Summer 7-5-2020

Using Text Mining and Sentiment Analysis To Explore Tourists Consumer Focus From Online Reviews – Taking Mausoleum Of The First Qin Emperor As Example

Hsiao-Ting Tseng
Department of Information Management, National United University, Miaoli, Taiwan

Lirong Xue
Department of Information Management, Tatung University, Taipei, Taiwan

Ming-Hsien Chen
Department of Information Management, Tatung University, Taipei, Taiwan, mhchen@ttu.edu.tw

Using Text Mining and Sentiment Analysis To Explore Tourists Consumer Focus From Online Reviews – Taking Mausoleum Of The First Qin Emperor As Example

Hsiao-Ting Tseng[1], *Lirong Xue*[2], *Ming-Hsien Chen*[2*]
[1]Department of Information Management, National United University, Miaoli, Taiwan
[2] Department of Information Management, Tatung University, Taipei, Taiwan

Abstract: With the development of the economy, high-quality free travel has become a mainstream leisure tourism method, and tourism-related information has also grown exponentially. Coupled with the diversity of information sources, tourist attraction consumers received a lot of fragmented information. Previous research pointed out that tourist attraction consumers' decision-making basis is increasingly relying on electronic word of mouth. However, the variety of reviews on the Internet makes it easier for tourist attraction consumers to make timely or even wrong judgments due to information integration errors. In order to solve the problems mentioned above, this research is based on big data text mining and sentiment analysis processing analysis, using the existing electronic travel review data to conduct mining analysis, in order to recommend the most useful review information to tourist attraction consumers, allowing tourist attraction consumers to make effective decisions. In other words, tourist attraction consumers can enable users to get advance reminders before making decision and presented with visualization. In this way, tourists who are consumers of tourist attractions can receive the information they need quickly and logically, and quickly make decision. Then, improve user satisfaction. Finally, results provide tourist attractions operators as a reference to improve and strengthen their core business contents and priorities.

Keywords: tourist attraction consumers, attraction information visualization, big data text mining, sentiment analysis

1. INTRODUCTION

1.1 Background

In the era of big data, people are constantly exploring how to find potential value from huge amounts of data [1]. It is estimated that unstructured materials account for about 80% of the current social data structure [2]. Therefore, analysis of unstructured data is inevitable [2]. Unlike structured data, unstructured data cannot be used directly in the database and must be reprocessed before it can be used.

This study is concerned that a large amount of unstructured data is generated every day in the website of the tourism industry, but most of these data are just scattered around the webpage or stored in the database and piled of ash, due to existed website design requirement dilemma (cannot show visualization of big data), which will be cleared after a period of time. If the data can be processed and analyzed, the value of the data can be tapped and the utilization rate of the data can be increased.

Most of the time, due to the geographical distance, in addition to the official introduction of the scenic spot, tourists can only learn more about the comprehensive information of the scenic spot by browsing travel reviews. Tourists write travel reviews to share their travel experiences, convey more information about the scenic spot, and reduce the unequal information among tourists about the scenic spot. However, everyone's focus and needs for play are different. Usually, after browsing dozens of reviews or travel notes, they only find the information they need.

In the operation mode of travel reviews, because reviews are written by tourists, the contents of these

* Corresponding author. Email: 12345@public.wh.hb.cn(Elizabeth Whitworth) , 123@163.com(Brian Whitworth)

reviews do not have a qualitative standard, and the quality is not uniform. In many cases, users need to filter out many meaningless information by themselves or repeat the browsing. Not only wasting time but not focusing.

The authors believe that the website does not make full use of the review information of tourists, and we can use the information visualization method to rearrange the review information into images with visualization. Users can quickly find the key information from the charts, without having to absorb and filter by themselves, at a glance, easy to consult, and the graphical communication form also enhances the fun elements, so that the information browsing is no longer boring[3].

