sustainability

MDPI

Article

A Methodological Workflow for Deriving the Association of Tourist Destinations Based on Online Travel Reviews: A Case Study of Yunnan Province, China

Tao Liu [1,2], Ying Zhang [3,*], Huan Zhang [2] and Xiping Yang [4,5]

1 College of Resources and Environment, Henan University of Economics and Law, Zhengzhou 450002, China; liutao@huel.edu.cn
2 Key Laboratory of New Materials and Facilities for Rural Renewable Energy (MOA of China), Henan Agricultural University, Zhengzhou 450002, China; zhanghuan5754@163.com
3 College of Economics and Management, Henan Agricultural University, Zhengzhou 450046, China
4 School of Geography and Tourism, Shaanxi Normal University, Xi'an 710119, China; xpyang@snnu.edu.cn
5 Shaanxi Key Laboratory of Tourism Informatics, Xi'an 710119, China
* Correspondence: nongdazhangying@126.com

Abstract: Insights into the association rules of destinations can help to understand the possibility of tourists visiting a destination after having traveled from another. These insights are crucial for tourism industries to exploit strategies and travel products and offer improved services. Recently, tourism-related, user-generated content (UGC) big data have provided a great opportunity to investigate the travel behavior of tourists on an unparalleled scale. However, existing analyses of the association of destinations or attractions mainly depend on geo-tagged UGC, and only a few have utilized unstructured textual UGC (e.g., online travel reviews) to understand tourist movement patterns. In this study, we derive the association of destinations from online textual travel reviews. A workflow, which includes collecting data from travel service websites, extracting destination sequences from travel reviews, and identifying the frequent association of destinations, is developed to achieve the goal. A case study of Yunnan Province, China is implemented to verify the proposed workflow. The results show that the popular destinations and association of destinations could be identified in Yunnan, demonstrating that unstructured textual online travel reviews can be used to investigate the frequent movement patterns of tourists. Tourism managers can use the findings to optimize travel products and promote destination management.

Keywords: online travel review; user-generated content; association rule; movement pattern of tourist

Citation: Liu, T.; Zhang, Y.; Zhang, H.; Yang, X. A Methodological Workflow for Deriving the Association of Tourist Destinations Based on Online Travel Reviews: A Case Study of Yunnan Province, China. *Sustainability* **2021**, *13*, 4720. https://doi.org/10.3390/su13094720

Academic Editor: Chia-Lin Chang

Received: 10 March 2021
Accepted: 21 April 2021
Published: 23 April 2021

1. Introduction

Spatial movement is an essential behavior of tourism activities. Tourist movement involves time, space, place, and scale, which are the basic elements of tourism geography. Tourist travel behavior can potentially imply the popularity of tourist attractions and the correlation among destinations. Moreover, investigating tourist travel behavior can help uncover the intrinsic characteristics of how tourists design their itineraries, thereby helping tourism agencies and industries in planning destination facilities, assessing tourism products, and exploiting tourism resources. Therefore, tourist movement patterns have been an important research topic in tourism geography.

Traditional approaches in investigating tourist movement patterns and destination characteristics usually utilize questionnaires, but the collection of this dataset is costly and time consuming [1]. Moreover, this method is limited in sample size and space–time resolution, making the analysis of tourist travel behavior from a comprehensive and broad perspective difficult. Fortunately, with the rapid development of information and the internet, numerous social media websites and applications (apps) allow tourists to share their own experiences and feelings (e.g., reviews or comments on a tourist attraction or

destination) about their travel [2–5]. These tourism-related user-generated contents (UGC) can be considered a valuable data source and open up new horizons for researchers to understand tourists' travel experiences well and create smart urban tourism [6,7]. UGC big data can be classified into two categories: (1) geo-tagged UGC, which is produced using location-aware devices that record the location information of tourists when they post their travel experiences as comments or photos on social apps (e.g., Twitter, Flickr, and Instagram); and (2) unstructured textual content without location coordinate information on public travel service websites (e.g., TripAdvisor and Ctrip), which allows tourists to share their comments about the quality of service and release reviews of their travel experience.

Geo-tagged UGC has received widespread attention from researchers in the fields of tourism, geography, and computer science because of its advantage in tracking the spatial and temporal activities of tourists [8,9]. The literature includes detecting tourism destinations or districts [10–12], characterizing tourist flows among destinations [13–15], visualizing the spatial and temporal patterns of tourists [16–19], and developing the recommendation model for tourist routes or attractions [20–22]. These studies show the powerful potential of geo-tagged big data in grasping insights into the spatial characteristics of tourist movement and destination correlation on an unparalleled spatial and temporal scale.

Unstructured textual UGC generates a body of descriptive texts, including the comments or reviews of travel experiences, implying the immediate perception of tourists on destinations or travel products [3,23]. For example, these online texts can be utilized to understand destination branding or image [24–26], identify the unique or specific attribute of destinations [27,28], understand the cooperation or similarities of attractions [29,30], explore tourist movement patterns [31], and analyze tourist sentiments [32,33]. In addition, based on the comments, tourism managers will understand how tourists and customers evaluate their service (i.e., electronic word-of-mouth (eWOM)), thereby giving them ideas on how to make their management or service targeted and intelligent [34–38]. Although the second data type does not track the variation in tourist locations, it can reflect the tourists' perception of travel activities.

Nowadays, tourists tend to plan a long journey and visit more destinations with the permission of time, economy, and physical condition. Moreover, the "time–space compression" effect brought by advanced transportation expands the radiation range of tourism and makes it possible to visit more destinations during a tour on a large spatial scale; thus, multi-destination tourism has now become a popular travel mode [39]. Characterizing tourist movement patterns among multiple destinations will help to understand the interaction among destinations, further helping to predict the next destination [40]. Therefore, further research on multi-destination relationships is necessary. In tourism, the association rules of tourist destinations, which can be embodied from movement patterns, can quantify the possibility of tourists visiting a destination after having traveled from another. A further understanding of such rules can help predict the destinations of tourists, thus benefiting the tourism economy. In addition, the association of destinations indicates the popular destination sets and their association. Therefore, tourism practitioners and managers can use the association rules to generate targeted strategies and travel products to promote destination management and provide improved services for tourists.

Currently, only a few studies have focused on tourism-related rules from UGC data. Rong et al. [41] implemented behavioral analysis on the association between web sharers and browsers and revealed the direct influence of eWOM. Based on geo-tagged UGC characteristics (e.g., geo-tagged photos, bluetooth tracking data), popular tourist attractions or destinations can be identified, and frequent mobility or sequential patterns can be extracted from geo-tagged travel diaries through association rule learning to understand the travel behavior and preferences of tourists [42]. In terms of data sciences in digital marketing, Saura (2021) presented a holistic overview of the framework, method, research topics, and performance metrics, and claimed that although the use of data sciences for decision-making and knowledge discovery has remarkably increased, the management of data sciences in digital marketing remained scarce [43]. Therefore, extracting knowledge from

user-generated dataset could provide useful strategies for improving tourism management and digital marketing.

