots with respect to different seasons

attractions change significantly with the alternation of seasons. Specifically, we take Nalati scenic spots for example and the results of STLDA and LDA for this attraction is also visually shown in Fig. 6. Nalati scenic spots, located in Xinjiang Uygur Autonomous Region, is famous for its unique natural scenery, dynamic culture and ethic customs. As can be seen from Table 4, the detected topics from STLDA for Nalati scenic spots are woods and blossom in spring and the corresponding probability value are 0.635 and 0.208 respectively. This shows that the representative topic features of Nalati scenic spots are woods and flowers in spring, which are consistent with the common sense that the beautiful natural scenery is prominent in spring of Nalati scenic spots. The topic generated from STLDA with highest probability in summer for Nalati scenic spots is cultural activity (0.494). Further accessing the relevant information, we see that the temperature is agreeable in summer of Nalati scenic spots, which is suitable for doing outdoor activities. Another interesting finding is that summer is also the peak tourist season to visit the Nalati scenic spots. Thus, the local Kazakhs, who are hospitable and excelling at dancing and singing, often hold a variety of folk activities in summer to show their colorful ethnic culture to the tourists. The topics found by STLDA in autumn are woods, maple leaves and harvest, while in winter are snowscape, entertainment and ice sports respectively. These observations also reveal that topics corresponding to various seasonal contexts generated from STLDA are capable of reflecting the features of the attraction with respect to different seasons in real life. For LDA model, the detected topic with highest probability is snowscape, followed by woods, village, entertainment and blossom. It is obvious that the topics

of Nalati scenic spots generated from LDA are different from those of STLDA. Even the same topics found by these two models such as snowscape, woods, entertainment and blossom, their probability values also differ. The topics detected from STLDA for the attraction are doubtlessly more comprehensive compared with LDA, which is helpful for tourists to fully utilize such knowledge to plan their trip and tour operators to precisely grasp features of tourist attractions so as to provide more targeted publicity and recommendation for tourists. Further analysis of Yuntai Mountain and Zhangjiajie National Forest Park suggest a similar result. The results of STLDA and LDA for Yuntai Mountain and Zhangjiajie National Forest Park are also visually shown in Fig. 7 and Fig. 8 respectively.

With summary of the analysis results of above tourist attractions using STLDA and LDA model, three observed conclusions can be made. Firstly, the topics found by STLDA and LDA for tourist attractions indeed have a certain degree of similarity, but the topic probability distributions are prominently different in these two cases. For instance, the results of STLDA and LDA for Yuntai Mountain both have topics such as woods and blossom. The corresponding probability value obtained from STLDA are 0.648 and 0.121, while in LDA are 0.220 and 0.176 respectively. Secondly, STLDA explicitly identifies topic clusters corresponding to various seasonal contexts, while the topic representation of LDA tend to be more general and less coherent, which can not reveal the seasonality of tourist attractions. Taking Zhangjiajie National Forest Park as an example, STLDA clearly detects and localizes the snowscape topic in winter and the maple leaves topic in autumn, but these topics are confusingly merged by LDA. Not modeling time can confound co-occurrence topic patterns and result in unclear topic representation for tourist attractions. Finally, STLDA clearly detects some other topics that are ignored by LDA model. For Nalati scenic spots, the topics found by STLDA model such as maple leaves, ice sports and cultural activity may be the representative features of this attraction in specific season, but are neglected by LDA model. If the topics whose occurrence probability less than 0.1 are examined from the results of LDA for Nalati scenic spots, we can see that these ignored topics whose probability values are 0.088, 0.043 and 0.0002 respectively. When considering the seasonal contextual information, the probability value of cultural activity topic increases from 0.0002 to 0.494, which makes the cultural activity topic prominent in summer of Nalati scenic spots. This difference comes from STLDA's assumption that takes the intrinsic seasonal features of each tourist attraction into consideration. Therefore STLDA can capture the potential season-dependent topics on a season level of fine-grained, while some of meaningful topics are filtered out in LDA due to their extremely low probability value on a coarse-grained level.

