International Journal of
Geo-Information

Article

Geo-Tagged Photo Metadata Processing Method for Beijing Inbound Tourism Flow

Wen Chen [1,2], Zhiyun Xu [1], Xiaoyao Zheng [3] and Yonglong Luo [1,3,*]

[1] School of Geography and Tourism, Anhui Normal University, Wuhu 241003, China; wchen@ahnu.edu.cn(W.C.); zhyxu@mail.ahnu.edu.cn(Z.X.);

[2] School of Mathematics and Computer Science, Tongling College, Tongling 244000, China

[3] Anhui Provincial Key Laboratory of Network and Information Security, Anhui Normal University, Wuhu 241003, China; zxiaoyao@ahnu.edu.cn;

[*] Correspondence: ylluo@ustc.edu.cn; Tel.: +86-0553-5910645

Received: 24 October 2019; Accepted: 02 December 2019; Published: 3 December 2019

Abstract: Technological advances have led to numerous developments in data sources. Geo-tagged photo metadata has provided a new source of mass research data for tourism studies. A series of data processing methods centering on the various types of information contained in geo-tagged photo metadata have thus been proposed; as a result, the development of tourism studies based on such data has advanced. However, an in-depth study of the data processing methods designed to conduct tourist flow prediction based on geo-tagged photo metadata has not yet been conducted. In order to acquire accurate substitutive data regarding inbound flows in cities, this paper introduces and designs several methods, including data screening, text data similarity calculation, geographical location clustering, and time series data modelling, in order to realize a data preprocessing model for inbound tourist flows in cities based on geo-tagged photo metadata. Wherein, the entropy filtering method was introduced to aid in determining whether the data were posted by inbound tourists; whether the inbound persons' activities were related to tourism was judged through the calculation of tag text similarity; an efficient clustering method based on geographic grid partition was designed for cases in which the tag values were empty; finally, the time series of the inbound tourist flows of a certain region and period were obtained through data statistics and normalization. For the empirical research, Beijing City in China was selected as the research case, after which the feasibility and accuracy of the methods proposed in this paper were verified through data correlation analysis between Flickr data and real statistical yearbook data, as well as analysis of the prediction results based on a machine learning algorithm. The data preprocessing method introduced and designed in this paper provides a reference for the study of geo-tagged photo metadata in the field of tourism flow prediction. These methods can effectively filter out inbound tourist flow data from geotag photo metadata, thus providing a novel, reliable, and low-cost research data source for urban inbound tourism flow forecasting.

Keywords: geo-tagged photo metadata; inbound tourism flow forecast; data preprocessing; flickr; Beijing; big data analytics; machine learning algorithm; data correlation analysis

1. Introduction

Due to the significant growth in demand for world tourism, people have become increasingly interested in tourism studies [1]. Naturally, tourism studies cannot be conducted without research data. Surveys and opinion polls are common methods used to collect tourism data, while statistical yearbooks are another frequently used data source [2,3]. However, all of these collection methods are time-consuming [4]. The results of surveys and related statistics are typically published within about eight weeks of data collection [5], and the release cycle for a statistical yearbook is even longer [6]. In addition, the high cost of tourism data collection is a problem that needs to be urgently addressed [7,8]. It is thus worth studying ways of accurately and promptly acquiring tourism data, as well as methods of analyzing and exploring the potential values by breaking through the traditional tourism data collection approaches [9].

The development of network and information technology has provided us with brand new avenues for data collection. Researchers have identified a close relationship between online data and socioeconomic activity. For example, Ginsberg et al. [10] use online search data to successfully predict the trends in flu outbreaks. Yang et al. [11] employ web search queries on Google and Baidu to predict tourism demand for Hainan, China; by doing so, they indicate that web searches not only reflect tendencies in tourism product preferences, but can also improve prediction accuracy. Li et al. [12] propose a new forecasting frame combined with search-trend data, which is applied to the forecasting of Beijing (formerly Romanized as Peking) tourism. Moreover, network sharing platforms such as Flickr, Panoramio, Pinterest, etc., offer mass geo-tagged photo shared by users. Hence, the geo-tagged photo metadata contained in EXIF (exchangeable image file format), including shooting time and geographic coordinates, can also provide a large body of research data for tourism studies [13,14]. This type of data is a kind of UGC data (User-Generated Content). Li et al. [13] found that about 47% of data used in tourism research is UGC data. For example, Kisilevich et al. [15] used data from photo-sharing sites such as Flickr and Panoramio to achieve point of view analysis, point of interest extraction, and comparison of tourist behavior patterns. Moreover, based on photo data from online photo sharing services and Wikipedia's knowledge interaction, Lucchese et al. [16] achieved personalized recommendations for tourist attractions, while Mou et al. [17] used geo-tagged photos to analyze the spatial and temporal distribution of tourists and the changes in inbound tourism flows. In the tourism industry, the widespread availability of big data opportunities has changed the traditional research approach substantially. The enormous amount of available data has made the information extraction phase more complex, such that it now requires advanced analytics techniques to perform [18,19].

To better exploit the value of geo-tagged photo metadata, a number of researchers have conducted helpful explorations of geo-tagged photo metadata processing. The technology used to process geo-tagged photo metadata includes text data processing, geographical location clustering, image identification, and time series data modelling. These four types of technology are used to process and analyze different types of data contained in geo-tagged photo metadata: namely, text tags, geotags, image contents, and time of image capture [14]. For example, Peng et al. [20] outline a new method for discovering popular tourist attractions by integrating spatial clustering and text mining methods in order to extract hotspots. Kisilevich et al. [21] present P-DBSCAN (Photo Density-Based Spatial Clustering of Applications with Noise), a new density-based clustering algorithm based on DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [22], which is used to analyze places and events using a collection of geo-tagged photos. Zhang et al. [23] analyzed the contents of tourism images through in-depth learning technology, applying this technology to identify specific tourism behaviors and explore tourist cognitions about certain tourist destinations. Furthermore, Xia et al. [24] detected and calculated events and trends in geo-tagged photo data by applying time series analysis and classification technology. Many of the above types of analytical methods are often used in combination. A single method of processing such data can affect the accuracy of the results. For example, Pantano et al. [25] used a single cluster analysis method to analyze pictures from Flickr and demonstrated the role of building appearance in tourism attraction. Giglio et al. [26] used a single cluster analysis method to analyze Flickr pictures in order to demonstrate the relationship between human mobility and tourist attractions. The above study directly used Flickr data from a certain region as an experimental object. However, these Flickr images were not necessarily taken by tourists; data unrelated to tourism has no positive effect on tourism research.