At present, most of the studies only utilized the unstructured textual UGC data (e.g., online travel reviews) to understand the tourists' mental response to travel activities. To the best of our knowledge, its application in investigating spatial multi-destination association has not been exploited in tourism research due to its limitation in the access location information of tourist destinations. Based on this research gap, the main aim of this study is to extract the association characteristics of destinations from unstructured textual UGC data, and intends to answer the research questions: how the unstructured textual UGC data could be used to quantify the spatial association rules among tourist destinations such as geo-tagged UGC data.

Therefore, this study investigates whether unstructured online textural UGC data can be used to understand the association among destinations (frequent rules among destinations) and extend the usage of unstructured UGC data from perceiving mental travel experiences to understanding the spatial movement patterns of tourists. Moreover, this study aims to exploit a new path to excavate the frequent association of destinations from unstructured online travel reviews. First, we develop a crawling program to collect popular destination and online travel reviews from a public commercial travel service website. Then, a text-matching algorithm is used to identify the destination sequence for each travel review in accordance with the co-occurrence of the destinations. Finally, we describe the main principle of the association rule learning method to derive the frequent travel patterns from the extracted destination sequence sets. The province of Yunnan in China is used for a case study to demonstrate the feasibility of the proposed method and gather insights into the travel behavior of tourists and the association of destinations.

The main originality and contribution of this study could be drawn from twofold. First, a methodological contribution is that we develop a workflow from collecting travel reviews to identify association rules among destinations from the text. Second, an empirical case study is conducted to help understand the spatial association among the main tourist destinations of Yunnan province in China. The remainder of the paper is organized as follows: Section 2 introduces the study area of Yunnan province. Section 3 describes the methodological workflow including dataset collection, extraction of destination sequences and mining association rules. The research results are shown in Section 4. Section 5 discusses the main findings. Finally, the conclusion is presented in Section 6.

2. Study Area

Yunnan Province (capital: Kunming) is located at the southwest border of China (Figure 1). It covers more than 390,000 square kilometers and includes 129 administrative counties. Recently, Yunnan has become one of the top tourist destinations in China because of the following merits: (1) The terrain of Yunnan is a mountainous plateau with an average elevation of 2000 m. Yunnan has many mountains, forests, lakes, and rivers, thus having many graceful natural resources and beautiful sceneries; (2) The climate is comfortable, and the annual temperature difference is small; hence, tourists can visit any time of the year; (3) The province has many historical and cultural resources because it has the most ethnic groups among all provinces of China. Yunnan is composed of different kinds of ethnic cultures with colorful customs, thus attracting many tourists. These abundant tourism resources attract tourists worldwide. According to statistics, more than 6.6 million tourists from overseas visit this province, generating more than 3.5 billion dollars in revenue in 2017. In addition, more than 560 million domestic tourists traveled to Yunnan in 2017, resulting in a revenue of more than 668.2 billion yuan. In recent years, tourism has become a new driving force to promote the economic development of Yunnan. Therefore, understanding the association among tourist destinations within the province is important for administrators and tourism agencies to develop strategies that can provide improved services for tourists.

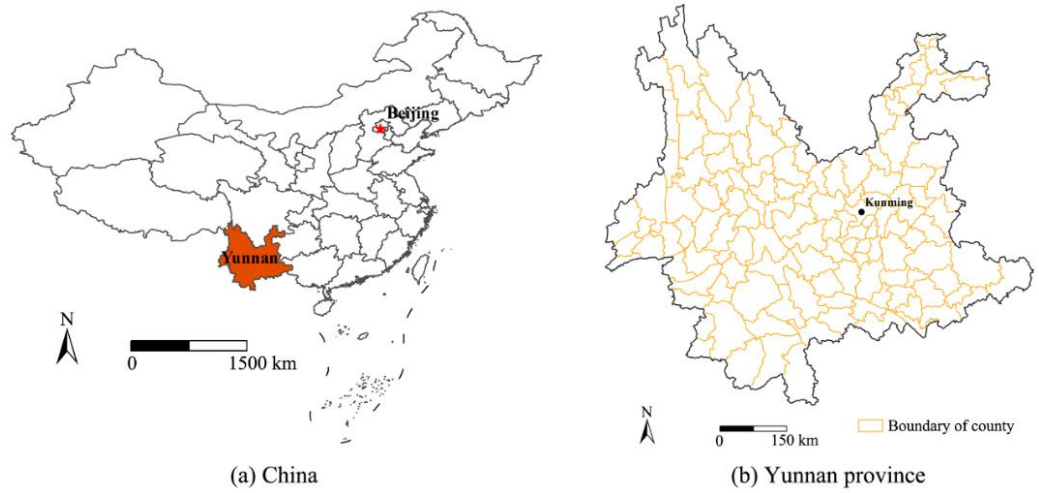

Figure 1. The study area of this research. (**a**) The China and (**b**) Yunnan province.

3. Methodology

In this section, we describe the methodological workflow of the study. First, we develop a crawler program to collect an online review dataset from an open-access tourism website. Second, we extract each tourist's travel destination sequences from the reviews. Finally, we present the main principle of mining association rules among popular tourist destinations. We implement the methodological workflow based on popular Python programming language.

3.1. Data Collection

The online review data used in this study is collected from Ctrip (https://www.ctrip.com/ accessed on 22 April 2021). This website is one of the largest internet platforms for Chinese tourists that provide full-scale services, including the list of attractions in a destination, ticket and hotel bookings, and popular travel route recommendations. Ctrip also allows tourists to leave their comments on the attractions or destinations and upload their travel photos, stories, or reviews, thereby providing a reference for other tourists who intend to travel to the same places. Although in text or photo format, the review usually records the travel experiences of tourists in detail, making it possible to find the destinations that were visited by the reviewer. Moreover, the sample size of online reviews is larger than the traditional questionnaire data. Therefore, reviews can be used to understand the association among destinations.

In this study, we develop a web crawling program to download the online reviews of tourists from Ctrip. We input the keyword "Yunnan" in the homepage of Ctrip to search the Yunnan-related homepage, which provides Yunnan's travel-related services (e.g., transport, attraction, accommodation, shopping, etc.). This study mainly focuses on reviews and popular destinations (Figure 2a). We first search the popular destination list in Yunnan and store the name of each destination into a destination set D. For each travel review, we capture information, such as the title, time (month of the tour), number of days in the tour, and content of the travel note (Figure 2b). After collecting this information, we generate a review information set *R*.

(a) Homepage of Yunnan in the Ctrip

(b) Example of a travel review

Figure 2. (**a**) Homepage of Yunnan in Ctrip and (**b**) example of a travel review homepage.