To show the dominancy of STLDA over the basic LDA model more intuitively, we evaluate the statistical

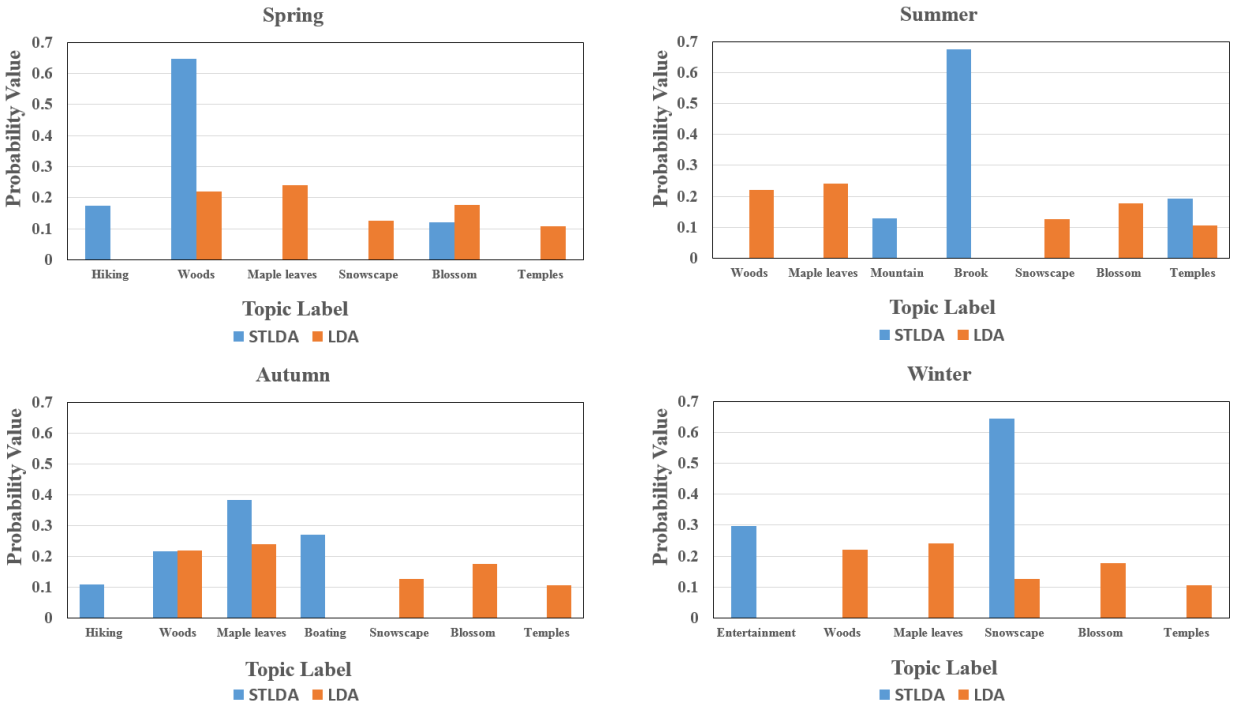


Fig. 7. The topic distribution for Yuntai Mountain with respect to different seasons