Studies using the above data processing methods have laid the foundation for the application of geo-tagged photo metadata, such that it is possible to extensively apply data of this kind in multiple research fields. (1) In the field of travel recommendation, Kou et al. [27] find that the order in which travelers visit specific locations can be determined from the timestamp and location shown in Flickr shared photos; on this basis, these author shave built a mixed model to estimate the probability that certain interesting places would be visited. Moreover, by utilizing numerous geographical labels and images with text annotation, Chen et al. [28] established a distributed geographical image retrieval and recommendation system. In addition, Xu et al. [29] proposed a recommendation method for individualized travel locations based on the logs and contents of geo-tagged photos. (2) In the field of tourist behavior analysis, Su et al. [30] collected tourist geographical data from Flickr for 333 prefecture-level cities in China and analyzed the geographical preferences of international tourists. Koylu et al. [31] established a computer vision algorithm based on a convolutional neural network combined with kernel density estimation to identify objects of interest, while human activity patterns were also inferred from geo-tagged photos on Flickr. Yang et al. [32] used geotagged photographs from Flickr to extract visitors' trajectories, identifying and analyzing the typical travel behavior patterns associated with different sightseeing preferences to improve travel recommendations. (3) Additionally, geo-tagged photo metadata has also been applied to studies in the tourism prediction field. However, most existing prediction studies based on geo-tagged photo metadata aim at predicting tourists' geographical locations [33,34], behaviors [35,36], etc. Using the keywords "geo-tagged photo metadata" and "forecast", we searched the literature database using Web of Science. The results show that few tourist flow prediction studies based on geo-tagged photo metadata have been conducted. One example of such a study is that of Miah et al.[14], who evaluated the applicability of big data analysis methods in tourist destination management, and further analyzed the feasibility of tourist flow prediction based on metadata from Flickr geo-tagged photos. This article provides an overview of the construction of a time series model of Melbourne's future tourism needs. However, it does not describe the specific processes and methods used to obtain tourism flows in the region, especially the lack of introduction of urban inbound data acquisition methods based on geo-tagged photo metadata.

Tourism has become an important sector in many countries, and reliable tourism demand forecasts are important for government and business for planning and investment purposes [37,38]. Based on the above analysis, it can be seen that there are some inherent defects of the data collection methods traditionally used for tourism studies [4–8]. Existing studies [17] point out that geo-tagged photo metadata obtained from Flickr can be adopted as a substitute data source rather than using traditional tourism data (such as survey data, statistical yearbooks, etc.). However, further research into the methods of processing geo-tagged photo metadata to facilitate tourist flow prediction is still required.

In light of the above research background, this paper takes the data provided by typical network sharing platforms like Flickr as the research object and studies the data processing methods required to provide accurate substitutive data pertaining to inbound tourist flows in cities. Moreover, based on analysis of the information contained in such data, the entropy filtering method is introduced to identify whether a user whose location value is empty is an inbound visitor. Judgment is made on whether foreigners' activities in a certain region are tourism-related through a calculation of the text similarity of tag texts. To address cases in which the tag value is empty, a clustering method based on geographic grid partition is designed to facilitate a judgement of whether the camera site was near a POI (Point of Interest). Finally, the two types of data, i.e., geo-tagged photo metadata and statistical yearbook data, are adopted as experimental data. Geo-tagged photo metadata is derived from the Flickr data processed in this article. The statistical yearbook data is obtained from the officially released data on urban inbound tourism flows. A prediction model for inbound tourist flow in cities is thus established based on three typical machine learning prediction models, after which the feasibility and accuracy of the data processing methods proposed in this paper are evaluated. The datasets and related data processing methods used in this article are listed in Table 1.

Table 1. Source and references.

Type	Name	Links / References
Data source	Flickr API	http://www.flickr.com/services/api/
	Beijing Statistical Yearbook data	http://www.bjstats.gov.cn/
Data screening	entropy filtering	Reference [39]
Text data processing	word2vec	Reference [40]
	TF–IDF	Reference [41]
Geographical location clustering	DBSCAN	Reference [22]
	P-DBSCAN	Reference [21]

2. Description of the Methodology

Flickr data are a typical example of geo-tagged photo metadata. Therefore, Flickr data is used in this paper to explore the characteristics of existing problems with such data, along with their solutions. An example of Flickr geo-tagged photo metadata is shown in Table 2. The latitude and longitude information can be used to determine whether the shooting location is within the currently predicted city range; the location information is submitted by the registered user, and this information can indicate whether the user is an overseas tourist. Moreover, the time information can be used to determine the time of travel; this information allows us to obtain the number of travel streams over a certain period of time. Finally, the tag information contains information about the photos themselves, such as a description of an attraction, a description of an event (parade, wedding, etc.), even invalid information (null value, description information not related to travel, etc.), etc. We can use this information to determine whether the photo in question is related to travel.

Table 2. Example geo-tagged Flickr photo metadata.

UserID	Latitude	Longitude	Location	Date & Time	Tags
12XXXXX23@N03	116.403284	39.924407	Beijing, China	2015/1/1 16:17:46	Forbidden City, Gold, China, Lions, Sculpture, Beijing, Peking, Bronze, Palace
92XXXXX4@N00	116.402925	39.932042	South Hamilton, Massachusetts, USA	2015/1/2 14:52:54	China, Beijing, Prospect Hill, Jingshan

Data from a certain region and a certain period can be screened out according to the longitude and latitude information and shooting time in Table 2. This is an easy process that will not be further described here. The example in Table 2 shows data in an ideal state; real data often contains uncertainties and noise. The problems existing in real data, along with some solutions, will be outlined in the next few subsections.

2.1. Screening of Domestic and Foreign Tourists

This paper focuses on analyzing the forecasting of inbound flows. Accordingly, we removed the tourist information pertaining to local tourists using the following methods.

1. If the location value in Table 2 indicates a domestic user, the photo information will be deleted. In addition, the location value in Table 2 can be filled out arbitrarily when a user registers on Flickr. Some users may have mistakenly filled out their location incorrectly, or deliberately entered an incorrect location. Furthermore, even though Flickr users can be confirmed to come from foreign countries by examining the location value, there are some foreigners who are living temporarily in a sightseeing destination; the photos uploaded by these foreigners therefore cannot be defined as tourism-related data. The Technical Handbook on the Collection and Presentation of Domestic and International Tourism Statistics published by the World Tourism Organization [42] defines international tourists as follows: people who remain in a tourist destination for more than one year

should not be defined as tourists. Therefore, we will analyze the timespan in which the pictures are taken. If photos are taken by a user in the same location for more than one year, these users can be defined as local residents. These photos will thus be deleted.