3.2. Extracting the Destination Sequences from Online Travel Reviews

This section introduces the process of extracting the destination sequences for each tourist from their online review. Destination set $D = \{d_1, d_2, \ldots, d_n\}$, where d_i represents the name of the destination, and n is the number of popular destinations in Yunnan. Review set $R = \{r_1, r_2, \ldots, r_l\}$, $r_j = \{t_j, m_j, d_j, c_j\}$, where l is the number of total reviews, and t_j, m_j, d_j, and c_j represent the corresponding title, month, days, and content of the review. The specific process of extracting destination sequences is illustrated in Figure 3. For example, assuming set D has five popular destinations, then c_i and c_j are the contents of reviews i and j (Figure 3a,b). For each destination, we first apply the text matching algorithm to identify the number of appearances of each destination in each content, where the figures in the brackets represent the total number of appearances for the corresponding destination in c_i and c_j (Figure 3c). However, tourists may visit destination a and mention other destinations in their reviews. For example, someone stopped to eat special food in destination b on their way to destination a and may write this experience in their review. In this case, b can be considered an affiliated destination. In general, the number of affiliated destinations mentioned in reviews is often very small, especially if the tourist does not intend to visit these destinations. Therefore, a threshold parameter is used to filter destinations with a small number of mentions and mitigate this issue. In Figure 3, destinations with a number

of appearances less than 2 are excluded to generate the ultimate destination sequences s_i and s_j (Figure 3d). In this manner, we can extract the destination sequences for each travel review in set R.

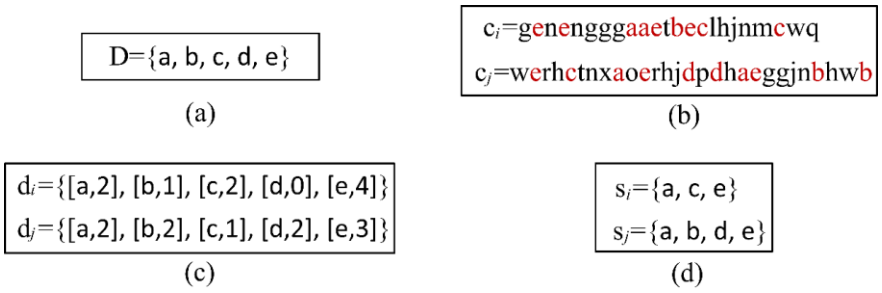

Figure 3. Extracting destination sequences. (**a**) destination sequences; (**b**) contents of reviews;(**c**) total number of appearances for the corresponding destination in c_i and c_j; (**d**) ultimate destination sequences s_i and s_j.

3.3. Mining Association Rules of Tourist Destinations

Association rule learning, which is widely used to identify the frequently purchased combination among commodities from transaction databases, can discover interesting relationships among variables in large databases. For example, rule $\{A, B\} \Rightarrow \{C\}$ indicates that if a customer buys products A and B, they are more likely to buy product C. Currently, association rules have been applied to tourism research to uncover the travel patterns of tourists. Li et al. (2010) incorporated both positive and negative association rules into understanding the HongKong residents' outbound travel characteristics [44]. Lee et al. (2013) applied the clustering and association rules to mine the areas of attraction and their associative patterns [45]. Versichele et al. (2014) present an empirical study on the mining of association rules in tourist attraction visits by using Bluetooth tracking data [46]. Qi and Wong (2014) adopted Apriori algorithm association rules mining to segment Macau's tourists and to predict tourists' preferences for the different local heritage attractions [47]. In addition, the association rule technique could also be utilized to develop a tourism recommendation system [48]. Therefore, it is feasible to transfer the association rule data mining technique to uncover the frequent tourist patterns.

An associate rule can be represented as $X \Rightarrow Y$, where $X, Y \subset I$ and $X \cap Y = \varnothing$, where X and Y are the left-hand side (LHS) and right-hand side (RHS), respectively, and I represents the item set. The associate rule indicates that if item X appears in a transaction, then item Y may appear in the same transaction with a certain probability. In this study, destination set D is the item set, and the transaction database is the extracted destination sequence set $S = \{s_1, s_2..s_m\}$, where $s_i \subseteq D$ represents the destination sequence that tourist i has visited during the tour (s_i is unordered among destinations in the sequence).

Three indicators are used to compare the effectiveness of association rules, namely, support, confidence, and lift. For a destination association rule $d_i \Rightarrow d_j$, the three indicators can be calculated as follows:

$$\begin{aligned} \text{support}(d_i \Rightarrow d_j) &= \frac{frq(d_i, d_j)}{number\ of\ total\ destination\ sequence}, \\ \text{confidence}(d_i \Rightarrow d_j) &= \frac{frq(d_i, d_j)}{frq(d_i)}, \\ \text{lift}(d_i \Rightarrow d_j) &= \frac{\text{support}(d_i \Rightarrow d_j)}{\text{support}(d_i) \times \text{support}(d_j)}. \end{aligned} \tag{1}$$

Equation (1) shows that the support of a rule reflects how frequently the destination union set of LHS and RHS appears in the total extracted destination sequences; the confidence of a rule indicates how frequently the destination union set of LHS and RHS appears in the destination sequences that contain LHS, and the lift of a rule is the ratio of the expected support if LHS and RHS are independent. Furthermore, support, confidence,

and lift quantify the significance, accuracy, and representativeness of the rule, respectively. Figure 4 presents an example of a calculation of the support, confidence, and lift of rule $a \Rightarrow b$. A rule is considered a strong association rule if the support, confidence, and lift are greater than the user-defined minimum thresholds $min_{support}$, $min_{confidence}$, and min_{lift}.

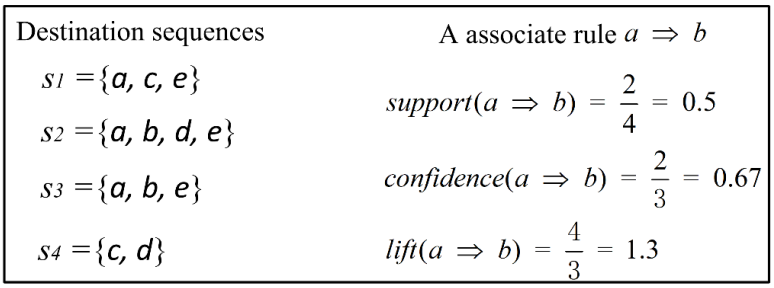

Figure 4. Sample calculation of support, confidence, and lift of a rule, where a, b, c, d, e represent the popular destinations.

4. Results Analysis

4.1. General Statistical Analysis

A total of 66 popular destinations and 12,752 travel reviews were collected from Ctrip. We excluded reviews that have no text description (i.e., photos only) because they are inappropriate for extracting destination sequences. A total of 7875 reviews were assessed and included in set R. Combining with set D, R will be used to extract the destination sequences using the process in Section 3.2 to generate the destination sequence set S.

In accordance with the generated S, statistical analysis is conducted involving D, where the popularity is defined as $p = n_d/n$, n_d represents the number of sequences containing destination d in S, and n is the total number of sequences (n = 7875). We then sort the destinations according to their popularity. Figure 5 shows the spatial distribution and statistical order of destinations according to their popularity. Table 1 lists the top 15 popular destinations. The top five popular destinations are Lijiang, Dali, Kunming, Xianggelila, and Luguhu, which are the main destinations that attract tourists in Yunnan. This result is consistent with Mafengwo, which is another travel social service platform in China, which demonstrates that the online reviews can be utilized to understand the characteristics of tourist destinations.