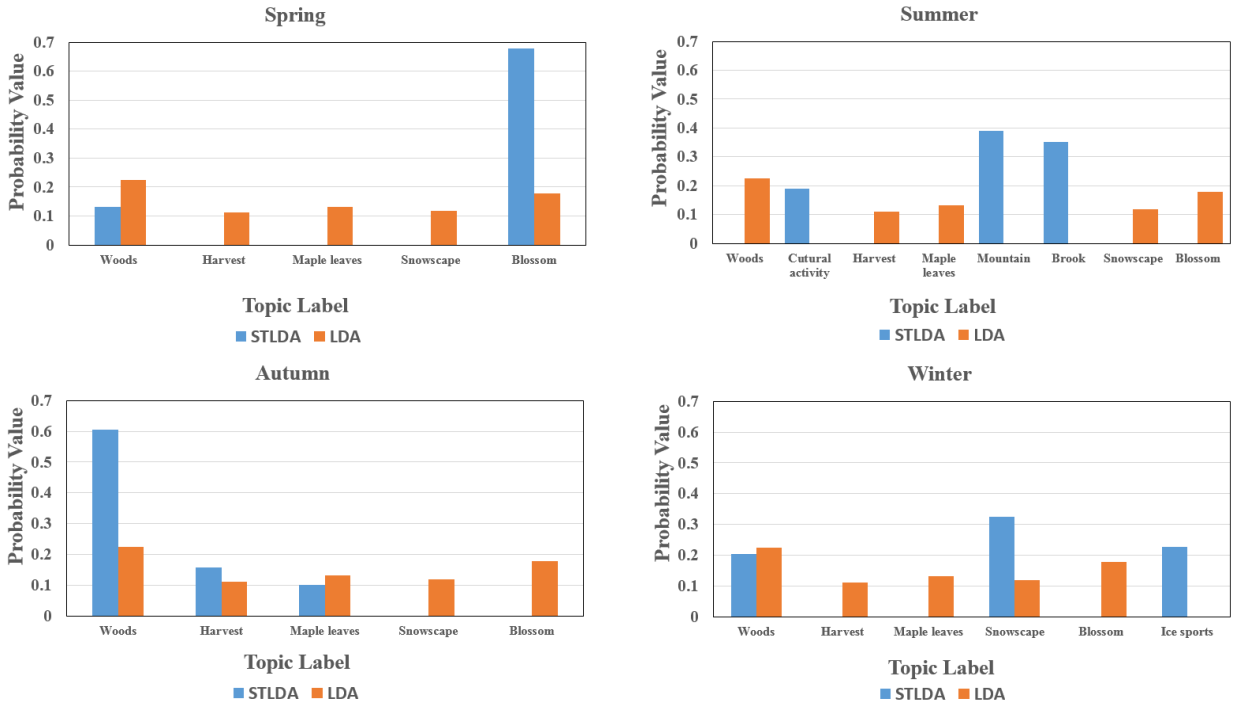


Fig. 8. The topic distribution for Zhangjiajie National Forest Park with respect to different seasons

properties of obtained topics and topic probability distributions of all 160 tourist attractions using these two models and the results are shown in Table 5. The topics whose occurrence probability larger than 0.1 are selected for each tourist attraction. In our experiments, we choose five statistical indicators, namely Richness, Coincidence, Diversity, Significance and Volatility. Richness indicator refers to the average number of topics detected from each tourist attraction. Coincidence indicator denotes the average coincident number of topics generated from these two models for each tourist attraction. Diversity indicator reflects the average number of extra topics generated from one model over the other model for each tourist attraction. Significance indicator represents the average highest topic probability value of each tourist attraction. Volatility indicator indicates the average standard deviation of topic probability distribution corresponding to each tourist attraction. The SP, SU, AU and WI in Table 5 denote the four seasons spring, summer, autumn and winter respectively.

Table 5
Comparison of statistical indicators between STLDA and LDA for attractions corpus

Statistical indicator	LDA	STLDA			
Richness	3.313	7.838			
Coincidence	2.856	2.856			
Diversity	0.456	4.981			
Significance	0.355	0.564(SP)	0.537(SU)	0.536(AU)	0.563(WI)
Volatility	0.085	0.134(SP)	0.128(SU)	0.139(AU)	0.148(WI)

From Table 5, it can be easily noticed that the average number of topics generated from STLDA is 7.838, while the value is 3.313 in LDA, indicating that our proposed model is capable of detecting more topics than LDA. The value of Coincidence indicator is 2.856, comparing with the Richness value 3.313 in LDA, we can see that the topics identified by STLDA can mostly cover the topics generated from LDA. Further referring to the Diversity indicator, the values corresponding to STLDA and LDA are 4.981 and 0.456, which implies that STLDA can capture 4.981 topics more than LDA for each tourist attraction on average, while fail to detect only 0.456 topics for each tourist attraction. Then, we inspect the Significance indicator of these two models. For STLDA, the average highest topic probability values of each tourist attraction are calculated in different seasons. As shown in this table, the indicator values with respect to different seasons in STLDA are 0.564, 0.537, 0.536 and 0.563 respectively, all larger than the value 0.355 in LDA. This result reveals that the topic with the highest probability is dominant in topics generated from STLDA for most attractions corresponding to specific season. In other words, the results of STLDA can give the prominent topic that a tourist attraction belongs to. However, the results of LDA are less significant compared with STLDA. Finally, we investigate the Volatility indicator. From Table 5, we can clearly see that the values of Volatility indicator calculated from the results of STLDA

