A Novel Evaluation Method of Tourist Attractions Based on Extracted Emotional Information and Semantic Analysis

Shiyu WANG, Liwei CHEN, Shenglang JIN

College of Tourism, Huangshan University, Huangshan, 245021 China.

***Abstract* — To improve the accuracy of evaluating the quality of tourist spots, we propose a method based on semantically analyzing extracted emotional information. The research steps are: i) we select scenic spots and determine the data to be used in the method, ii) we build an index system and classification method, iii) we evaluate the class probability of scenic spots by calculating feature items and class joint probability of the scenic spot, iv) we use a Naive Bayes classifier to forecast the class of scenic spots, v) we establish a latent semantic model to evaluate the scenic spots, vi) we use expert interviews and phone surveys to carry out field investigation on 52 scenic spots, and vii) collect and tabulate the experimental results. The results show that the evaluation of tourist spots is consistent with objective reality, and verifies the practicality and efficiency of our method.**

Keywords - emotional information extraction; semantic analysis; tourist spots; naive bayes

I. INTRODUCTION

Traditional method for evaluation of tourist spots is to confirm the classification system of tourist resources, and then establish an evaluation model on this basis, so that the scientific classification may be given to tourist spots based on expert scoring on tourist spots under objective survey. At present, the quality class of Chinese tourist spots is divided into 5 classes, including AAAAA-class, AAAA-class, AAA-class, AA-class and A-class tourist spots. Emergence of GIS technology makes researchers combine GIS technology with the model of tourist resources evaluation so as to realize functions of evaluation, forecasting, planning and decision-making, etc. of tourist resources.

In recent years, with emergence of location services and social network sites, the user may conveniently record the existing location via mobile computing device with location function; therefore, a large number of location check-in data have been generated in the Internet. Many scholars have obtained obvious landmarks of the city by means of analysis, extraction and excavation of location check-in data, and are to excavate and recommend tourist spots and route through Mahalanobis distance discriminant method, weighted Mahalanobis distance method or Bayes classification algorithm. These evaluation methods are different from methods for evaluation of tourist spots, and lay more emphasis on selection and preference of the public as well as personalized recommendation under demands of users themselves.

Here, with intention of consideration of two perspective and combination between two indexes of "quality class of tourist spots" and "check-in number of tourist spots", establish an index system under consideration of perspectives of both "expert" and "the public", and simultaneously excavate tourist spots favored by men or women and those favored by natives or foreigners, and whether the tourist spots have low and peak seasons and whether the distribution between low and peak seasons is similar, as well as other funny problems from detailed information of gender, region and check-in time, etc. of users, so that the evaluation on tourist spots is more comprehensive and user-friendly to widely meet demands of different users.

II. DESCRIPTION OF EXPERIMENTAL SUBJECT

A. Sampling Information of Tourist Spots Within the Scope

Among 5138 national tourist spots (in total) of A-5A and non-A, make simple random sampling without replacement based on different classes; the total number of simple random sampling is 6, while the single sample capacity is 25. Then, with samples as initial seed file, get Microblog number, check-in number, picture number and other information from Sina Microblog, and the total number of records is 150. Each record includes five data items of name of scenic spots, quality class, Microblog number, check-in number and picture number.

B. Check-in Information of Tourist Spots Location

Take 179 tourist spots recommended in Baidu Travel – Zhengzhou Tourist Spots and tourist spots recorded in Zhengzhou Tourist Administration as the initial seed file, and then carry out crawling on the Microblog website of Sina, to only acquire the check-in, picture and Microblog number for the tourist spots with check-in number less than 100, not acquiring the detailed check-in user information. Finally, 58 tourist spots in and around the Zhengzhou City are selected, with total check-in number of 70,859, of which 70,685 cases contain the detailed information (the final acquiring time is August 15, 2014), including 4 3A scenic spots and 2 2A scenic spots.

Acquired check-in user data includes 8 items of user ID, user name, gender, area, number of attentions, number of fans, Microblog number and check-in time.

C. Construction and Classification of Index System

Here, chose the "quality class of tourist spot" represents the data index in expert's perspective, as well as "check-in number" as the data index in "public" perspective. Because the "quality class of tourist spot" is the ordinal level variable, while "check-in number" is the interval level variable, it is required to convert the "check-in number" from the interval level variable into "few", "adaptable" and "many" ordinal level variable. Then, respectively take two indexes as axis X and Y, and construct the rectangular coordinate system, as shown in Fig. 1.

