Postprint of a Sentiment Analysis Model for Temporal Features of Online Travelogues

Authors: Cheng Cuiqiong, Xu Jian

Date: 2017-11-08T00:00:00+00:00

Abstract

[Objective] Through sentiment analysis of online travel blogs, this study aims to discover the temporal distribution patterns of tourists' emotional tendencies toward destinations.

[Application Background] An increasing number of individuals collect information and formulate travel plans by browsing extensive online travel blogs. Such blogs have become a crucial reference for tourists when selecting destinations and travel timing, while also presenting business opportunities for service providers.

[Method] This paper proposes a sentiment analysis model oriented toward the temporal characteristics of online travel blogs to analyze temporal variation patterns of tourist emotions. The model comprises five modules: (1) collection of online travel blog text content and travel time data, (2) preprocessing of blog text, (3) sentiment annotation, (4) statistical analysis of blog sentiment feature scores by time period, and (5) analysis of temporal characteristics of blog sentiment. Experiments were conducted on the model using crawled travel blogs from four types of destinations.

[Results] Among the seven emotion categories, the mean score of [Positive] emotion consistently and substantially exceeds other emotions across all months for each destination, demonstrating relative stability; [Positive], [Joy], and [Negative] exhibit considerable fluctuations across different months; the temporal fluctuation of sentiment is not correlated with the corresponding volume of blogs, indicating that the traditional classification of peak and off-peak seasons is not associated with tourists' actual emotional experiences.

[Conclusion] The model can effectively reflect temporal fluctuations of tourist emotions at destinations, thereby providing a novel information reference channel for tourism managers and potential tourists in information acquisition.

Full Text

Preamble

ChinaXiv Partner Journal, Issue 2, 2017, No. 2
A Sentiment Analysis Model Based on Temporal Characteristics of Travel Blogs*
Cheng Cuiqiong, Xu Jian
(School of Information Management, Sun Yat-Sen University, Guangzhou 510006)

Abstract

[**Objective**] This study aims to identify temporal distribution patterns of tourists' emotional tendencies toward destinations through sentiment analysis of online travel blogs.
[**Context**] An increasing number of travelers rely on extensive travel blogs to gather information and plan their trips. Travel blogs have become an important reference for tourists when selecting destinations and travel times, while also providing business opportunities for service providers.
[**Methods**] We propose a sentiment analysis model based on the temporal characteristics of travel blogs to analyze temporal variation patterns in tourist emotions. The model comprises five modules: collection of travel blog text content and temporal data, text preprocessing, sentiment annotation, statistical calculation of sentiment feature scores by time period, and temporal characteristic analysis of blog sentiments. We conducted experiments using blogs from four types of destinations.
[**Results**] Among seven emotion categories, the mean score for [Good] consistently far exceeded other emotions across all destinations and months, remaining relatively stable. The emotions [Good], [Joy], and [Disgust] showed the greatest fluctuation across months. Emotional fluctuations over time were not correlated with the volume of corresponding travel blogs, indicating that traditional peak/off-season classifications are not related to tourists' actual emotional experiences.
[**Conclusions**] The model effectively reflects fluctuations in tourist emotions toward destinations over time, thereby providing a new information channel for tourism managers and potential travelers.

Keywords: Travel Blogs; Sentiment Analysis; Sentiment Lexicon; Temporal Characteristics
Classification Number: G350

1. Introduction

Seasonal fluctuations in tourist attractions and imbalances in visitor flow or service quality between peak and off-peak seasons affect tourist emotions, resulting in different travel experiences across time periods and corresponding changes in emotional tendencies toward destinations. With the development of social media, an increasing number of tourists share their experiences through online travel blogs, expressing opinions and emotional tendencies toward destinations. Travel blog data provides both textual content and travel timing information, accurately reflecting tourists' opinions and emotional tendencies toward destinations across different time periods. Conducting temporal characteristic analysis of travel blogs, combined with fine-grained sentiment classification, can reveal patterns of various emotions over time. This approach offers tourism managers a new information source to identify time periods with strong positive emotions but low visitor numbers, enabling them to adjust promotional strategies. To encourage positive online word-of-mouth, managers can identify periods with strong negative emotions and optimize resource allocation accordingly. For potential tourists, this provides a new search channel based on temporal sentiment tendencies, satisfying diverse information needs and helping them select optimal destinations and timing according to their expectations.

This study employs a dictionary-based sentiment analysis method to annotate sentiments in tourist blogs, analyzing emotional fluctuations across different time periods to discover temporal patterns in emotional tendencies, thereby providing new information channels for tourism managers and potential travelers.