all larger than that of LDA in different seasons, which indicates that the topic probability values generated from STLDA in each season have remarkable difference. But the topics found by LDA tend to have relatively uniform probability value and consequently there is no obvious discrimination between these topics. For simplicity, we take the Yuntai Mountain for example. The topic with the highest probability 0.645 detected from STLDA in winter is snowscape and the standard deviation of topic probability distribution is 0.174, indicating that the snowscape topic highly belongs to the winter of Yuntai Mountain. In LDA, the highest topic probability value is only 0.240 and the standard deviation of topic probability distribution is merely 0.052. When judging the prominent topic of Yuntai Mountain from the results of LDA in practical applications such as recommending interesting attractions to tourists, the tour operators may get confused due to its relatively uniform probability distribution, and thus can not provide accurate recommendations for potential tourists.

Further deeply investigate the topic probability distributions of all 160 tourist attractions generated from STLDA model, we find that not all tourist attractions have clearly seasonality in every season, but all exhibit seasonal characteristics to some extent. For example, Wuyuan, the most beautiful village of China, have prominent topics in spring, summer and autumn. However, in winter it presents a uniform probability distribution over the topics and the corresponding probability values of topics all less than 0.1, indicating that it does not have remarkable seasonal topic features in winter. Similarly, Thousand Islet Lake shows seasonal characteristics in spring, autumn and winter and it does not present distinct seasonality in summer.

These comparative results indicate that the seasonal contextual information captured by STLDA plays a very important role in forming better topic representation of tourist attractions. By including seasonal contextual information, STLDA can model the variations of topic occurrence that reflect the changing seasonal contexts, which makes STLDA more suitable and effective for topic detection of tourist attractions. An immediate and obvious merit of our proposed model is that this helps us understand more precisely what remarkable topic features are corresponding to specific seasons for a given tourist attraction, which can provide a good opportunity for both tourists and tour operators to fully utilize such information in various decision tasks that are season dependent. The most straightforward application is to apply STLDA to the field of personalized attraction recommendation. For example, a tourist plans to travel during the Labour Day. The Labour Day of China is on May 1st which is spring. Besides, according to the Chinese law, the holidays last for three days, which is one of the peak travel seasons every year. At this time, the tourist intend to see flowers. Taking the data shown in Table 4 as example, the probability values of blossom topic in Nalati scenic spots, Yuntai Mountain and Zhangjiajie National Forest Park generated from LDA are 0.102, 0.176 and 0.179 respectively and there is no

distinct difference between these probability values. When providing recommendation for the tourist based on the results of LDA, the tour operator would lack obvious pertinence. But according to the results of STLDA, it is apparently that Zhangjiajie National Forest Park should be recommended to him since the probability value of blossom topic up to 0.678 in spring. Recommending different tourist attractions to users exposed at specific seasonal contexts on the basis of different seasonal topic features of attractions can achieve higher satisfaction for tourists and realize more profit for tour operators undoubtedly.

5. Conclusions

It is desperately needed to outline tourist attractions from massive travel-related information available on the Web, with the aim of providing decision support for both tourists and tour operators. Thematic analysis for a given tourist attraction provides us a good opportunity to obtain the high-level concepts that reflect the attributes of the attraction. However, the common sense that tourist attractions tend to show distinct features corresponding to different seasons has been neglected by aforementioned probabilistic topic models, which should be considered as a valuable reference for improving topic representation of tourist attractions. Our work addressed this gap and proposed the STLDA model which can model the variations of topic occurrence that reveal the changing seasonal contexts with consideration of the seasonal contextual information exisiting in attraction description documents. Then, we developed an inference algorithm using Gibbs sampling to learn the posterior distributions and model parameters of our proposed model. Finally, to our best of knowledge, there is no similar research result on seasonal topic features of tourist attractions. Therefore, an empirical study that compares the performance of STLDA with the original LDA model was conducted on real-life textual data of selected tourist attractions. Experimental results demonstrate that STLDA outperformed LDA in terms of perplexity and the seasonal contextual information contributed to the improved performance of STLDA. The results also show that STLDA model can effectively capture the season-dependent topics and the topic representations for tourist attractions are much more comprehensive and representative compared with the basic LDA model.

Acknowledgements

This research is supported by National Natural Science Foundation of China (No. 71671038).