2. Moreover, some location values are empty. To solve this problem, we carry out data screening through the entropy-filtering method suggested in [39]; this is explained in more detail below.

$$\mathrm{E}(u) = -\sum_{i}^{\mathrm{Mon}(u)} P_i(u)\log P_i(u), \tag{1}$$

$$P_i(u) = \frac{D_i(u)}{\sum_{i}^{Mon(u)} D_i(u)}, \tag{2}$$

$D_i(u)$ denotes the days on which user u has taken photos in month i, while $\mathrm{Mon}(u)$ is the timespan during which user u has taken photos, expressed as a number of months. $P_i(u)$ is the probability that user u takes a photo in the i-th month; this figure is equal to the number of photos taken in the i-th month divided by the total number of shots in $\mathrm{Mon}(u)$ months. For example, if the number of photos posted by a user each month over three months are 2, 1, and 2, then, $\mathrm{E(u)} = -\left(\frac{2}{5}\right)\log\left(\frac{2}{5}\right) - \left(\frac{1}{5}\right)\log\left(\frac{1}{5}\right) - \left(\frac{2}{5}\right)\log\left(\frac{2}{5}\right) = 1.522$. We thus set a threshold value to measure whether or not the user is a foreigner. This threshold-setting process will be outlined in more detail in the experiment in Section 3.

2.2. Screening of Data Irrelevant to Tourism

The photos uploaded to Flickr differ in terms of their themes. The data acquired in Section 2.1 are simply the activity data of foreigners in a certain region. However, what is of interest in this paper is the issue of tourist flow prediction; thus, there is a need to judge whether the foreigners' activities are related to tourism. The tag text in Table 2 is the descriptive information of the text type provided for photos by Flickr users. Based on the similarity calculation method [43], it can be determined whether the theme of the photo is related to tourism. There are two possible situations regarding the tag value: either the tag contains effective text description, or the tag value is empty. Two solutions for these two tag value cases are presented below.

1. For the situation in which the tag has a value, the tourism correlation of the tag text is analyzed through a text analysis method involving the following steps:

(1) Establish the tourism text corpus. The language materials appearing in the actual use of language are stored in this corpus. The corpus can be compiled from search engine data, social network data, or Wikipedia [43]. The tourism text corpus established here will be used in the calculation of text similarity in Step (4).

(2) Establish the dictionary of 'stop words' and remove these from the tag text. The tag text contains some words that are irrelevant to our study, such as camera model, f-number, and so on. These terms are referred to as stop words, and their set is called the dictionary of stop words. To improve the efficiency and accuracy of the follow-up operation, it is necessary to establish the dictionary of stop words and to remove these stop words from the tag text. We counted the word frequency for tag text. Following manual screening, words with a higher word frequency that are unrelated to tourism are included in the stop word dictionary.

(3) Acquire subject terms of the tag text via TF–IDF (Term Frequency–Inverse Document Frequency). In [41], a method for acquiring subject terms via TF–IDF was proposed. As a statistical approach, TF–IDF is used to evaluate the degree of importance of the words in the text. The most important word in the text is the subject term.

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d)\times\mathrm{idf}(t,D), \tag{3}$$

$$\mathrm{idf}(t,D) = \log\frac{\mathrm{N}}{|\{d\in D, t\in d\}|}, \tag{4}$$

In Equations (3) and (4), tf denotes the count of label t in text d, while idf represents the information provided by label t in database D and N is the record count of database D. Typically, the label with the highest frequency in each text is selected as the text label. The text label is the subject term of the tag text referred to in this article.

(4) Calculate the tourism correlation. Computer science has provided an effective text similarity calculation method for natural language processing applications. Based on the tourism text corpus established in Step (1), the text similarity between the subject terms of the tag text and the word "tourism" can be determined by means of the text similarity calculation method. Data exhibiting a similarity greater than the predetermined threshold value will be treated as tourism-related data.

2. For the situation in which the tag value is empty, the following assumption is proposed to enable the largest amount of tourism-related data to be acquired: namely, if the tag value of a photo is empty and the filming location is near a POI, the photo will be deemed tourism-related. These data were analyzed by means of geographical location clustering. Photos whose clustering center is near a POI are considered to be tourism related. However, traditional clustering algorithms have been found to be inefficient in dealing with massive amounts of big data; accordingly, we propose a clustering method based on geographic grid partitioning. The steps of the algorithm are as follows:

(1) Divide the grid. In light of the application scenario of this paper, we divide the dataset of a certain area into multiple sub-areas according to geographical location; the dataset of each sub-area serves as a subset of clusters.

(2) Subset clustering. We use P-DBSCAN to cluster each subset. Since no intersection between the subsets exists, the clustering process of each subset can be conducted in a multi-threaded parallel mode, which can effectively improve the clustering efficiency. Finally, all of the cluster center points of each subset are summarized into candidate sets.

(3) Hierarchical clustering. Several problems can arise when engaging in subset clustering based on meshing. First, it is easy for subset clustering to fall into a local optimal solution; second, clusters close to grid boundaries are likely to belong to the same cluster as a cluster outside the boundary. Therefore, for the cluster center points of all subsets generated in step (2), we perform clustering again to obtain the final clustering results.

(4) Reclassify. We compare each object in the original Flickr dataset with the cluster center in the hierarchical clustering results. When the spacing is below a certain distance threshold, we can consider the object as belonging to the cluster. A point that does not belong to any cluster is considered an outlier.

2.3. Data Statistics and Data Normalization

After the above preprocessing operation is complete, data about inbound tourism of a certain region and period can be screened out. After the monthly statistics relating to the preprocessed Flickr geo-tagged photo metadata are generated, time series prediction data is formed. Next, as the order of magnitude of the value is far lower than the actual number of inbound tourists, we conduct data normalization (by means of Equation (5)) in order to improve prediction accuracy and avoid potential value problems such as the dominant occupation by larger values [44].

$$x_i{}' = \frac{x_i - x_{min}}{x_{max} - x_{min}}, \tag{5}$$

In Equation (5), $x_i{}'$ represents the result of normalization, while x_i represents the data that needs to be normalized; moreover, x_{max} and x_{min} represent the maximum and minimum value in the dataset respectively. In this article, normalization is used to map real-life tourist flow data and Flickr data to the range of 0–1. The normalization calculation process is conducted as follows: for example, assume that the maximum and minimum values of inbound tourism by month in a certain region in a certain year are 200 and 2000 respectively, and that the current number of people in a certain month is 1100; thus, according to Equation (5), the result of normalization of the data is: $x_i{}' = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{1100-200}{2000-200} = 0.5$.