Table 1. Top 15 popular destinations in Yunnan.

Order	Name	*p*	Order	Name	*p*	Order	Name	*p*
1	Lijiang	0.628	6	Xishuangbanna	0.07	11	Yuanyang	0.017
2	Dali	0.395	7	Deqin	0.069	12	Puer	0.016
3	Kunming	0.311	8	Shilin	0.055	13	Chengjiang	0.014
4	Xianggelila	0.247	9	Tengchong	0.039	14	Jianshui	0.013
5	Luguhu	0.245	10	Mile	0.018	15	Luoping	0.012

We also investigate the temporal characteristics of the destination sequences. Figure 6 displays the statistical percentage of sequences in months, indicating which month the tourists are more likely to visit Yunnan. Results show decentralization from January to December. One possible reason is that the climate is comfortable for tourists to travel any time of the year. July has the maximum proportion of sequences, which is summer vacation in China. Students and parents with children choose to travel to Yunnan during this period. October has the second-largest percentage, which may be due to the Chinese National Days (1–7 October) when a majority of workers take a vacation and plan a trip.

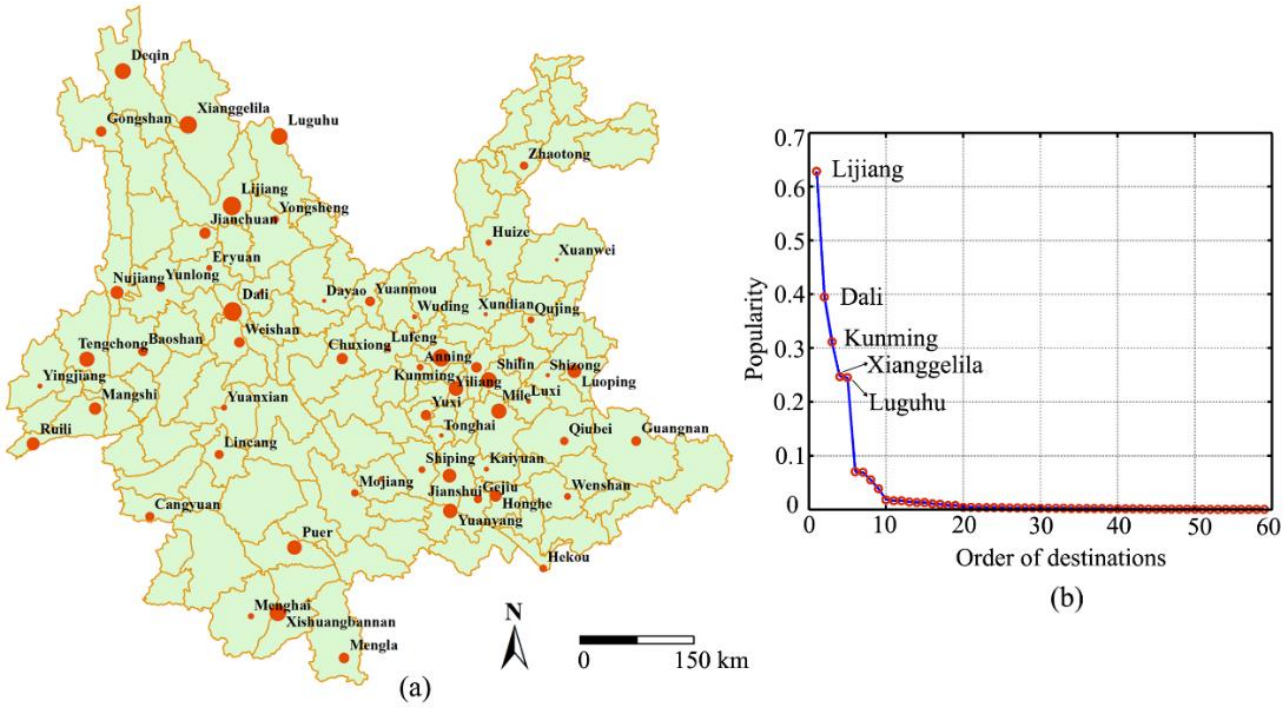

Figure 5. (**a**) Spatial distribution of the popularity of destinations (the bigger the red dot, the more popular the destination); (**b**) Popularity of the destinations.

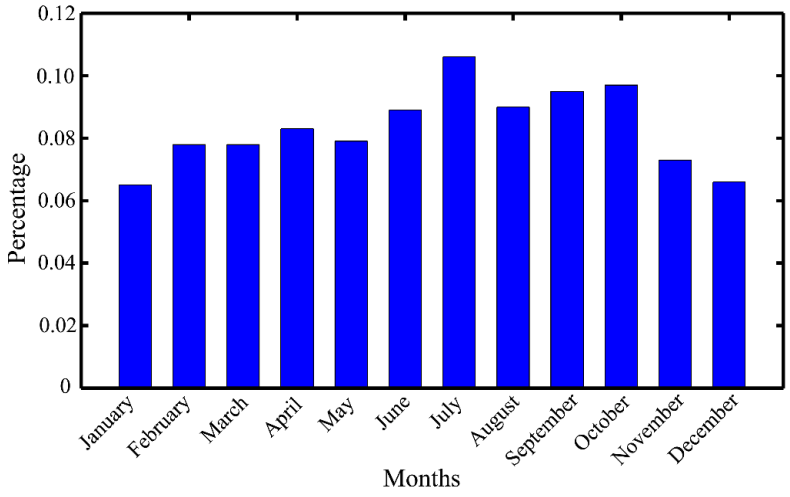

Figure 6. Monthly percentage of tourists.

Figure 7 shows the percentage of the number of days that tourists spend in Yunnan. The results show that approximately 90% of tourists spend a maximum of ten days in Yunnan. The number of tourists who travel for only one day accounts for more than 20%. This percentage may be problematic because we find that some tourists who only stay for one day have reviewed several destinations. Moreover, based on our knowledge of Yunnan, visiting several destinations in the province in one day is impractical for tourists; therefore, this finding might contain some errors. The number of tourists who spend 5, 6, and 7 days in the province is higher than on other days. We also calculate the distribution of tourists according to the number of days spent and number of destinations (Figure 8). We exclude the data for one day to reduce the errors for the following analysis. The findings reveal that most people are more likely to spend less than 10 days visiting five or fewer destinations in Yunnan. Moreover, more than 36% of tourists are willing to spend several days staying in one destination only. This preference may be attributed to the following reasons: (1) a

popular destination usually covers multiple attractive scenic spots, and tourists have to schedule several days to travel to these attractions; and (2) some destinations, such as Lijiang and Dali, are famous for their slow, leisurely, lazy life pace, thereby attracting people who live in large cities to relax and escape from the hustle and bustle of urban life. Therefore, among the tourists who stay in one destination when visiting Yunnan, more than 55% prefer Lijiang or Dali for their destination (Figure 9).