References

[1] N. Pervin, F. Fang, A. Datta, K. Dutta, D. Vandermeer, Fast, scalable, and context-sensitive detection of trending topics in microblog post streams, ACM Transactions on Management Information Systems 3 (4) (2013) 1–24. doi:10.1145/2407740.2407743.

[2] T. Takahashi, R. Tomioka, K. Yamanishi, Discovering emerging topics in social streams via link-anomaly detection, IEEE Transactions on Knowledge and Data Engineering 26 (1) (2014) 120–130. doi:10.1109/TKDE.2012.239.

[3] H. Yin, B. Cui, Y. Sun, Z. Hu, L. Chen, Lcars: A spatial item recommender system, ACM Transactions on Information Systems 32 (3) (2014) 11:1–11:37. doi:10.1145/2629461.

[4] B. Skaggs, L. Getoor, Topic modeling for wikipedia link disambiguation, ACM Transactions on Information Systems 32 (3) (2014) 1–24. doi:10.1145/2633044.

[5] Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, L. Zhang, Summarizing tourist destinations by mining user-generated travelogues and photos, Computer Vision and Image Understanding 115 (3) (2011) 352 – 363, special issue on Feature-Oriented Image and Video Computing for Extracting Contexts and Semantics. doi:http://dx.doi.org/10.1016/j.cviu.2010.10.010.

[6] D.-Y. Yeh, C.-H. Cheng, Recommendation system for popular tourist attractions in taiwan using delphi panel and repertory grid techniques, Tourism Management 46 (2015) 164–176. doi:http://dx.doi.org/10.1016/j.tourman.2014.07.002.

[7] Q. Hao, R. Cai, C. Wang, R. Xiao, J.-M. Yang, Y. Pang, L. Zhang, Equip tourists with knowledge mined from travelogues, in: Proceedings of the 19th International Conference on World Wide Web, WWW'10, ACM, New York, NY, USA, 2010, pp. 401–410. doi:10.1145/1772690.1772732.

[8] Q. Hao, R. Cai, X.-J. Wang, J.-M. Yang, Y. Pang, L. Zhang, Generating location overviews with images and tags by mining user-generated travelogues, in: Proceedings of the 17th ACM International Conference on Multimedia, MM'09, ACM, New York, NY, USA, 2009, pp. 801–804. doi:10.1145/1631272.1631418.

[9] J. Shen, C. Deng, X. Gao, Attraction recommendation: Towards personalized tourism via collective in-

telligence, Neurocomputing 173, Part 3 (2016) 789–798. doi:http://dx.doi.org/10.1016/j.neucom.2015.08.030.

[10] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, R. A. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (6) (1990) 391–407.

[11] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'99, ACM, New York, NY, USA, 1999, pp. 50–57. doi:10.1145/312624.312649.

[12] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[13] O. Arbelaitz, I. Gurrutxaga, A. Lojo, J. Muguerza, J. M. Perez, I. Perona, Web usage and content mining to extract knowledge for modelling the users of the bidasoa turismo website and to adapt it, Expert Systems with Applications 40 (18) (2013) 7478 – 7491. doi:http://dx.doi.org/10.1016/j.eswa.2013.07.040.

[14] S. Jiang, X. Qian, J. Shen, T. Mei, Travel Recommendation via Author Topic Model Based Collaborative Filtering, Springer International Publishing, Cham, 2015, pp. 392–402. doi:10.1007/978-3-319-14442-9_45.

[15] Y.-Y. Chen, A.-J. Cheng, W. H. Hsu, Travel recommendation by mining people attributes and travel group types from community-contributed photos, IEEE Transactions on Multimedia 15 (6) (2013) 1283–1295. doi:10.1109/TMM.2013.2265077.

[16] S. Tao, D. Rohde, J. Corcoran, Examining the spatial-temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap, Journal of Transport Geography 41 (2014) 21–36. doi:http://dx.doi.org/10.1016/j.jtrangeo.2014.08.006.

[17] H. Stromberg, O. Rexfelt, I. M. Karlsson, J. Sochor, Trying on change-trialability as a change moderator for sustainable travel behaviour, Travel Behaviour and Society 4 (2016) 60–68. doi:http://dx.doi.org/10.1016/j.tbs.2016.01.002.