Sentiment analysis, also known as opinion mining, analyzes and extracts subjective viewpoints, emotions, and polarity from user-generated content to determine text sentiment categories [1]. Methods 主要分为两大类: dictionary-based approaches and machine learning approaches.

(1) Dictionary-based sentiment analysis requires constructing a sentiment dictionary that accurately identifies emotion categories and intensity values for sentiment words, where dictionary quality significantly impacts analysis effectiveness. Hu et al. [2] derived sentiment words and their polarities from WordNet, determining sentence sentiment orientation based on divergence between positive and negative scores. In Chinese research, Zhu et al. [3] used HowNet to calculate semantic similarity and related fields, proposing two methods for lexical sentiment orientation computation with high accuracy. Shi et al. [4] constructed a fuzzy sentiment word ontology for microblog sentiment computation, generating public sentiment curves for emergency events.

(2) Machine learning-based sentiment analysis can be formulated as a text classification problem, with common models including Naive Bayes, Maximum Entropy, centroid vector classification, K-nearest neighbors, and Support Vector Machines (SVM). Pang et al. [5] applied Naive Bayes, Maximum Entropy, and SVM to movie review sentiment classification, finding SVM achieved the best performance with up to 80% accuracy. In Chinese research, Xu et al. [6] used Naive Bayes and Maximum Entropy for sentiment classification of news and comments, finding Maximum Entropy generally outperformed Naive Bayes.

Sentiment analysis applications in tourism primarily focus on service reviews and attraction reviews. In service reviews, Banić et al. [7] used KNIME to analyze hotel online reviews, obtaining sentiment scores for specific features and overall hotel sentiment, providing decision support for potential customers and improvement directions for managers. In destination reviews, Wang et al. [8] designed a feature-opinion pair extraction algorithm based on dependency relations for Shanxi attraction reviews, effectively mining tourist opinion information. Zheng [9] applied pointwise semantic analysis to extract sentiment phrases from destination reviews, analyzing tourist-reported problems and proposing suggestions.

Tourist emotions toward destinations vary not only by individual but also by time. Jin [10] discovered dynamic changes in emotional intensity across different travel stages through semi-structured diaries. Hu et al. [11] used ROST CM6 to analyze cycling travel blogs, identifying peak experience segments based on sentiment fluctuation patterns. Yu [12] constructed a tourist sentiment analysis model based on microblog big data, analyzing temporal variation characteristics of tourist emotions in Xi' an. Li et al. [13] designed VisTravel, a visual analytics system for tourism network opinions that effectively displays temporal changes in tourist sentiment through sentiment analysis of travel blogs and reviews.

Significant research and application opportunities remain for sentiment analysis in tourism. Most current research employs binary sentiment classification (positive/negative) and focuses primarily on service reviews. In contrast, this study introduces temporal characteristics into destination sentiment research using fine-grained sentiment classification to construct a sentiment analysis model based on travel blog temporal features, better reflecting temporal variation patterns in tourist emotions toward destinations.

3. Sentiment Analysis Model Based on Temporal Characteristics of Travel Blogs

Existing tourism sentiment analysis primarily focuses on user reviews of tourism products or services. This study argues that travel blogs can be mined more deeply by introducing temporal characteristics to discover temporal patterns in tourist emotions, thereby providing decision-making references for potential tourists and suggestions for tourism product planning and marketing strategies across different time periods. The proposed sentiment analysis model based on temporal characteristics of travel blogs analyzes temporal variation patterns in tourist emotions toward destinations based on blog text content. The model architecture is shown in Figure 1 [Figure 1: see original paper].

The processing workflow comprises five modules: travel blog data collection, text preprocessing, sentiment annotation, sentiment feature score calculation by time period, and temporal characteristic analysis.

(1) Data Collection. First, specify the time range and destinations, then use web crawlers to collect blog data meeting these criteria: complete data elements (destination, title, travel date, text content) and travel dates within the specified period.

(2) Text Preprocessing. Some blogs describe multiple destinations, which may confound sentiment toward the target destination. To improve sentiment targeting and reduce interference from co-occurring destinations, blogs mentioning multiple destinations are removed. Information extraction then captures titles, travel dates, and text content as information units stored in a database for experimental use.

(3) Sentiment Annotation. Construct an appropriate sentiment dictionary and match sentiment words against blog texts, recording positions where sentiment words appear. Since negation affects sentiment categories, identify each sentiment word' s category and intensity from the dictionary, then check for negation words within the preceding five characters. If present, apply sentiment shift accordingly, and record the identified sentiment category and intensity.