2.4. Prediction Model

The machine learning prediction algorithm is represented by a neural network (NN), while support vector regression (SVR) is applied extensively in the tourism prediction field [1]. In addition, Huang et al. [45] proposed the Extreme Learning Machine (ELM) algorithm, which was proven to achieve better performance than traditional machine learning algorithms. A comparison was drawn between the ELM algorithm and classical machine learning algorithms in the academic literature, after which the superiority of the ELM algorithm was again verified [46,47]. Therefore, ELM has been widely applied to prediction studies across multiple fields with excellent results [48,49]. Several prediction algorithms, including NN, SVR, and ELM, are introduced in this paper, while the time series of inbound tourist flows and officially issued statistical yearbook tourism data following normalization (see Section 2.3) are adopted as learning data for the prediction model. The feasibility and accuracy of the data processing methods proposed in this paper are verified through contrast experiments.

3. Case Study

3.1. Selection of Case Location

Among the countries that play host to global tourists, China remains at the global forefront when it comes to the scale of inbound tourism [50]. As the capital of China and one of the country's most important and famous historical cities, Beijing is home to numerous scenic spots and historical sites. The work of collating inbound tourism statistics was started early in Beijing, so it has relatively integrated historical data [51] that provide valid comparative data for the present research. Consequently, Beijing was selected as the study case in this paper.

3.2. Data Acquisition and Preprocessing

Tian'anmen is the center of Beijing. We take the latitude and longitude of Tian'anmen as the central point of the sampling, with the maximum measurement range set to 32 km. The crawler code was edited with Python based on the Flickr API. In total, 349,665 pieces of photo data from 2007 until 2016 were collected; among these, 34,160 were from 2015 and 47,075 were from 2014.

3.2.1. Screening of Domestic and Foreign Tourists

In line with the analysis in Section 2.1, we removed the data uploaded by local residents and those who are resident in the tourist destination. Table 3 presents the resulting dataset. The proportion of deleted photos accounts for about 47% of the total. Among them, photos taken by foreigners living in tourist destinations accounted for 3.35%.

Table 3. Dataset summary.

Year	Raw	Filtered	Users	Tags
2015	34,160	16,154	3,655	17,232
2014	47,075	21,636	3,754	19,086

The threshold E(u) next needs to be determined so that we can distinguish whether a user whose location value is empty is an inbound visitor. The literature [39] considers that tourists remain in the tourist area for a relatively short period of time (for example, one week). Therefore, the date on which the image was shot should normally be within the same month, or up to several consecutive months. Local residents can take images during almost all months. This article applies the entropy filtering method to distinguish between the shooting data of residents and tourists. E(u) is calculated on the basis of different values of k using Equation (1) and (2).

Thirty percent of the 2015 data was randomly selected as the test data for the E(u) threshold. It can be seen from the experimental results in Table 4 that 2.0 is the optimal threshold value, and the filter rate has reached 79.2%. That is to say, when we set the threshold to 2.0, 79.2% of the data released by overseas tourists can be effectively screened out. Therefore, we chose the threshold with the highest filtering rate as the optimal threshold.

Table 4. Filtering accuracy of user types with different thresholds.

Month count(k)	**3**	**4**	**5**	**6**	**7**
E(u)	$-3\times\left(\frac{1}{3}\right)\log\left(\frac{1}{3}\right)$	$-4\times\left(\frac{1}{4}\right)\log\left(\frac{1}{4}\right)$	$-5\times\left(\frac{1}{5}\right)\log\left(\frac{1}{5}\right)$	$-6\times\left(\frac{1}{6}\right)\log\left(\frac{1}{6}\right)$	$-7\times\left(\frac{1}{7}\right)\log\left(\frac{1}{7}\right)$
Threshold value	1.6	2.0	2.3	2.6	2.8
Filter rate	74.4%	79.2%	72.4%	71.0%	70.3%

3.2.2. Screening of Data Irrelevant to Tourism

1. Tourism correlation analysis of tag text

(1) Establish the tourism text corpus. TripAdvisor (https://www.tripadvisor.in/) is a world-famous tour recommendation website containing hundreds of millions of comments and suggestions from tourists. We used Python's BeautifulSoup library to capture travel commentary text data pertaining to popular sites in Beijing from the TripAdvisor website. We performed a simple cleaning operation (e.g., deleting emoji and so on) on the text data set. This text can be used directly as a word2vec corpus without the need for additional operations [51]. The text of comments provided by users from different language countries forms an important part of the travel text corpus. Therefore, we use Google Translate to translate texts from different languages into English for subsequent experiments.

(2) Establish the stop word dictionary and remove stop words from the tag text. The dictionary of stop words established in this paper contains specific words frequently used on Flickr, such as Instagram-related, app-related, and camera parameter-related terms. These stop words are irrelevant to the content of the research. For example, according to the word frequency statistics of the tag text, the top 10 stop words with the highest word frequency are as follows: Beijing, China, CN, uploaded, Instagram app, picture, iphoneography, format, lens, film. Because of their high word frequency, they are very likely to be the subject of tag text. Therefore, stop words in the tag text were identified through matching with the contents of the stop word dictionary and subsequently removed. The first column of Table 5 displays the tag text descriptions of a group of photos uploaded by a user following removal of the stop words. The text contains descriptions of several photos of a single activity; thus, repeated descriptive terms exist, and these repeated terms highlight the theme of the text. In addition, users from different countries use different languages on Flickr, meaning that tag texts are written in various languages. Non-English descriptions in the tag text were all translated into English via Google Translate.

(3) Acquire the subject terms of the tag text via TF–IDF. The data ratio of effective text descriptions contained in the tag text is as follows: 70.74% in 2014; 73.57% in 2015. Tag text subject terms in the Flickr data from 2014 and 2015 were acquired via TF–IDF. The second column of Table 5 presents the subject terms acquired and the calculation results. Words with the highest TF–IDF value were selected as the subject terms of the texts; some texts may have several subject terms.