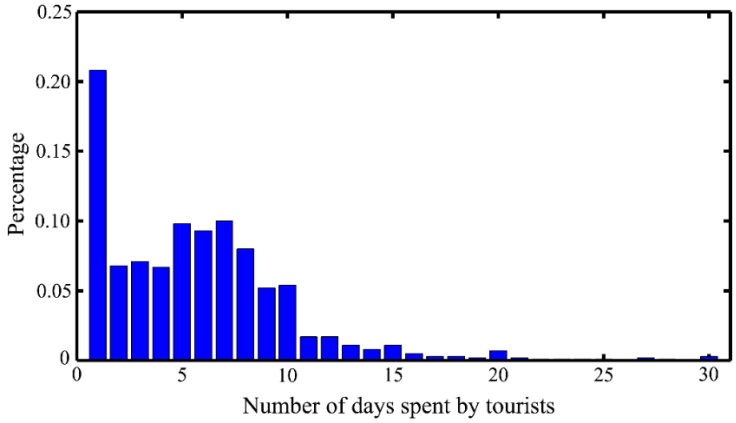

Figure 7. Percentage of days spent by tourists during travel.

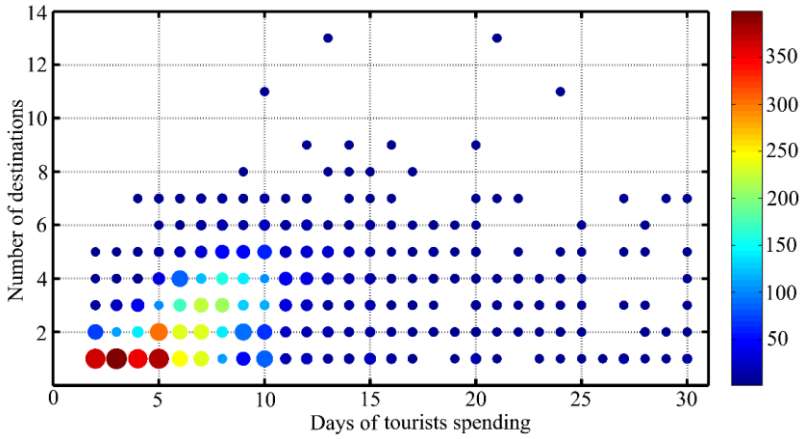

Figure 8. Distribution of tourists according to days spent and the number of destinations (larger dots denote more number of tourists).

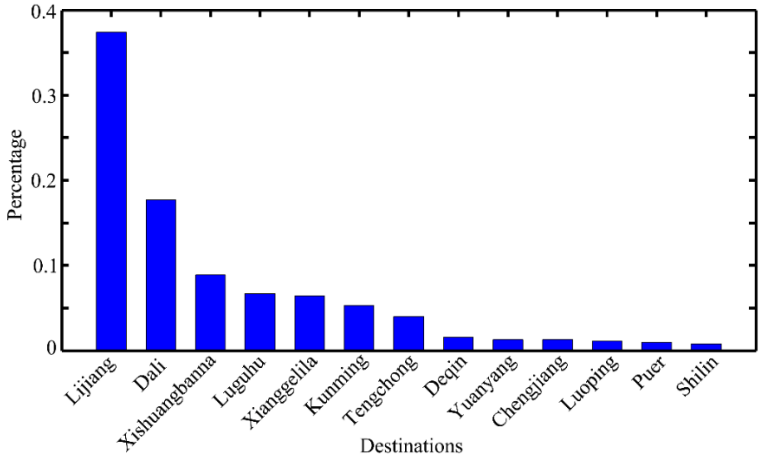

Figure 9. Distribution of tourists staying in one destination in Yunnan.

4.2. Association Rules among Tourist Destinations

In this section, the classical Apriori algorithm is utilized to mine the strong association rules among the popular tourist destinations from set S. The 4110 sequences containing two or more destinations are used to perform association analysis because sequences involving only one destination cannot be used to analyze the relationship among destinations. The minimum confidence $min_{confidence} = 0.6$, and the minimum lift $min_{lift} > 1.0$. Minimum support is difficult to determine because it depends on the characteristics of the database, and users usually set it through trial and error. Referring to the study of Vu et al. (2017), we set the minimum support $min_{support} = 0$. The top 31 association rules are selected to discuss the travel behavior of tourists in Yunnan. We classify these rules into groups according to the number of destinations in LHS.

Table 2 shows the association rules with one destination in LHS. Several relatively strong association rules are identified between Lijiang and other five destinations ($r_{1–5}$). Regarding the support of rules, many sequences containing Dali and Lijiang (support = 50.5%), indicating that more than half of the tourists plan to visit the two destinations during a tour. Regarding confidence, if travelers intend to visit one of the five destinations (Dali, Kunming, Luguhu, Xianggelila, and Deqin), the probability (more than 80% confidence) that Lijiang will be included in their tour is high, especially in Luguhu and Xianggelila with a confidence of 98.2% and 93.3%, respectively. One possible explanation for this high confidence is the convenience of the road connecting Lijiang and Luguhu or Xianggelila and the short distance between them as compared with destinations, such as Dali or Kunming (Figure 5a). In addition, a strong association exists between Dali and Kunming, indicating that some travelers intend to visit the two destinations in their travel.

Table 2. Association rules with one destination in LHS.

Rule ID	Association Rules	Support	Confidence
r_1	Dali ⇒ Lijiang	0.505	0.877
r_2	Kunming ⇒ Lijiang	0.399	0.800
r_3	Luguhu ⇒ Lijiang	0.379	0.982
r_4	Xianggelila ⇒ Lijiang	0.355	0.933
r_5	Deqin ⇒ Lijiang	0.089	0.843
r_6	Kunming ⇒ Dali	0.350	0.701
r_7	Dali ⇒ Kunming	0.350	0.615

Table 3 displays the 13 association rules with two destinations in LHS. Dali, Lijiang, and Kunming are more likely to be included in a tour by travelers (support = 30.2%), that is, if travelers choose to visit any two destinations among the three, they are likely to visit the remaining one ($r_{1–3}$). Rule $r_{4–7}$ shows that travelers will visit Lijiang if they visited Dali and Luguhu/Xianggelila or Luguhu and Kunming/Xianggelila (confidence = 100%). If travelers plan to visit Kunming and Luguhu/Xianggelila, they are more than 70% likely to travel to Dali during their journeys ($r_{8–9}$). Meanwhile, travelers who visit Dali and Luguhu/Xianggelila are likely to visit Kunming ($r_{10–11}$). For rule $r_{12–13}$ covering Lijiang, Xianggelila, and Deqin, it is apparent that Lijiang or Xianggelila is more likely to be considered as a stop when travelers plan to visit the other two destinations.