(4) Statistical Calculation of Sentiment Feature Scores. Aggregate all identified sentiment types and intensities in each blog text, calculating the sum of intensities for each sentiment category to obtain average scores per blog. Then select an appropriate time unit (quarter, month, week, etc.), sum the average scores of each sentiment category across all blogs in each time period, and calculate overall average sentiment scores for each time period.

(5) Temporal Characteristic Analysis. Analyze sentiment data from different destinations to identify emotional fluctuations across time periods, discover temporal patterns in tourist sentiment tendencies, and provide new information channels for tourism managers and potential travelers.

4. Experiments on Temporal Sentiment Analysis of Travel Blogs

To validate the model, we collected travel blog data from Mafengwo.com [12] and analyzed typical destinations of different types to discover temporal sentiment patterns.

4.1 Data Collection and Preprocessing

To ensure representativeness, we selected four distinctive tourist cities in different regions: Harbin (northeast resource-based), Huangshan (natural landscape-based), Chongqing (urban sightseeing-based), and Sanya (southern coastal-based). We used the "LocoySpider" crawler tool [14] to collect blog data from Mafengwo.com [15] in November 2016, covering the period from January 1, 2013, to December 31, 2015. This yielded 1,347 Harbin blogs, 1,671 Huangshan blogs, 1,916 Chongqing blogs, and 2,349 Sanya blogs. To ensure accuracy, blogs mentioning multiple destinations were removed, resulting in 821 Harbin blogs, 969 Huangshan blogs, 1,234 Chongqing blogs, and 2,183 Sanya blogs. Destination information, blog content, and corresponding travel dates were structurally extracted and stored in a database.

4.2 Sentiment Annotation

(1) Sentiment Dictionary Selection
Compared to machine learning methods, dictionary-based scoring is simpler, more efficient, and suitable for engineering applications, though its effectiveness heavily depends on dictionary quality [16]. Major Chinese sentiment dictionaries include HowNet , NTUSD , and the Chinese Emotion Ontology . This study employs the Chinese Emotion Ontology published by Dalian University of Technology' s Information Retrieval Laboratory in 2012, which classifies emotions into seven categories: Joy, Good, Anger, Sadness, Fear, Disgust, and Surprise, containing 27,352 Chinese sentiment words.

(2) Sentiment Annotation Process
Match sentiment words from the ontology against blog texts, recording occurrence positions. For negation cases, we adopt Du Zhenlei' s [17] sentiment shift method: if a negation word appears within the preceding five characters, apply sentiment shift and record the new category and adjusted intensity.

For example, in "Honestly, both the dumplings and noodles weren' t very tasty; without the vinegar and chili, I couldn' t eat them," the sentiment word "tasty" belongs to [Good] category with intensity 3. Due to the negation "not,"sentiment shifts to [Disgust] with reduced intensity of 0.6.

4.3 Sentiment Feature Score Calculation

(1) Emotion Mean Calculation
The mean score eliminates the influence of blog volume, accurately reflecting overall sentiment characteristics for a month. For month Moni, we average the [Joy] emotion means across all blogs to obtain MonAvgJoyi, as shown in Formula (1):

$$MonAvgJoyi = \frac{\sum_{k=1}^{n} AvgJoy_k}{n}$$

where AvgJoy_k is the mean [Joy] score for blog Dk in month Moni (the sum of [Joy] word intensities divided by total sentiment word occurrences), and n is the number of blogs in month Moni.

(2) Emotion Standard Deviation
Standard deviation of emotion means across months reflects fluctuation magnitude. Using the standard deviation formula [18] for [Joy] across months yields StdJoy, as shown in Formula (2):

$$StdJoy = \sqrt{\frac{\sum_{i=1}^{12} (MonScoJoy_i - \overline{MonScoJoy})^2}{12}}$$

where MonScoJoy_i is the [Joy] sentiment score for month Moni.

5. Experimental Results Analysis

We extracted two sentiment feature indicators for each emotion category: monthly blog emotion means and monthly standard deviations. The mean reflects overall sentiment tendency for a month, while standard deviation indicates fluctuation magnitude across months.

Since all blog texts contain sentiment words, they represent subjective content suitable for quantitative sentiment feature analysis. Representing the blog corpus as D = {D1, D2, ···, Dn}, we calculated sentiment feature scores for seven emotion categories by month. As calculation methods are identical across categories, we use "Joy" as an example.