(4) Calculate the text similarity. The Python gensim package was used to realize the word2vec model of the tourism text corpus [51]. word2vec is an open-source word vector calculation tool developed by Google. It maps each word into a vector and gathers similar words in the vector space according to the vector similarity. Therefore, it is able to effectively measure the words similarity. The word2vec model of the tourism text corpus contains a large number of words related to tourism. Based on this model, the results of similarity calculation between the subject terms of the tag text and the word "tourism" can be gained. According to [41], the higher the value of the cosine distance between two vectors composed of words , the more similar these two words will be. The third column of Table 5 presents the results of the text similarity calculation. We randomly selected some experimental results under different thresholds for manual screening. The results show that when the threshold is set to 0.83629805, the correlation between the obtained data and "tourists" reaches a maximum of 90.17%. Accordingly, we chose this threshold and can thus obtain as much relevant prediction data as possible.

Table 5. Tag text and its processing results.

Tags	TF–IDF results	Cosine Distance
stadium landmark sports light night architecture Olympics landmark architecture	Architecture: 0.4902234 landmark: 0.4902234	architecture: 0.1135169 landmark: 0.1577416
forbidden palace architecture landmark sight river reflection square landmark landmark attraction temple	landmark: 0.5363209	landmark: 0.1577416
zoo animal mammal panda zoo animal mammal nature deer zoo street river winter zoo animal mammal nature monkey	zoo: 0.6059326	zoo: 0.1999467

2. Cluster matching

As shown in Table 2, there are 16,154 and 21,636 instances of data from 2014 and 2015 respectively remaining after filtration. According to the previous analysis, the proportions of empty tag values in these data are 29.26% and 26.43%. The 10,445 records of empty tag values in the two years were adopted as the current set of experimental data.

P-DBSCAN is used as a comparison algorithm. P-DBSCAN requires multiple parameters to be set manually. Generally speaking, the parameter that facilitates relatively even distribution of clustering results is usually selected as the optimum parameter after several experiments are conducted [21]. After many comparison experiments were conducted, the parameters of P-DBSCAN could be finally determined as EPS = 600 m and MinOwners = 5. The algorithm proposed in this paper is named clustering algorithm based on geographic grid partition. Centering on Tian'anmen Square in Beijing, China, we obtained Flickr data for a square area with a side length of 32 km. We divide this square area into 32*32 grids with a side length of 1 km and treat the data in each grid as a subset of the data. The Havg average distance [52] is defined as the distance of the hierarchical clustering. In our experiments, it is assumed that the distance between the clustering center and POI is smaller than 150 meters, and that all shooting sites contained in this cluster are related to tourism. The comparison algorithm and the algorithm proposed in this paper are implemented using C++. The configuration of the running platform is as follows: Intel i5 2.5 Ghz processor, 8 GB memory, Win10 system. Experimental results are presented in Table 6. Since our proposed algorithm utilizes multi-thread parallel computing, its efficiency is nearly six times greater than that of P-DBSCAN, while the clustering results of the two algorithms are similar. It can therefore be seen that our proposed algorithm not only ensures clustering quality, but also improves the operating efficiency.

Table 6. Execution time of two methods.

Algorithm	Execution Time (ms)	Number of Matching Records
Clustering algorithm based on geographic grid partition	16,328	897,172
P-DBSCAN	91,736	829,448

3.3. Data Correlation Analysis

The number of inbound tourists was summarized by month using preprocessed data in Section 3.2, and normalization was also conducted. Finally, the time series data of inbound tourist flows were generated. To confirm whether the Flickr geo-tagged photo metadata is correlated with the actual inbound tourist flow, the data from 2014 to 2015 is taken as an example for use in calculating the cosine distance of the data after normalization. The cosine distance [53] adopts the cosine of two vector angles in the vector space to measure the difference between two individual vector, as follows:

$$\cos\alpha = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}, \quad (6)$$

In the above computational equation, A_i and B_i stand for two vectors, while $\cos\alpha$ stands for represents the cosine distance. When the cosine of two vector angles is 1, the two vectors coincide with each other completely; when it is close to 1, the two vectors are similar. The smaller the cosine value, the less correlated the two vectors are. The two vectors calculated by cosine distance are as follows: inbound tourist flow data obtained from the preprocessed Flickr geo-tagged photo metadata and inbound tourist flow data in the official statistical yearbook. For a certain year, the above data is a sequence of 12 values, which represent the normalized results of the number of inbound tourists per month for 12 months. Finally, we calculated that the cosine distance between the preprocessed data and the actual statistical yearbook data is 0.9198486; this shows that a strong correlation exists between the preprocessed Flickr geo-tagged photo metadata and the actual statistics, and further proves the feasibility of engaging in tourism forecasting based on Flickr geo-tagged photo metadata.

3.4. Forecast Results Analysis

We conducted the following two sets of experiments to prove the feasibility and accuracy of our proposed methods from a prediction accuracy perspective. Data preprocessed according to the methods outlined above, along with statistical yearbook data on inbound tourism in Beijing were adopted as the experimental data. We use a cross-validation approach: two types of data from 2014 to 2015 were used as training data, while data for the six months from January 2016 to June 2016 is taken as the test sample. Tourism forecasting is conducted using several methods, namely NN, SVR, and ELM. RBF (Radial Basis Function) is selected as the kernel function of the SVR, and the parameters are as follows: sigma = 0.1 and C = 10; the manually set parameter for NN is nhid = 10; the manually set parameter of ELM is nhid = 12. A sigmoid function is taken as the activation function. The above algorithms are implemented in Python sklearn.

Four evaluation indexes—the mean absolute error (MAE), root-mean-square error (RMSE), mean absolute percentage error (MAPE), and correlation coefficient (R) between the forecast value and actual value—are adopted to facilitate a comparison of the differences between all prediction models. The specific definitions of these indexes are as follows:

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)|, \quad (7)$$

$$\text{RMSE}=\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n}, \quad (8)$$

$$\text{MAPE}=\frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)/y_i| \times 100\%, \quad (9)$$

$$\text{R}=\frac{\sum_{i=1}^{n} y_i\hat{y}_i}{\sqrt{\sum_{i=1}^{n} {y_i}^2}\sqrt{\sum_{i=1}^{n} {\hat{y}_i}^2}}, \quad (10)$$

In the above equations, y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of forecast samples. MAE, RMSE, and MAPE reflect the deviation between the actual value and the predicted value. The lower the value, the closer the actual value will be to the predicted value, resulting in higher prediction accuracy. Moreover, R takes the correlation between the actual value and the predicted value into account: the closer R is to 1, the higher the correlation between the actual value and the predicted value will be, resulting in higher prediction accuracy.

Table 7 presents the prediction accuracy of the three types of prediction algorithms on the preprocessed Flickr geo-tagged photo metadata.