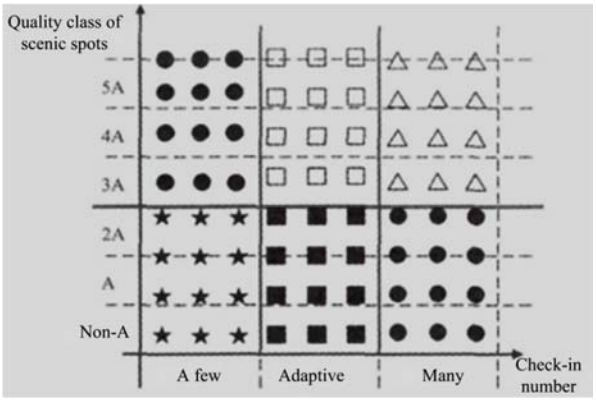

Figure 1. Classification of tourist spots based on "quality class of scenic spots" and "check-in number"

All scenic spots can be divided into the following 5 classes here. 1) Scenic spots at high quality class and with relative more check-in number can be classified into highly-recommended scenic spots undoubtedly; 2) Scenic spots at high quality class and with check-in number fit with level can be classified into traditional classical scenic spots; 3) Scenic spots at low quality class and with check-in number fit with level can be classified into traditional common scenic spots; 4) While scenic spots at high quality class but with relatively few check-in number, or scenic spots at low quality class but with relatively more check-in number, such kinds are provided with mutual contradictions in two indexes, which may be "a lie under great reputation", or maybe "unknown for its hiding in deep mountains", such scenic spots can be classified into scenic spots worthy of exploring; 5) Scenic spots at low quality class and with relatively few check-in number can be classified into scenic spots that will be chosen with cautions.

III. LATENT SEMANTIC SCENIC SPOT EVALUATION BASED ON NAIVE BAYES CLASSIFIER

A. Emotion Polarity Classification based on Naive Bayes

The basic idea of naïve Bayes classifier is to evaluate the class probability of given scenic spots by calculating the feature item of scenic spot and class joint probability, to predict the class of scenic spots. According to the Naive Bayes idea, it is finally required to obtain the class C_i, as well as the probability $P\left(C_i\left|Doc\right.\right)$ when the scenic spot is given, and so that to judge the class of scenic spot in accordance with the size of conditional probability $P\left(C_i\left|Doc\right.\right)$. According to the conditional probability formula, it can be obtained:

$$P\left(C_i\left|Doc\right.\right)=\frac{P\left(Doc\left|C_i\right.\right)\times P\left(C_i\right)}{P\left(Doc\right)} \tag{1}$$

The Naive Bayes classifier is achieved on the basis of one assumption: The labels in the scenic spot are mutually independent, namely, label feature of scenic spot relies on the class, while having no relation with other words at label context of its scenic spots or the length of scenic spot. According to this assumption, the probability of scenic spot under class C_i can be expressed as:

$$P\left(Doc\left|C_i\right.\right)=\prod_{t_j\in V}P\left(Doc\left(t_j\right)\left|C_i\right.\right) \tag{2}$$

$$P\left(Doc\right)=\sum_i\left[P\left(C_i\right)\prod_{t_j\in V}P\left(Doc\left(t_j\right)\left|C_i\right.\right)\right] \tag{3}$$

Of which, t_j represents the feature vector of scenic spot, and $P\left(C_i\right)$ represents the probability of scenic spots in class C_i occupies the total scenic spots, while $P\left(Doc\left(t_j\right)\left|C_i\right.\right)$ represents the probability of feature t_j in the scenic spot appear under the given class C_i.

$$\begin{aligned}P\left(C_i\left|Doc\right.\right)&=\frac{P\left(Doc\left|C_i\right.\right)\times P\left(C_i\right)}{P\left(Doc\right)}\\&=\frac{P\left(C_i\right)\times\prod_{t_j\in V}P\left(Doc\left(t_j\right)\left|C_i\right.\right)}{\sum_i\left[P\left(C_i\right)\prod_{t_j\in V}P\left(Doc\left(t_j\right)\left|C_i\right.\right)\right]}\end{aligned} \tag{4}$$

Finally, the value of conditional probability $P\left(Doc\left(t_j\right)\left|C_i\right.\right)$ can be estimated by utilizing Lapras:

$$P\left(Doc\left(t_j\right)\left|C_i\right.\right)=\frac{1+N\left(Doc\left(t_j\right)\left|C_i\right.\right)}{2+\left|D_{c_i}\right|} \tag{5}$$

Of which, $N\left(Doc\left(t_j\right)|C_i\right)$ refers to the number of scenic spots where the feature t_j appears in the scenic spots in class C_i, and $\left|D_{c_i}\right|$ refers to the total number of scenic spots in class C_i. The concrete emotion polarity classification flow is as shown in Fig. 2.