For the association rules with three destinations in LHS (Table 4), eight strong rules are found from the sequences. Travelers have a high probability of visiting Kunming if they plan to visit Lijiang, Dali, and Luguhu/Xianggelila ($r_{1–2}$). Dali ($r_{3–4}$) is likely to be a stop during the journeys if travelers travel to one of the two destination combinations (Lijiang, Luguhu, and Kunming; Lijiang, Kunming, and Xianggelila). In addition, rules $r_{5–8}$ demonstrate that a very high chance (approximately 100% confidence) that Lijiang will be visited when travelers plan to visit one of the four destination combinations (Dali, Luguhu, and Kunming; Dali, Kunming, and Xianggelila; Luguhu, Kunming, and Xianggelila; Dali, Luguhu, and Xianggelila). Moreover, all rules contain Lijiang, indicating that Lijiang is an indispensable destination when travelers intend to visit four or more destinations.

Table 3. Association rules with two destinations in LHS.

Rule ID	Association Rules	Support	Confidence
r_1	Lijiang, Dali ⇒ Kunming	0.302	0.605
r_2	Dali, Kunming ⇒ Lijiang	0.302	0.863
r_3	Lijiang, Kunming ⇒ Dali	0.302	0.756
r_4	Dali, Luguhu ⇒ Lijiang	0.192	0.980
r_5	Dali, Xianggelila ⇒ Lijiang	0.181	0.962
r_6	Luguhu, Kunming ⇒ Lijiang	0.176	0.992
r_7	Luguhu, Xianggelila ⇒ Lijiang	0.158	0.988
r_8	Luguhu, Kunming ⇒ Dali	0.133	0.733
r_9	Kunming, Xianggelila ⇒ Dali	0.129	0.704
r_{10}	Dali, Luguhu ⇒ Kunming	0.133	0.679
r_{11}	Dali, Xianggelila ⇒ Kunming	0.129	0.685
r_{12}	Xianggelila, Deqin ⇒ Lijiang	0.078	0.836
r_{13}	Lijiang, Deqin ⇒ Xianggelila	0.078	0.879

Table 4. Association rules with three destinations in LHS.

Rule ID	Association Rules	Support	Confidence
r_1	Lijiang, Dali, Luguhu ⇒ Kunming	0.133	0.690
r_2	Lijiang, Dali, Xianggelila ⇒ Kunming	0.126	0.697
r_3	Lijiang, Luguhu, Kunming ⇒ Dali	0.133	0.737
r_4	Lijiang, Kunming, Xianggelila ⇒ Dali	0.126	0.715
r_5	Dali, Luguhu, Kunming ⇒ Lijiang	0.133	0.996
r_6	Dali, Kunming, Xianggelila ⇒ Lijiang	0.126	0.979
r_7	Luguhu, Kunming, Xianggelila ⇒ Lijiang	0.092	1.000
r_8	Dali, Luguhu, Xianggelila ⇒ Lijiang	0.085	0.994

Three association rules have four destinations in LHS and one in RHS (Table 5). Only 6.4% of the sequences contain the top five popular destinations. Lijiang is the destination that travelers are sure to visit. Kunming (confidence = 75.2%) or Dali (confidence = 69.8%) is more likely to be visited if travelers plan to visit the other four destinations ($r_{1–2}$).

Table 5. Association rules with four destinations in LHS.

Rule ID	Association Rules	Support	Confidence
r_1	Lijiang, Dali, Luguhu, Xianggelila ⇒ Kunming	0.064	0.752
r_2	Dali, Luguhu, Kunming, Xianggelila ⇒ Lijiang	0.064	1.000
r_3	Lijiang, Luguhu, Kunming, Xianggelila ⇒ Dali	0.064	0.698

5. Discussion

With the rapid development of information and technology, user-generated data has experienced explosive growth in various fields. Based on these user-generated data, data-driven innovation become available and has led to the emergence and development of some new products and business models in the digital market [49]. In tourism, various user-generated data have been used to understand tourist travel patterns and the service quality of tourism, which help exploit new tourism products and improving management efficiency [25,34]. This study addresses this line and attempts to discover spatial frequent association rules among popular destinations from user-generated textual online reviews.

Based on the analysis of the results, some characteristics can be drawn as follows. First, based on the support of the association rules, it can be concluded that Lijiang, Dali, Luguhu, Xianggelila, and Kunming are the top five popular destinations in Yunnan. These destinations, especially Lijiang (which is an indispensable destination in Yunnan), are considered by most of the travelers who have not been to these places. From the perspective of spatial distribution, four out of these five popular destinations are mainly located in the northwest of Yunnan, forming the overall characteristics "dense in the west

and sparse in the east, dense in the north and sparse in the south", which is consistent with previous research results [50]. In addition, the strong correlation among the five major tourist destinations further indicates a spatial monopoly in important destinations [51]. Second, the distance between destinations and traffic accessibility are key factors that affect travel plans, as illustrated by the rules in Table 2 (i.e., the closer the distance between the destinations, the larger the confidence of the rules). For example, Lijiang, Luguhu, and Xianggelila are three destinations with relatively close distances, and the rules containing these destinations usually have high confidence. Moreover, the traffic conditions from Lijiang to the other two destinations are convenient. Therefore, travelers usually include the three destinations in their journeys. In the context of all-for-one tourism in China, the first task that must be executed to improve the role of other destinations in Yunnan is to break the monopoly of popular destinations. This goal can be achieved by developing advanced transport facilities and networks to establish branch connections between the popular destinations and their nearby destinations.

Specifically, one main contribution of this study towards the existing literature is to propose a methodological workflow to extract spatial association characteristics among popular destinations from these unstructured textual travel reviews. Although this paper takes Yunnan province, China as a case study to demonstrate the feasibility of the workflow, the proposed method could be extended to other areas. Moreover, it is not limited to spatial scale, which means that it is feasible to utilize the method to analyze the association characteristics of attractions within a city, or quantify the connection of cities within a country. Currently, the travel social service websites have become an important part of tourists, from planning their journeys before the tour to updating their comments or experience after the tour. Therefore, it is very convenient to access this unstructured textual UGC to investigate the travel behavior of tourists. Most previous literature utilizes this type of data to extract meaningful knowledge based on the view of tourism marketing, e.g., analyzing destination image, evaluating tourists' satisfaction about hotels or tourism products. This study attempts to discover useful geographical knowledge from the unstructured textual reviews. Although it is effortless to extract spatial movement of tourists using geo-related tracking datasets such as mobile phone data and social media data (Flickr and Twitter), these datasets are inaccessible for most tourism researchers in China. Furthermore, some social media datasets do not include all of the places that tourists visit during their tours. For example, someone visits a destination but does not post a message on the social media application, then the destination would possibly be ignored in the analysis. Generally, the online travel reviews are available from public travel service websites, and record the destinations and attractions as well as tourists' experience in detail. Therefore, this study makes a new attempt to employ unstructured UGC data for understanding the geographical movement patterns of tourists. It demonstrates that although no location information is available in travel reviews, reviews can also be used as a resource to investigate tourist movement patterns, thereby providing a new method to study the spatial movement of tourists and their destination association. Therefore, the proposed method helps enrich data analysis technique in the field of tourism, and infer spatial associative characteristics among popular destination from textual description generated by visitors.