Table 7. Forecast results based on Flickr geo-tagged photo metadata.

Forecast Model	Kernel Function/ Activation Function	Parameters	MAE	RMSE	MAPE	R
NN	/	*nhid* = 10	0.1151518	0.1388751	7.528479e-05	0.8205143
SVR	RBF	*sigma* = 0.1,*C* = 10	0.1184485	0.1451613	0.003048146	0.8168386
ELM	sigmoid	*nhid* = 12	0.0769539	0.0957949	0.0007842781	0.9188205

Table 8 lists the prediction results of the three types of prediction algorithms under the same parameters on the officially published statistical yearbook. Through comparing the contents of Tables 7 and 8, we can determine that the three prediction algorithms with the same parameters are similar in terms of prediction accuracy on the preprocessed Flickr geo-tagged photo metadata and the statistical yearbook data. We can therefore conclude that the preprocessed Flickr data in this paper can be applied to the prediction of monthly urban tourist inflow data when the prediction models and parameter settings are the same, and that relatively satisfying prediction results can be derived. Moreover, the ELM algorithm's prediction accuracy is higher than that of the other algorithms; accordingly, it is not only suitable for processing statistical yearbook data, but also for processing the alternative data presented in this paper.

Table 8. Forecast results based on statistical yearbook data.

Forecast Model	Kernel Function/ Activation Function	Parameters	MAE	RMSE	MAPE	R
NN	/	*nhid* = 10	0.1710794	0.1986154	0.06896336	0.8067608
SVR	RBF	*sigma* = 0.1,*C* = 10	0.1869958	0.2210619	0.06836812	0.7692361
ELM	sigmoid	*nhid* = 12	0.1069969	0.1306419	0.06172149	0.9541696

4. Discussion

Geo-tagged photo metadata contained in EXIF data have provided a large body of research data for use in tourism studies. This paper proposed the use of geo-tagged photo metadata for inbound tourism flow prediction. To obtain inbound tourism data from geo-tagged photo metadata, several data processing methods were introduced and designed. The geo-tagged photo metadata analysis method has been widely studied and applied in previous related research. As far as the data screening method is concerned, Sun et al. [39] applies the entropy filtering method to distinguish between the shooting data of residents and tourists, and thus implements a tourism recommendation system based on geo-tagged photo metadata. This paper also implements this data screening method to distinguish between domestic and foreign tourists whose registration information is empty, and subsequently verifies the correctness of this method via experiments. In terms of text processing technology, moreover, Hu et al. [54] designed a text processing method for generating text summaries of online travel forums and social network comment data. Furthermore, Miah et al. [14] considers that the text in geo-tagged photo data contains specific keywords, which may reflect certain priorities, interests, and/or motivations when visitors take photos. This article uses text processing software to analyze such text. In this paper, we use word2vec [51] and TF–IDF [42] to achieve the similarity calculation of tag text and further obtain the tourism-related data through the similarity calculation results. The geographical location clustering method has also received widespread attention from researchers. Li et al. [13] pointed out that the clustering methods for analyzing the metadata in tourism research can be classified into three main categories: namely, centroid-based, density-based, and connectivity-based methods. At the same time, the research also points out that the centroid-based clustering method requires multiple scans of data to find the best clustering center, resulting in the inefficient processing of big data. Density-based clustering methods have been deemed more suitable for geo-tagged photo big data. Among these methods, DBSCAN [22] is a representative density-based clustering algorithm; however, additional research has shown that P-DBSCAN [21] is more efficient for clustering geo-tagged photo data. Connection-based clustering methods have also been applied in tourism research. For example, Oender et al. [55] used the connectivity-based clustering method to study the merging of multi-destination itineraries based on Flickr geo-tagged photo data. In this paper, an efficient clustering method based on geographic grid partition was designed for cases in which the tag values were empty. This approach successfully integrates the advantages of density-based and connection-based clustering methods, and thus provides an effective solution for the clustering of geo-tagged photo big data.

The Flickr data for the Chinese capital of Beijing was taken as the study case in this paper to realize relevant data preprocessing operations. Similarity calculation was conducted between geo-tagged photo metadata after preprocessing, as well as actual statistical yearbook data on inbound tourist flows. It was thereby proven that, following preprocessing, this data could approximately substitute for the real inbound tourist flow data. In addition, through the use of three typical machine learning prediction models (SVR, NN, and ELM), it was verified that geo-tagged photo metadata after preprocessing could be used to achieve a prediction accuracy equivalent to that of actual inbound tourist flow-related statistical yearbook data.

In this paper, a type of new, reliable, and low-cost data has been identified and used for the study of inbound tourist flow prediction. Moreover, a simple and effective solution is proposed for conducting tourist flow prediction for regions lacking in historical data. The Flickr data is a typical EXIF data. As the data processing methods proposed in this paper are applicable to all EXIF data, they can be generalized to studies of other network sharing platform data. In addition, the research contents of this paper involve only the processing of inbound tourist flow data. The data screening scheme proposed in this paper can also be applied to obtain local tourist data. By contrast, local visitor data does not need to consider foreigners who temporarily reside in tourist destinations; therefore, the processing of such data is slightly less difficult than the research content of this paper.

It should be noted that there are still some shortcomings of the present research. Firstly, the sample distribution of geo-tagged photo metadata is unbalanced, meaning that not all regions have sufficient samples available. However, as noted in the research achievements of [56], the digital footprints of city tourists exist across several data sources; the data from these data sources can be complementary, and can thus represent one or more tourist activities when considered together. Therefore, city tourists can be analyzed based on several data sources, meaning that the defect of the single data source can be addressed. In addition, there is still scope for further research on the processing of text labels. For example, this article does not consider the similarity between words when building a dictionary of stop words. Similar words can be obtained by calculating the distance between words; however, simply selecting a high-frequency word that is unrelated to the subject of this study as a stop word will result in similar words with low word frequency being ignored. Thus, we can use N-Gram[57], Word2vec [51], etc., to calculate the distance between words in order to achieve a more accurate stop word dictionary.