[18] R. Butler, Seasonality in tourism: Issues and implications, The Tourist Review 53 (3) (1998) 18–24. doi:http://dx.doi.org/10.1108/eb058278.

[19] X. Wang, A. McCallum, Topics over time: a non-markov continuous-time model of topical trends, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'06, ACM, New York, NY, USA, 2006, pp. 424–433. doi:10.1145/1150402.1150450.

[20] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine Learning, ICML'06, ACM, New York, NY, USA, 2006, pp. 113–120. doi:10.1145/1143844.1143859.

[21] H.-M. Lu, Detecting short-term cyclical topic dynamics in the user-generated content and news, Decision Support Systems 70 (2015) 1–14. doi:http://dx.doi.org/10.1016/j.dss.2014.11.006.

[22] Q. Liu, E. Chen, H. Xiong, Y. Ge, Z. Li, X. Wu, A cocktail approach for travel package recommendation, IEEE Transactions on Knowledge and Data Engineering 26 (2) (2014) 278–293. doi:10.1109/TKDE.2012.233.

[23] Y. Kim, K. Shim, Twilite: A recommendation system for twitter using a probabilistic model based on latent dirichlet allocation, Information Systems 42 (2014) 59 – 77. doi:http://dx.doi.org/10.1016/j.is.2013.11.003.

[24] G. Heinrich, Parameter estimation for text analysis, Technical report, Fraunhofer Institut fur Graphische Datenverarbeitung (2009).

[25] D. Ramage, D. Hall, R. Nallapati, C. D. Manning, Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, EMNLP'09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 248–256.

[26] D. M. Blei, J. D. Lafferty, A correlated topic model of science, The Annals of Applied Statistics 1 (1) (2007) 17–35. doi:10.1214/07-AOAS136.

[27] M. Paul, R. Girju, A two-dimensional topic-aspect model for discovering multi-faceted topics, Urbana 51 (61801) (2010) 36.

[28] Y. Liu, A. Niculescu-Mizil, W. Gryc, Topic-link lda: Joint models of topic and author community, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, ACM, New York, NY, USA, 2009, pp. 665–672. doi:10.1145/1553374.1553460.

[29] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI '04, AUAI Press, Arlington, Virginia, United States, 2004, pp. 487–494.

[30] C. Andrieu, N. de Freitas, A. Doucet, M. I. Jordan, An introduction to mcmc for machine learning, Machine Learning 50 (1) (2003) 5–43. doi:10.1023/A:1020281327116.

[31] T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences 101 (suppl 1) (2004) 5228–5235. doi:10.1073/pnas.0307752101.

[32] A. Daud, J. Li, L. Zhou, F. Muhammad, Knowledge discovery through directed probabilistic topic models: a survey, Frontiers of Computer Science in China 4 (2) (2010) 280–301. doi:10.1007/s11704-009-0062-y.

[33] C. Lin, Y. He, R. Everson, S. Ruger, Weakly supervised joint sentiment-topic detection from text, IEEE Transactions on Knowledge and Data Engineering 24 (6) (2012) 1134–1145. doi:10.1109/TKDE.2011.48.

[34] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (4) (2012) 77–84. doi:10.1145/2133806.2133826.

[35] S. Bird, Nltk: The natural language toolkit, in: Proceedings of the COLING/ACL on Interactive Presentation Sessions, COLING-ACL'06, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006, pp. 69–72. doi:10.3115/1225403.1225421.

[36] A. McCallum, X. Wang, A. Corrada-Emmanuel, Topic and role discovery in social networks with experiments on enron and academic email, Journal of Artificial Intelligence Research 30 (2007) 249–272. doi:10.1613/jair.2229.

[37] B. Rosner, R. J. Glynn, M.-L. T. Lee, The wilcoxon signed rank test for paired comparisons of clustered data, Biometrics 62 (1) (2006) 185–192. doi:10.1111/j.1541-0420.2005.00389.x.

[38] CNTA, Classification, investigation and evaluation of tourism resources, Tech. rep., China National Tourism Administration (2003).

[39] A. Leask, Progress in visitor attraction research: Towards more effective management, Tourism Management 31 (2) (2010) 155 – 166. doi:http://dx.doi.org/10.1016/j.tourman.2009.09.004.