Applying the processing methods from Section 4 to our dataset yielded sentiment feature data for Harbin, Huangshan, Chongqing, and Sanya. Based on this data, we analyzed temporal distribution of blog volumes and emotional changes to discover temporal sentiment patterns, providing new reference channels for tourism managers and potential travelers.

5.1 Temporal Distribution of Travel Blog Volume

Extracting explicit travel dates from blog content reveals monthly blog volumes that partially reflect visitor numbers. Figure 2 [Figure 2: see original paper] shows temporal variations for the four destinations.

Overall monthly patterns show relatively high volumes in September-October and lower volumes in November-December, corresponding to China' s holiday schedule. Peak months vary by destination, indicating tourists plan destinations according to month. Distribution patterns differ significantly: Sanya peaked in September with 274 blogs, while Harbin' s lowest was April (21 blogs) and highest December (191 blogs), showing large temporal fluctuations and limited appeal outside winter. Huangshan and Chongqing show similar "M-shaped" distributions with spring and autumn peaks, closely related to local climate. Sanya maintains relatively high volumes year-round, only dropping in November-December, indicating mature tourism development.

5.2 Temporal Variation Analysis of Destination Sentiment

(1) Range of Emotion Means

Figure 3 [Figure 3: see original paper] displays monthly sentiment variation for the four destination types.

Across all destinations, [Good] emotion means are consistently highest and relatively stable, followed by [Joy] and [Disgust] fluctuating around 1.0, while the other four emotions remain below 0.5. [Good] includes respect, praise, affection, and wishes, indicating tourists' overall positive feelings and that blogs tend to recall pleasant experiences, generating positive online word-of-mouth.

Temporal characteristic 1: Among seven emotion categories, [Good] emotion means are consistently highest across all destinations and months, far exceeding other emotions and remaining relatively stable.

(2) Emotion Fluctuation Magnitude
Figure 4 [Figure 4: see original paper] shows standard deviations of monthly emotion means.

Standard deviation distributions are similar across destinations: [Good], [Joy], and [Disgust] show noticeable variation, while [Anger], [Sadness], [Fear], and [Surprise] show smaller fluctuations. This indicates varying expressive power among the seven emotions in temporal analysis.

Emotional fluctuation differs substantially from blog volume distribution. December is Harbin' s most popular month with highest visitor flow, yet not the month with strongest positive emotions. Periods exist with strong positive emotions but low visitor numbers, revealing development potential.

Temporal characteristic 2: Among seven emotion categories, [Good], [Joy], and [Disgust] show the greatest fluctuation across months, while [Anger], [Sadness], [Fear], and [Surprise] show minimal variation.

(3) Relationship Between Emotional Fluctuation and Blog Volume
Since [Good], [Joy], and [Disgust] reflect monthly emotional changes, we analyze their temporal patterns for each destination. Using Harbin as an example, Figure 5 [Figure 5: see original paper] shows the relationship between these three emotions and blog volume.

Temporal characteristic 3: Emotional fluctuation over time is not correlated with blog volume, indicating that traditional peak/off-season classifications are unrelated to tourists' actual emotional experiences.

5.3 Comparative Analysis of Different Destination Types

Harbin (northeast resource-based) and Sanya (southern coastal-based) represent different destination types with distinct temporal patterns. Figures 6 [Figure 6: see original paper] and 7 [Figure 7: see original paper] show their temporal variations.

Harbin exhibits more significant emotional fluctuations than Sanya, indicating destination-type differences in temporal sentiment patterns. Harbin shows higher positive emotions ([Good] and [Joy]) in January, July, November, and December (summer and winter), reflecting its 避暑 (summer cooling) and 冰雪 (winter snow/ice) attractions. However, [Disgust] is also high in winter, suggesting service improvement opportunities. April-May show below-average [Good] and [Joy] with rising [Disgust], indicating insufficient spring tourism product development.

Sanya' s [Good] emotion mean (annual average 2.47) exceeds Harbin' s with smaller fluctuations, indicating more stable and mature tourism development. May-June shows declining [Good] and [Joy] with rising [Disgust], likely due to hot weather and typhoons. December shows high [Good] and [Joy]; tourists seeking recreation should choose October-December (high [Joy]), while those preferring leisure sightseeing should select April, July, August, or December (high [Good]).