Additionally, the following issues merit further study in future work:

- Geo-tagged photo metadata contains multiple types of information: text tags, geotags, image contents, and time of image capture [14]. The research scope of this paper did not extend to cover the analysis of image contents. The image analysis techniques represented by in-depth learning algorithms have already obtained promising new results. If the content of images can be analyzed using effective image analysis techniques of this kind, relevant decision support information can be acquired.
- Multi-level tourism demand analysis has attracted some interest from decision-makers in the field, as this approach can provide information that is more detailed and diversified than the total tourism demand [1]. Tourists from different countries and regions can be screened out according to the registered addresses of users on the network sharing platform. From this, we can conduct research on tourism behavior analysis, tourism recommendation, and tourism flow forecasting for tourists from different sources.
- The inbound tourism industry is a very sensitive industry; various emergencies, along with global economic and political turmoil, could very well exert a huge influence on the inbound tourism industry [1]. In the future, a study could be conducted to assess tourist flow prediction in emergency situations based on geo-tagged photo metadata.

In future research, we hope to continue to extend the work conducted in this paper. We will study the processing of geo-tagged photo metadata and other complementary digital footprint data (Twitter, Weibo, etc.) [56] from different spatial granularities (countries, cities, tourist attractions, etc.) in order to provide a reliable, low-cost data source for tourism research.

Author Contributions: Conceptualization, Wen Chen, Zhiyun Xu and Yonglong Luo; Methodology, Wen Chen; Software, Wen Chen; Validation, Wen Chen and Xiaoyao Zheng; Data curation, Wen Chen; Writing—original draft preparation, Wen Chen; Writing—review and editing, Wen Chen , Zhiyun Xu and Yonglong Luo ; Upervision, Yonglong Luo; Pproject administration, Yonglong Luo; Funding acquisition, Wen Chen, Xiaoyao Zheng and Yonglong Luo.

Funding: This research was funded by the National Natural Science Foundation of China under Grants 61972439, 61672039, 61702010, 61772034, Tongling College Talent Research Fund Project under Grants 2017tlxyrc07, Natural Science Foundation of Anhui Province: 1808085MF172.

Acknowledgments: The authors would like to acknowledge all of reviewers and editors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Song, H.; Li, G. Tourism demand modelling and forecasting—A review of recent research. *Tour. Manag.* **2008**, *29*, 203–220.
2. Mckercher, B.; Lau, G. Movement Patterns of Tourists within a Destination. *Tour. Geogr.* **2008**, *10*, 355–374.
3. Asakura, Y.; Iryo, T. Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. *Transp. Res. Part Policy Pract.* **2007**, *41*, 684–690.
4. Vu, H.Q.; Li, G.; Law, R.; Ye, B.H. Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. *Tour. Manag.* **2015**, *46*, 222–232.
5. *Technical Handbook on the Collection and Presentation of Domestic and International Tourism Statistics*; World Tourism Organization: Madrid, Spain,1981.
6. Girardin, F.; Fiore, F.D.; Ratti, C.; Blat, J. Leveraging explicitly disclosed location information to understand tourist dynamics: A case study. *J. Locat. Based Serv.* **2008**, *2*, 41–56.
7. Jones, B. Reforming the System? In *A Review of Australian Tourism Statistics*; Bureau of Tourism Research: Canberra, Australia, 1996.
8. Finn, M.; Walton, M.; Elliott-White, M. *Tourism and Leisure Research Methods: Data Collection, Analysis, and Interpretation*; Pearson Education: Essex, England, 2000.
9. Qin, J.; Li, L.P.; Tang, M.D. Exploring the spatial characteristics of Beijing inbound tourist flow based on geotagged photos. *Acta Geogr. Sin.* **2018**, *73*, 1556–1570.
10. Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S.; Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature* **2009**, *457*, 1012.
11. Yang, X.; Pan, B.; Evans, J.A.; Lv, B. Forecasting Chinese tourist volume with search engine data. *Tour. Manag.* **2015**, *46*, 386–397.
12. Li, X.; Pan, B.; Law, R.; Huang, X. Forecasting tourism demand with composite search index. *Tour. Manag.* **2017**, *59*, 57–66.
13. Li, J.; Xu, L.; Tang, L.; Wang, S.; Li, L. Big data in tourism research: A literature review. *Tour. Manag.* **2018**, *68*, 301–323.
14. Miah, S.J.; Vu, H.Q.; Gammack, J.; McGrath, M. A Big Data Analytics Method for Tourist Behaviour Analysis. *Inf. Manag.* **2017**, *54*, 771–785.
15. Kisilevich, S.; Krstajic, M.; Keim, D.; Andrienko, N.; Andrienko, G. Event-Based Analysis of People's Activities and Behavior Using Flickr and Panoramio Geotagged Photo Collections. In Proceedings of the IEEE 2010 14th International Conference Information Visualisation, London, UK, 26–29 July 2010; pp. 289–296.
16. Lucchese, C.; Perego, R.; Silvestri, F.; Vahabi, H.; Venturini, R. How Random Walks Can Help Tourism. In Proceedings of the Advances in Information Retrieval, Barcelona, Spain, 1–5 April 2012; pp. 195–206.
17. Mou, N.; Yuan, R.; Yang, T.; Zhang, H.; Tang, J.; Makkonen, T. Exploring spatio-temporal changes of city inbound tourism flow: The case of Shanghai, China. *Tour. Manag.* **2020**, *76*, 103955.
18. Centobelli, P.; Ndou, V. Managing customer knowledge through the use of big data analytics in tourism research. *Curr. Issues Tour.* **2019**, *22*, 1862–1882.