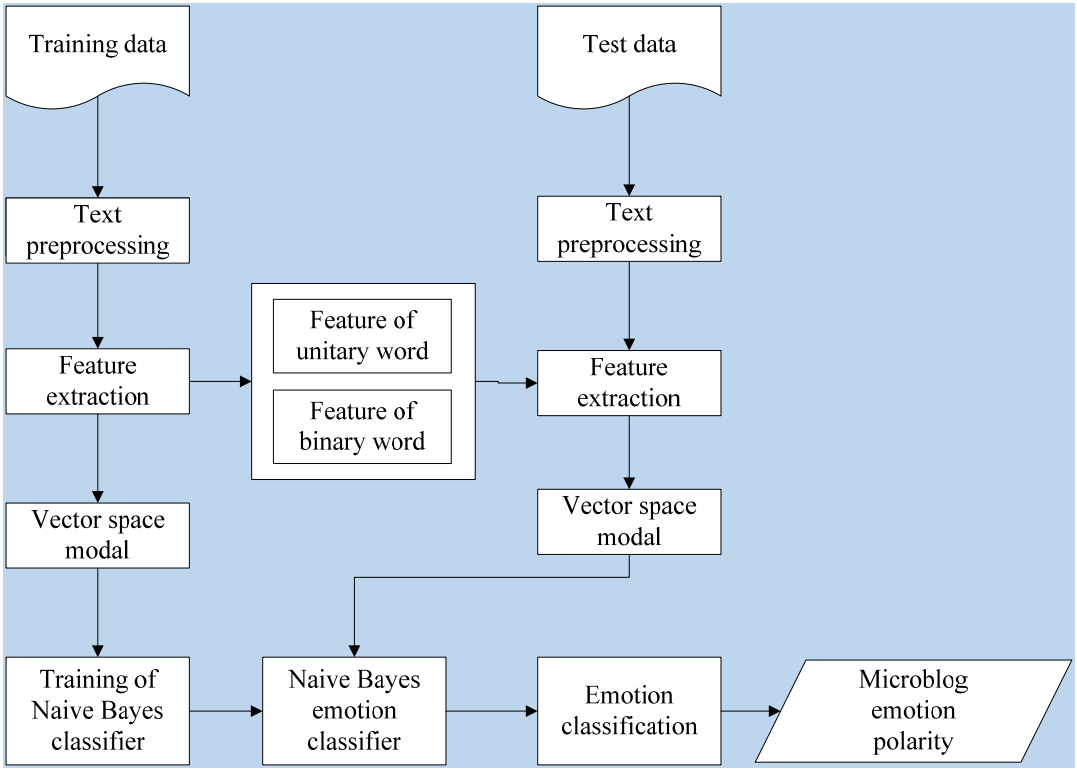

Figure 2. Emotion polarity classification flow based on naive Bayes.

B. Latent Semantic Model of Scenic Spot Evaluation

(1) Screen the scenic spots with high attention. Count the forward number *N_rel* and comment number *N_com* of every scenic spot. When choosing the Microblog with high attention at one time period, an index is required to be used for choosing the topic with certain attention. After observing and analyzing, it can be found that, after one topic attracts the net citizen to focus on, its forward and comment number will gradually increase and so that spread within a very short time, which reflects the hot degree of the topic to some extent. Therefore, there is:

Definition 1: The hot degree H_T of (hot) Microblog *b* is its weighted sum of forward number *N_rel* and comment number *N_com*, namely

$$H_T = \left(\alpha N\_rel + \beta N\_com\right) b_i \tag{6}$$

Of which, the calculation form of N_relb_i or N_combi_i is:

$$N\_rel\left|N\_comb_i = \begin{cases} 0, N < 10^2 \\ 0.3, 10^2 < N < 10^3 \\ 0.5, N \geq 10^3 \end{cases}\right. \tag{7}$$

Of which, α and β are regulatory factors withα+β=1, and N is the Microblog numbers of scenic spots, while $H_T \in [0,1]$ is the hot degree of every topic. In addition, only the Microblogs with $H_T \geq 0.3$ are made the subsequent processing, namely, the Microblogs with *N_relbi* and *N_combi* of 0 or in very small number do not have the condition of hot topic. The definition of hot degree supplies a macro standard for preliminary judge whether a topic can be a hot topic, which is be irrelevant to the Microblog content.