6. Conclusions

The association rules of tourist destinations quantify the possibility of tourists visiting a destination when they have traveled to one or more different destinations. Therefore, a deep knowledge of the travel behavior of tourists and the association of destinations can provide insights into how tourists schedule their journeys during the tour, thereby helping managers or industries to take effective measures for improving their services and meeting the demands of tourists. Currently, UGC big data from tourism-related social websites and apps offer great opportunities to examine the movement patterns of tourists from an unprecedented perspective. Based on previous studies that utilized geo-tagged UGC data to understand frequent movement patterns of destinations, we derive the association of

destinations using unstructured online textual travel reviews. Yunnan province is used for the case study. We collect travel reviews from the website of a public travel service and propose an extraction process for destination sequences from these reviews. In addition, we identify association rules using the Apriori algorithm. Results show that some popular destinations and frequent association rules among destinations in Yunnan can be uncovered using unstructured textual travel reviews.

Multi-destination tours have become a popular travel mode. Thus, examining the association among destinations can grasp the overall characteristics of destinations in a given area. An understanding of the relationship among tourist destinations could generate some potential implications for governmental agencies and tourist industries. Based on these association rules information, tourist administrative staffs could make corresponding traffic strategies such as setting up extra trains or special trains between these destinations with a high association. Tourist industries could develop some new tourist products or tourist routes by integrating customers' time and interests. In addition, these online travel service agents could enrich their recommendation system, such as recommending the next destinations when visitors are sightseeing other destinations according to these association rules. Therefore, it is useful for extracting the association rules among popular tourist destinations, improving tourist experience, and developing a sustainable smart tourism industry.

However, one main limitation of this UGC data is the lack of attribute information, such as income, age, and preferences. These attributes can be utilized to reveal the influencing factors of travel behavior. Therefore, further studies can combine the UGC data with traditional survey and geo-tagged UGC data to understand tourist movement patterns and their potential influencing factors.

Author Contributions: Conceptualization, T.L. and Y.Z.; methodology, T.L.; software, Y.Z. and X.Y.; validation, X.Y., H.Z.; formal analysis, T.L.; investigation, X.Y.; resources, X.Y.; data curation, X.Y.; writing—original draft preparation, T.L.; writing—review and editing, X.Y., H.Z.; visualization, T.L.; supervision, H.Z.; project administration, T.L., X.Y.; funding acquisition, T.L., Y.Z. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (grant number 41801376, 41801373). China Postdoctoral Science Foundation (No. 2020M682293); Open Research Fund of state key laboratory of information engineering in surveying, mapping and remote sensing, Wuhan University (18S03, 20S03); Key Research Projects of Henan Higher Education Institutions (19A420004); Henan Province Philosophy and Social Science Planning Project (Grant No.2019CJJ079); Henan Provincial Department of Education Humanities and Social Sciences Research Project (Grant No. 2021-ZZJH-164); Henan Agricultural University Science and Technology Innovation Fund Project (Grant No. KJCX2020B02).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

References

1. Shi, B.; Zhao, J.; Chen, P.-J. Exploring urban tourism crowding in Shanghai via crowdsourcing geospatial data. *Curr. Issues Tour.* **2017**, *20*, 1186–1209. [CrossRef]
2. Shoval, N.; Isaacson, M. Tracking tourists in the digital age. *Ann. Tour. Res.* **2007**, *34*, 141–159. [CrossRef]
3. Xiang, Z.; Gretzel, U. Role of social media in online travel information search. *Tour. Manag.* **2010**, *31*, 179–188. [CrossRef]
4. Li, J.; Xu, L.; Tang, L.; Wang, S.; Li, L. Big data in tourism research: A literature review. *Tour. Manag.* **2018**, *68*, 301–323. [CrossRef]
5. Zhang, J. Big data and tourism geographies—an emerging paradigm for future study? *Tour. Geogr.* **2018**, *20*, 899–904. [CrossRef]
6. Lu, W.; Stepchenkova, S. User-Generated Content as a Research Mode in Tourism and Hospitality Applications: Topics, Methods, and Software. *J. Hosp. Mark. Manag.* **2015**, *24*, 119–154. [CrossRef]
7. Brandt, T.; Bendler, J.; Neumann, D. Social media analytics and value creation in urban smart tourism ecosystems. *Inf. Manag.* **2017**, *54*, 703–713. [CrossRef]
8. Shoval, N.; Ahas, R. The use of tracking technologies in tourism research: The first decade. *Tour. Geogr.* **2016**, *18*, 587–606. [CrossRef]