6. Conclusion

Travel blogs are becoming an important tourism information source, providing references for potential tourists and business opportunities for service providers. This study proposes a sentiment analysis model based on travel blog temporal characteristics, comprising five modules: data collection, preprocessing, sentiment annotation, statistical calculation, and result analysis. The model reveals temporal patterns in tourist emotions:

1. Among seven emotion categories, [Good] emotion means are consistently highest across all destinations and months, far exceeding other emotions and remaining stable.
2. [Good], [Joy], and [Disgust] show the greatest fluctuation across months, while [Anger], [Sadness], [Fear], and [Surprise] show minimal variation.
3. Emotional fluctuation is not correlated with blog volume, indicating traditional peak/off-season classifications are unrelated to actual emotional experiences.

Analysis of four destination types confirms these patterns. Months with above-average [Good] and [Joy] indicate high satisfaction and development potential. Months with high [Disgust] are not recommended for tourists; managers should analyze blog content to identify causes and implement improvements. Periods with strong positive emotions but low visitor numbers represent opportunities for potential tourists. The model thus provides new information channels for tourism managers and travelers by analyzing temporal sentiment fluctuations.

A limitation is that while blogs provide rich information, some emotional expressions may be inadequate, as highly dissatisfied tourists rarely post blogs. Future research will incorporate additional information sources to further refine temporal sentiment patterns and enhance reference value.

References

[1] Yang Ligong, Zhu Jian, Tang Shiping. Survey of Text Sentiment Analysis[J]. Journal of Computer Applications, 2013, 33(6): 1574-1578.

[2] Hu M, Liu B. Mining and Summarizing Customer Reviews[C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2004.

[3] Zhu Yanlan, Min Jin, Zhou Yaqian, et al. Semantic Orientation Computing Based on HowNet[J]. Journal of Chinese Information Processing, 2006, 20(1): 14-20.

[4] Shi Wei, Wang Hongwei, He Shaoyi. Study on Public Sentiment Based on Microblogging Platform[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(11): 1171-1178.

[5] Pang B, Lee L, Vaithyanathan S. Thumbs up?: Sentiment Classification Using Machine Learning Techniques[C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. 2002.

[6] Xu Jun, Ding Yuxin, Wang Xiaolong. Sentiment Classification for Chinese News Using Machine Learning Methods[J]. Journal of Chinese Information Processing, 2007, 21(6): 95-100.

[7] Banić L, Mihanović A, Brakus M. Using Big Data and Sentiment Analysis in Product Evaluation[C]//Proceedings of the 36th International Convention on Information & Communication Technology Electronics & Microelectronics (MIPRO). 2013.

[8] Wang Suge, Wu Suhong. Feature-Opinion Extraction in Scenic Spots Reviews Based on Dependency Relation[J]. Journal of Chinese Information Processing, 2012, 26(3): 116-121.

[9] Zheng Wenying. Sentiment Analysis of Travel Destination Reviews in Chinese[D]. Harbin: Harbin Institute of Technology, 2010.

[10] Jin Cheng. Research on the Dynamics and Change Mechanism of Tourist Emotions[D]. Guangzhou: South China University of Technology, 2015.

[11] Hu Chuandong, Li Lumiao, Luo Shangkun. Cycling Tourists' Experience of Scenic Byways Based on Content Analysis of Travel Blogs: A Case Study of the Sichuan-Tibet Section of National Highway 318[J]. Tourism Tribune, 2015, 30(11): 99-110.

[12] Yu Jing. Research on Tourist Emotion and Spatio-temporal Variation Based on Microblog Big Data[D]. Xi' an: Shaanxi Normal University, 2015.

[13] Li Q, Wu Y, Wang S, et al. VisTravel: Visualizing Tourism Network Opinion from the User Generated Content[J]. Journal of Visualization, 2016, 19(3): 1-14.

[14] LocoySpider [CP/OL]. [2016-11-15]. http://www.locoy.com/.

[15] mafengwo.com [DB/OL]. [2016-11-15]. http://www.mafengwo.cn/.

[16] He Tao. Research on Chinese Blog Information Gathering and Opinion Retrieval[D]. Hefei: University of Science and Technology of China, 2009.

[17] Du Zhenlei. Sentiment Analysis Towards Microblog Short Text[D]. Beijing: Beijing Information Science and Technology University, 2013.

[18] Standard Deviation [EB/OL]. [2016-11-15]. https://en.wikipedia.org/wiki/Standard_{deviation}.

Author Contributions

Cheng Cuiqiong, Xu Jian: Conceived research ideas and designed the study; Cheng Cuiqiong: Conducted experiments, collected, cleaned, and analyzed data; Cheng Cuiqiong, Xu Jian: Drafted and revised the manuscript.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is available in the online version of the journal at http://www.infotech.ac.cn.

Received: 2016-10-07
Revised: 2016-11-30

Note: Figure translations are in progress. See original paper for figures.