The following latent semantic analysis analyze the information implied behind the characters from the micro view, and find the hot topic through the two-phase scenic-spot clustering algorithm.

(2) Construct the latent semantic space. Carry out the Chinese words segmentation and stop-word removal for the Microblogs screened at the first step with ICTCLAS. Build the index for scenic spots after word segmentation, and obtain the lexical item of 71514×18456 dimension – scenic-spot matrix $\boldsymbol{A}$ (formula (1)), of which *aij* represents the weight of *i*th label in *j*th scenic spot. Because the Microblog scenic spot is short and in large number, and the label entry appearing in the single scenic spot is very limited, normally, $\boldsymbol{A}$ is the sparse matrix. In the $\boldsymbol{A}$, as the feature entry weight has multiple different calculation methods, the calculation of in this paper applies the most common TF-IDF method with optimal effect at present, with calculation method as follows:

$$a_{ij} = \frac{tf_{ij} \times lb\left(N/n_i + 0.01\right)}{\sqrt{\sum_{j=1}^{t} f_{ij} \times lb\left(N/n_i + 0.01\right)^2}} \tag{8}$$

Of which, tf_{ij} represents the frequency of *i*th label in *j*th scenic spot, and *N* is the total number of scenic spots, while *ni* is the number of scenic spot containing the entry *i*. Calculate *aij* with formula (6), carry out the SVD processing with SVDLIBC developed by Massachusetts Institute of Technology for matrix $\boldsymbol{A}$, and obtain the approximate matrix $\boldsymbol{A}k$ of matrix $\boldsymbol{A}$, of which front line $\boldsymbol{U}k$ and $\boldsymbol{V}k$ of $\boldsymbol{U}$ and $\boldsymbol{V}$ are respectively served as word vector and scenic spot vector. This is the latent semantic analysis. And then carry out the further processing on this basis.

(3) The flow of two-phase clustering algorithm is as follows:

① $S = RandomS(D)$ //Random draw the scenic spot sample from the database *D*.

②Divide the sample *S* into *n* isometric parts, and carry out the local cluster in minimum mean distance $dist\left(c_i, c_j\right) = \frac{1}{n_i n_j} \sum_{p \in c_i} \sum_{p \in c_j} \left|p - p'\right|$ for any two data item *p* and *p′* in the cluster *ci* and *cj*.

③Remove the isolated scenic spots by random sampling.

④The new cluster represents the point $w.rep = p + \alpha \times \left(w.mean - p\right)$ (constriction factor: α = 0.5).

⑤Respectively calculat the value $Q(c)$ collected at last cluster in every layer, and draw the layer enabling the target function $Q(c)$ value to be the minimum among them.

⑥*LabRecord*(*D*) //Label every data sample with corresponding cluster label.

⑦Evaluate the mean value of all samples in every class *C*, and obtain the corresponding cluster center.

⑧According to the value K and cluster center obtained from ④ and ⑤, execute the *K*-means algorithm and obtain the cluster result of hot topic. End.

In the above ③, the removal of isolated point is divided as two phases: (1) In the clustering process, if a cluster increases too slowly (with point number as the threshold value), then, remove it; (2) When the clustering is nearly ended, remove the extremely small cluster.