19. Mariani, M.; Baggio, R.; Fuchs, M.; Hoeepken, W. Business intelligence and big data in hospitality and tourism: A systematic literature review. *Int. J. Contemp. Hosp. Manag.* **2018**, *30*, 3514–3554.
20. Peng, X.; Huang, Z.A. Novel Popular Tourist Attraction Discovering Approach Based on Geo-Tagged Social Media Big Data. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 216.
21. Kisilevich, S.; Mansmann, F.; Keim, D. P-DBSCAN: A Density Based Clustering Algorithm for Exploration and Analysis of Attractive Areas Using Collections of Geo-tagged Photos. In Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application, Washington, DC, USA, 21–23 June 2010; pp. 381–384.
22. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD* **1996**, *96*, 226–231.
23. Zhang, K.; Chen, Y.; Li, C. Discovering the tourists' behaviors and perceptions in a tourism destination by analyzing photos' visual content with a computer deep learning model: The case of Beijing. *Tour. Manag.* **2019**, *75*, 595–608.
24. Xia, C.; Schwartz, R.; Xie, K.; Krebs, A.; Langdon, A.; Ting, J.; Naaman, M. CityBeat: Real-time Social Media Visualization of Hyper-local City Data. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 7–11 April 2014; pp. 167–170.
25. Pantano, E.; Dennis, C. Store buildings as tourist attractions: Mining retail meaning of store building pictures through a machine learning approach. *J. Retail. Consum. Serv.* **2019**, *51*, 304–310.
26. Giglio, S.; Bertacchini, F.; Bilotta, E.; Pantano, P. Machine learning and point of interests: Typical tourist Italian cities. Curr. Issues Tour. 2019, 1–13, doi:10.1080/13683500.2019.1637827
27. Kou, N.M.; Hou, L.U.; Yang, Y.; Gong, Z. Travel topic analysis: A mutually reinforcing method for geo-tagged photos. *Geoinformatica* **2015**, *19*, 693–721.
28. Chen, L.; Gao, Y.; Xing, Z.; Jensen, C.S.; Chen, G. I2RS: A Distributed Geo-Textual Image Retrieval and Recommendation System. *Proc. Vldb Endow.* **2015**, *8*, 1885–1888.
29. Xu, Z.; Chen, L.; Guo, H.; Lv, M.; Chen, G. User similarity-based gender-aware travel location recommendation by mining geotagged photos. *Int. J. Embed. Syst.* **2018**, *10*, 356–365.
30. Su, S.; Wan, C.; Hu, Y.; Cai, Z. Characterizing geographical preferences of international tourists and the local influential factors in China using geo-tagged photos on social media. *Appl. Geogr.* **2016**, *73*, 26–37.
31. Koylu, C.; Zhao, C.; Shao, W. Deep Neural Networks and Kernel Density Estimation for Detecting Human Activity Patterns from Geo-Tagged Images: A Case Study of Birdwatching on Flickr. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 45.
32. Yang, L.; Wu, L.; Liu, Y.; Kang, C. Quantifying Tourist Behavior Patterns by Travel Motifs and Geo-Tagged Photos from Flickr. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 345.
33. Li, L.-J.; Jha, R.K.; Thomee, B.; Shamma, D.A.; Cao, L.; Wang, Y. Where the Photos Were Taken: Location Prediction by Learning from Flickr Photos. Advances in Computer Vision and Pattern Recognition. In *Large-Scale Visual Geo-Localization*; Zamir, A.R., Hakeem, A., Van Gool, L., Shah, M., Szeliski, R., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 41–58.
34. Baraglia, R.; Muntean, C.I.; Nardini, F.M.; Silvestri, F. LearNext: Learning to Predict Tourists Movements. In Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management, San Francisco, CA, USA, 27 October–1 November 2013; pp. 751–756.
35. Arain, Q.A.; Memon, H.; Memon, I.; Memon, M.H.; Shaikh, R.A.; Mangi, F.A. Intelligent travel information platform based on location base services to predict user travel behavior from user-generated GPS traces. *Int. J. Comput. Appl.* **2017**, *39*, 155–168.
36. Cai, G.; Lee, K.; Lee, I. *A Framework for Mining Semantic-Level Tourist Movement Behaviours from Geo-Tagged Photos. In Proceedings of the AI 2016: Advances in Artificial Intelligence*; Kang, B.H., Bai, Q., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 519–524.
37. Peng, B.; Song, H.; Crouch, G.I.; Witt, S.F. A Meta-Analysis of International Tourism Demand Elasticities. *J. Travel Res.* **2015**, *54*, 611–633.
38. Kulendran, N.; Shan, J. Forecasting China's Monthly Inbound Travel Demand. *J. Travel Tour. Mark.* **2002**, *13*, 5–19.
39. Sun, Y.; Fan, H.; Bakillah, M.; Zipf, A. Road-based travel recommendation using geo-tagged images. *Comput. Environ. Urban Syst.* **2015**, *53*, 110–122.
40. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv: abs/1301.3781.

41. Weiss, S.M.; Indurkhya, N.; Zhang, T.; Damerau, F. *Text Mining: Predictive Methods for Analyzing Unstructured Information*, 1st ed.; Springer Publishing Company, Incorporated: New York, United States, 2010.
42. Wei, X. *Introduction to Tourism*; China Forestry Publishing House: Beijing, China, 2000.
43. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2013; pp. 3111–3119.
44. Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *Acm Trans. Intell. Syst. Technol.* **2011**, *2*, 27.
45. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: A new learning scheme of feedforward neural networks. *Neural Netw.* **2004**, *2*, 985–990.
46. Cheng, G.-J.; Cai, L.; Pan, H.-X. Comparison of Extreme Learning Machine with Support Vector Regression for Reservoir Permeability Prediction. In Proceedings of the 2009 International Conference on Computational Intelligence and Security, Beijing, China, 11–14 December 2009; pp. 173–176.
47. Huang, G.-B.; Wang, D.H.; Lan, Y. Extreme learning machines: A survey. *Int. J. Mach. Learn. Cybern.* **2011**, *2*, 107–122.
48. Hassan, S.; Khosravi, A.; Jaafar, J.; Khanesar, M.A. A systematic design of interval type-2 fuzzy logic system using extreme learning machine for electricity load demand forecasting. *Int. J. Electr. Power Energy Syst.* **2016**, *82*, 1–10.
49. Sokolov-Mladenovic, S.; Milovancevic, M.; Mladenovic, I.; Alizamir, M. Economic growth forecasting by artificial neural network with extreme learning machine based on trade, import and export parameters. *Comput. Hum. Behav.* **2016**, *65*, 43–45.
50. China Tourism Academy. *Annual Report of China Inbound Tourism Development 2013*; Tourism Education Press: Beijing, China, 2013; pp. 2–3.
51. Wu, M.-Y.; Wall, G.; Tong, Y. Research on China's Inbound Tourism: A Comparative Review. *J. China Tour. Res.* **2019**, *15*, 320–339.
52. Johnson, S.C. Hierarchical clustering schemes. *Psychometrika* **1967**, *32*, 241–254.
53. Sarkar, I.N. A vector space model approach to identify genetically related diseases. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 249–254.
54. Hu, Y.-H.; Chen, Y.-L.; Chou, H.-L. Opinion mining from online hotel reviews—A text summarization approach. *Inf. Process. Manag.* **2017**, *53*, 436–449.
55. Oender, I. Classifying multi-destination trips in Austria with big data. *Tour. Manag. Perspect.* **2017**, *21*, 54–58.
56. Henar Salas-Olmedo, M.; Moya-Gomez, B.; Carlos Garcia-Palomares, J.; Gutierrez, J. Tourists' digital footprint in cities: Comparing Big Data sources. *Tour. Manag.* **2018**, *66*, 13–25.
57. Suen, C.Y. N-Gram Statistics for Natural Language Understanding and Text Processing. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *1*, 164–172.