9. Salas-Olmedo, M.H.; Moya-Gómez, B.; García-Palomares, J.C.; Gutiérrez, J. Tourists' digital footprint in cities: Comparing Big Data sources. *Tour. Manag.* **2018**, *66*, 13–25. [CrossRef]
10. Zhou, X.; Xu, C.; Kimmons, B. Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. *Comput. Environ. Urban Syst.* **2015**, *54*, 144–153. [CrossRef]
11. Shao, H.; Zhang, Y.; Li, W. Extraction and analysis of city's tourism districts based on social media data. *Comput. Environ. Urban Syst.* **2017**, *65*, 66–78. [CrossRef]
12. Giglio, S.; Bertacchini, F.; Bilotta, E.; Pantano, P. Using social media to identify tourism attractiveness in six Italian cities. *Tour. Manag.* **2019**, *72*, 306–312. [CrossRef]
13. 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. [CrossRef]
14. Chua, A.; Servillo, L.; Marcheggiani, E.; Moere, A.V. Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tour. Manag.* **2016**, *57*, 295–310. [CrossRef]
15. Raun, J.; Ahas, R.; Tiru, M. Measuring tourism destinations using mobile tracking data. *Tour. Manag.* **2016**, *57*, 202–212. [CrossRef]
16. Girardin, F.; Calabrese, F.; Fiore, F.D.; Ratti, C.; Blat, J. Digital Footprinting: Uncovering Tourists with User-Generated Content. *IEEE Pervasive Comput.* **2008**, *7*, 36–43. [CrossRef]
17. Kádár, B. Measuring tourist activities in cities using geotagged photography. *Tour. Geogr.* **2014**, *16*, 88–104. [CrossRef]
18. Cheng, M.; Edwards, D. Social media in tourism: A visual analytic approach. *Curr. Issues Tour.* **2015**, *18*, 1080–1087. [CrossRef]
19. Batista e Silva, F.; Marín Herrera, M.A.; Rosina, K.; Ribeiro Barranco, R.; Freire, S.; Schiavina, M. Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources. *Tour. Manag.* **2018**, *68*, 101–115. [CrossRef]
20. Santos, F.; Almeida, A.; Martins, C.; Gonçalves, R.; Martins, J. Using POI functionality and accessibility levels for delivering personalized tourism recommendations. *Comput. Environ. Urban Syst.* **2019**, *77*, 101173. [CrossRef]
21. Sun, X.; Huang, Z.; Peng, X.; Chen, Y.; Liu, Y. Building a model-based personalised recommendation approach for tourist attractions from geotagged social media data. *Int. J. Digit. Earth* **2019**, *12*, 661–678. [CrossRef]
22. Wan, L.; Hong, Y.; Huang, Z.; Peng, X.; Li, R. A hybrid ensemble learning method for tourist route recommendations based on geo-tagged social networks. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 2225–2246. [CrossRef]
23. Zeng, B.; Gerritsen, R. What do we know about social media in tourism? A review. *Tour. Manag. Perspect.* **2014**, *10*, 27–36. [CrossRef]
24. Stepchenkova, S.; Morrison, A.M. The destination image of Russia: From the online induced perspective. *Tour. Manag.* **2006**, *27*, 943–956. [CrossRef]
25. Költringer, C.; Dickinger, A. Analyzing destination branding and image from online sources: A web content mining approach. *J. Bus. Res.* **2015**, *68*, 1836–1843. [CrossRef]
26. Marine-Roig, E. Measuring Destination Image through Travel Reviews in Search Engines. *Sustainability* **2017**, *9*, 1425. [CrossRef]
27. Toral, S.L.; Martínez-Torres, M.R.; Gonzalez-Rodriguez, M.R. Identification of the Unique Attributes of Tourist Destinations from Online Reviews. *J. Travel Res.* **2017**, *57*, 908–919. [CrossRef]
28. McCreary, A.; Seekamp, E.; Davenport, M.; Smith, J.W. Exploring qualitative applications of social media data for place-based assessments in destination planning. *Curr. Issues Tour.* **2020**, *23*, 82–98. [CrossRef]
29. McKenzie, G.; Adams, B. A data-driven approach to exploring similarities of tourist attractions through online reviews. *J. Locat. Based Serv.* **2018**, *12*, 94–118. [CrossRef]
30. Yang, Y. Understanding tourist attraction cooperation: An application of network analysis to the case of Shanghai, China. *J. Destin. Mark. Manag.* **2018**, *8*, 396–411. [CrossRef]
31. Jin, C.; Cheng, J.; Xu, J. Using User-Generated Content to Explore the Temporal Heterogeneity in Tourist Mobility. *J. Travel Res.* **2018**, *57*, 779–791. [CrossRef]
32. Ye, Q.; Zhang, Z.; Law, R. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Syst. Appl.* **2009**, *36*, 6527–6535. [CrossRef]
33. Valdivia, A.; Luzón, M.V.; Herrera, F. Sentiment Analysis in TripAdvisor. *IEEE Intell. Syst.* **2017**, *32*, 72–77. [CrossRef]
34. Duan, W.; Cao, Q.; Yu, Y.; Levy, S. Mining Online User-Generated Content: Using Sentiment Analysis Technique to Study Hotel Service Quality. In Proceedings of the 2013 46th Hawaii International Conference on System Sciences, Wailea, HI, USA, 7–10 January 2013; pp. 3119–3128.
35. Sparks, B.A.; Perkins, H.E.; Buckley, R. Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behavior. *Tour. Manag.* **2013**, *39*, 1–9. [CrossRef]
36. Fuchs, M.; Höpken, W.; Lexhagen, M. Big data analytics for knowledge generation in tourism destinations—A case from Sweden. *J. Destin. Mark. Manag.* **2014**, *3*, 198–209. [CrossRef]
37. Park, E.; Kang, J.; Choi, D.; Han, J. Understanding customers' hotel revisiting behaviour: A sentiment analysis of online feedback reviews. *Curr. Issues Tour.* **2020**, *23*, 605–611. [CrossRef]
38. Hu, N.; Zhang, T.; Gao, B.; Bose, I. What do hotel customers complain about? Text analysis using structural topic model. *Tour. Manag.* **2019**, *72*, 417–426. [CrossRef]
39. Wu, L.; Zhang, J.; Fujiwara, A. A tourist's multi-destination choice model with future dependency. *Asia Pac. J. Tour. Res.* **2012**, *17*, 121–132. [CrossRef]

40. Yang, Y.; Fik, T.; Zhang, J. Modeling sequential tourist flows: Where is the next destination? *Ann. Tour. Res.* **2013**, *43*, 297–320. [CrossRef]
41. Rong, J.; Vu, H.Q.; Law, R.; Li, G. A behavioral analysis of web sharers and browsers in Hong Kong using targeted association rule mining. *Tour. Manag.* **2012**, *33*, 731–740. [CrossRef]
42. Vu, H.Q.; Li, G.; Law, R.; Zhang, Y. Travel Diaries Analysis by Sequential Rule Mining. *J. Travel Res.* **2017**, *57*, 399–413. [CrossRef]
43. Saura, J.R. Using data sciences in digital marketing: Framework, methods and performance metrics. *J. Innov. Knowl.* **2021**, *6*, 92–102. [CrossRef]
44. Li, G.; Law, R.; Rong, J.; Vu, H. Incorporating Both Positive and Negative Association Rules into the Analysis of Outbound Tourism in Hong Kong. *J. Travel Tour. Mark.* **2010**, *27*, 812–828. [CrossRef]
45. Lee, L.; Cai, G.; Lee, K. Mining points of interest association rules from Geo-tagged photos. In Proceedings of the 2013 46th Hawaii International Conference on System Sciences, Wailea, HI, USA, 7–10 January 2013; pp. 1580–1588.
46. Versichele, M.; Groote, L.; Bouuaert, M.C.; Neutens, I.M.; Weghe, N.V. Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium. *Tour. Manag.* **2014**, *44*, 67–81. [CrossRef]
47. Qi, S.; Wong, C.U.I. An Application of Apriori Algorithm Association Rules Mining to Profiling the Heritage Visitors of Macau. In *Information and Communication Technologies in Tourism 2015*; Tussyadiah, I., Inversini, A., Eds.; Springer: Cham, Switzerland, 2015; pp. 139–151. [CrossRef]
48. Gandhi, M. An enhanced approach for tourism recommendation system using hybrid filtering and association rule mining. *Natl. J. Syst. Inf. Technol.* **2015**, *8*, 1–8.
49. Saura, J.R.; Ribeiro-Soriano, D.; Palacios-Marqués, D. From user-generated data to data-driven innovation: A research agenda to understand user privacy in digital markets. *Int. J. Inf. Manag.* **2021**, 102331. [CrossRef]
50. Sun, Y.; Shi, C.Y.; Tang, W.W.; Liu, J. Research on the spatial network characteristics of travel itinerary in Yunnan province. *Hum. Geogr.* **2016**, *1*, 147–153. (In Chinese)
51. Tang, L.; Wei, J.; Zhao, M. Features of Regional Travel Itineraries Complex Networks: Taking Fujian Province as an Example. *Tour. Trib.* **2014**, *29*, 57–66. (In Chinese)