IV. EXPERIMENT ANALYSIS

According to the above classification and information mining methods, taking the check-in number of 52 scenic spots in Zhengzhou and its surrounding areas as the experimental samples, scenic spots classification, gender factors analysis, regional factors analysis and time factors analysis are conducted to them respectively. Firstly, according to the index system under consideration of perspectives of both “expert” and “the public”, the 52 scenic spots are classified. Where, the scenic spots of Shaolin Temple, Yellow River scenic spot, Zhengzhou Century Joy Park and other scenic spots are classified as the traditional and classical type; ZZICEC, Fantawild Advaenture, ERQI, Zhongyuan Tower and other scenic spots are classified as the type worth exploring; Jiangnanchun Hot Spring, Zhengzhou Green Expo Garden, Zhenzhou OPARARA and other scenic spots are classified as the traditional and ordinary type； the Yellow River Garden Entrance, Zhengzhou Marine Museum and other scenic spots are classified as the key recommendation type. In the gender factors analysis, most of the evaluation results are in accordance with the general cognition; namely, the male prefers the “manly” scenic spots, such as the Shaolin Temple, Zhongyuan Tower and others, while the female prefers the “gentle” scenic spots, such as the Jiangnanchun Hot Spring, Sinian Garden and others. However, a kind of interesting phenomenon is that the female prefers the scenic spots such as playgrounds, zoos and marine attractions, which may be due to the fact that the female takes more responsibility of bringing up children, thus leading to a preference for children scenic spots by more women; however, the specific reasons need to be further studied. In the regional factors analysis, outsiders prefer those local representative scenic spots with relatively higher reputation. In the time factor analysis, the vast majority of scenic spots have the slack and peak season, which is not suitable for few scenic spots such as Henan Art Center, Zhengzhou Museum and etc. According to the similarity of the change curve for the check-in number in each month of the scenic spots, the level clustering is conducted, finally getting the five major categories of tourism slack and peak seasons.

A: indicate that the peak tourist season is concentrated in winter or the Spring Festival, such as the scenic spots of Jiangnanchun Hot Spring, Songshan Shaolin Ski Resort, or Town God's Temple in Zhengzhou, as shown in Fig. 3.

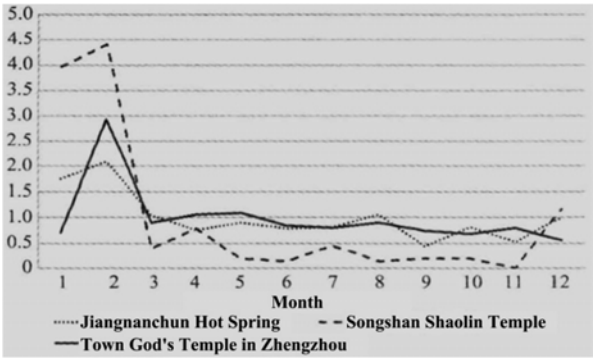

Figure 3. Peak tourist season concentrated in the winter or the spring festival

B: indicate that the peak tourist season is concentrated in the summer, such as the OPARARA, the Mala Bay of Yellow River Valley, Qingping Mountain Park and other cooling and refreshing scenic spots, as shown in Fig. 4.

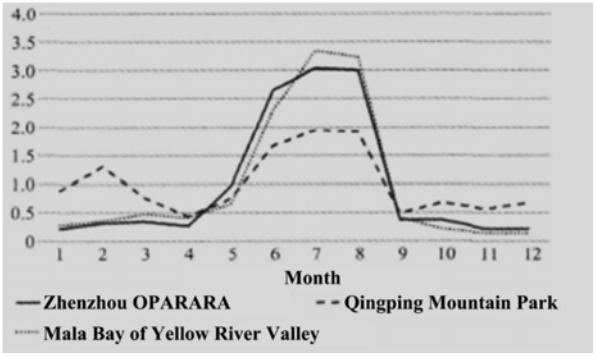

Figure 4. Peak tourist season concentrated in the summer

C: Indicate that the peak tourist season is concentrated in the specific one or two peaks; for example, peak season of the scenic spot of Yanming Lake is mainly concentrated in October, which is the time for the hot selling of hairy crab, while the Yanming Lake is famous for hairy crab; as shown in Fig. 5.

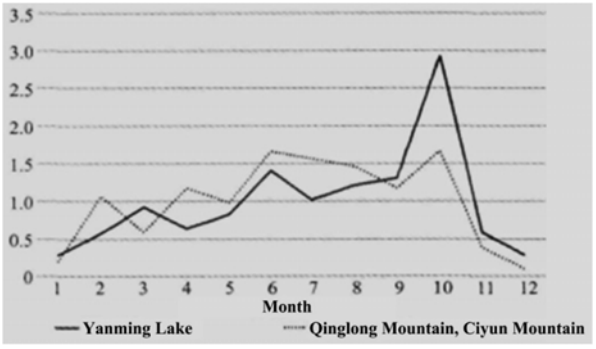

Figure 5. 1~2 Peaks of tourist peak season

D: Indicate that several small peaks will be presented at peak tourist season without strong overall regularity, but mainly focused on April to October that is applicable for outdoor activity, as shown in Fig. 6.

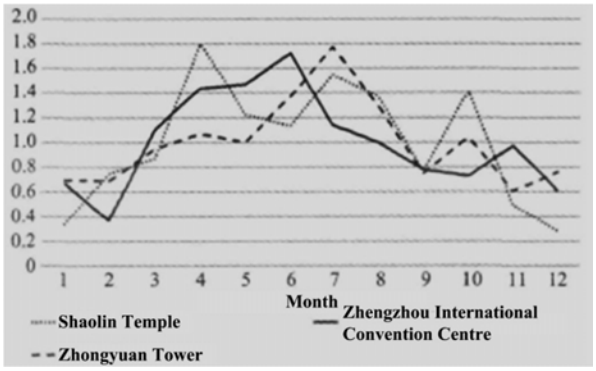

Figure 6. Several peaks figure in peak tourist season

52 scenic spots in experimental result are verified one by one through method of expert interviews, phone research and field visit. Except for the fact that it is difficult to verify partial details (for example, staff at scenic spot of Shaolin Temple tells that peak tourist season of Shaolin Temple is April to November of every year through phone research and interview, but it is difficult to answer difference of every month), the majority of evaluation results conform to objective and actual condition of the scenic spots, which verifies practicability and effectiveness of method mentioned here further.

V. CONCLUSION

This paper puts forward a method for evaluation of tourist spots based on emotional information extraction and semantic analysis, evaluates class probability of given scenic spots by utilizing Naive Bayes classifier and calculating characteristic item of scenic spots and joint probability of class and predicts class of scenic spots. Through latent semantic analysis method, high dimension, synonymy and ambiguity problem in traditional vector space modal is solved. Result shows that evaluation to tourist spots conforms to objective and actual condition basically, which verifies practicability and effectiveness of the method further.

ACKNOWLEDGMENTS

Projects of Philosophy and Social Sciences Research of Anhui Province. (No. AHSKQ2014D59)Coupling Research on Agricultural Industrialization and Ecotourism Development in Huizhou; Project of Philosophy and Social Sciences Planning of Anhui (NO. AHSKQ2014D27)Study on the Transformation of Regional Tourism in South Anhui from the Perspective of Ecological Civilization.

REFERENCES

[1] Prebensen N K, Woo E J, Chen J S. Motivation and involvement as antecedents of the perceived value of the destination experience.[J]. Journal of Travel Research, (2013), 52(2):253-264.

[2] Juster R P, Smith N G, Ouellet É. Sexual orientation and disclosure in relation to psychiatric symptoms, diurnal cortisol, and allostatic load.[J]. Psychosomatic Medicine, (2013), 75(2):103-16.

[3] Quijano-Sanchez L, Recio-Garcia J A, Diaz-Agudo B. Social factors in group recommender systems[J]. Acm Transactions on Intelligent Systems & Technology, (2013), 4(1):1199-1221.

[4] Wong I [I K A, Yimking [Wan Y K P W. A Systematic Approach to Scale Development in Tourist Shopping Satisfaction Linking Destination Attributes and Shopping Experience[J]. Journal of Travel Research, (2013), 52(1):29-41.

[5] Nawijn J, Mitas O, Lin Y Q. How do we feel on vacation? A closer look at how emotions change over the course of a trip.[J]. Journal of Travel Research, (2013), 52(2):265-274.

[6] Brown A, Kappes J, Marks J. Mitigating theme park crowding with incentives and information on mobile devices.[J]. Journal of Travel Research, (2013), 52(4):426-436.

[7] Cheng T M, Wu H C, Huang L M. The influence of place attachment on the relationship between destination attractiveness and environmentally responsible behavior for island tourism in Penghu, Taiwan[J]. Journal of Sustainable Tourism, (2013), 21(8):1166-1187.

[8] Cheng Q, Su B, Tan J. Developing an evaluation index system for low-carbon tourist attractions in China– A case study examining the Xixi wetland[J]. Tourism Management, (2013), 36(3):314-320.

[9] Li X, Kaplanidou K. The impact of the 2008 Beijing Olympic Games on China's destination brand: a U.S.-based examination.[J]. Journal of Hospitality & Tourism Research, (2013), 37(2):237-261.

[10] Versichele M, Groote L D, Bouuaert M C. Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium[J]. Tourism Management, (2014), 44(13):67-81.

[11] Wu J C, Li W. The ripple effect from destroyed attractions to undestroyed attractionsafter natural disasters: example of attractions in Sichuan province after the Wenchuan Great Earthquake.[J]. Tourism Tribune, (2013), 117(30):6489-6507.

.