According to previous research, humans have a much faster absorption rate of image data than text data [4]. In order to make it easy for users to understand the comments, we collected the tourist comments and used Mausoleum of the First Qin Emperor as a case study, collect its related reviews for analysis and then visualization.

2. RESEARCH PURPOSE

The content of tourist comments generally includes feelings about the weather, traffic, people flow, tickets, and their own subjective experience. The authors hope to extract and count the high-frequency vocabulary through big data text analysis and sentiment analysis; perform positive and negative sentiment analysis between the contexts of the review content; use word vectors to build a multi-dimensional model to find out the relevance to the central topic vocabulary that appear more frequently. After the above processing, we hope to get a visualized image of tourist attraction that can help users, reduce user reading, summarize high-frequency information, and reduce subjective influence of reviews.

By understanding the visual images of reviews, consumers of tourist attractions can find out in the keywords section whether they are interested in the attraction, such as 'family tours', 'places of interest', 'shopping' and other needs; The score and the number of comments are obtained, and everyone's impression score of the scenic spot; from the comment cloud word map, you can see some of the important points that people often mention about scenic spots. The aim of this study are as follows:

(1) To integrate fragmentation information of tourism attraction.

(2) To help users quickly understand the tourism attraction.

(3) To improve reading interest, effectiveness and efficiency of tourist attraction consumers.

3. LITERATURE REVIEW

3.1 Visual Communication of Tourist Information

With the advent of the information age, the amount of data has grown tremendously, and people have used data visualization to statistically classify and present many structured data in charts and reports. But gradually, the data visualization no longer meets people's needs. We generate more and more textual data, which need to be processed. Information data images gradually enter people's vision.

Data visualization and infographics are two similar professional field names [5]. In simple terms, the difference between the two is that data visualization is to organize, process, and classify structured data through statistical charts, while information visualization is to visualize the logical relationship between information in an intuitive way. Show it like a form.

Willian Playfair published the book "The Commercial and Political Atlas" in 1786 [6]. The data-based diagrams in the book became the buds of the image processing of the data. Since then, the visual infographics have become people's reconstruction. Important method for abstract data and unstructured complex information. The concept of information visualization was first proposed by Stuart Card, Mackinlay and George Robertson in 1989. He directly reflects how information visualization technology can improve people's access to information.

Maximum capacity.

3.2 Development and Researches of Cloud Word Graph

The concept of tag cloud was proposed by Fernanda B. Viégas in the 2008 paper Tag Clouds and the Case for Vernacular Visualization and has been active in socially oriented websites [7]. They use tags to index and visualize information.

Stepchenkova [8] and others studied the image of Russian travel destinations through online surveys. By comparing the relevant information on Russian travel on the US travel website and Russian travel website, using high-frequency word analysis, etc. The information on Russia's definition of tourism image is incomplete.

William [9] analyzes the image of Seoul's tourist destinations through comparative semiotics of visual performance. He mainly studies the image of Seoul's destination by copying previous research and comparing it with traditional projection images in printed brochures and guides. Zhang, et al. [10] used text analysis methods such as word segmentation and word frequency statistics to analyze online reviews of books to provide decision-making basis for online bookstores.

4. METHODOLOGY

This article mainly collects travel review data, organizes, cleans, and processes word segmentation, and uses the obtained words for keyword extraction, sentiment analysis, and word vector calculation. We collected and analyzed the travel reviews generated by specific sites in a website within three months. The purpose is to enhance the value of reviews, construct regional keywords, and help users quickly understand the public 's reviews of attractions to reduce user decisions. time. We use web crawlers to crawl the user's travel review data on the website and use IBM SPSS 22.0 as analysis tool to clean and manage the data. The processed data is used for word segmentation, calculation of vectors, and sentiment analysis to obtain cloud word maps, keywords, and comment sentiment scores. The cloud word map is displayed in the comment area, keywords are used for 'labels', and comment sentiment scores assist in the calculation of partitions. Here is the research process of this study:

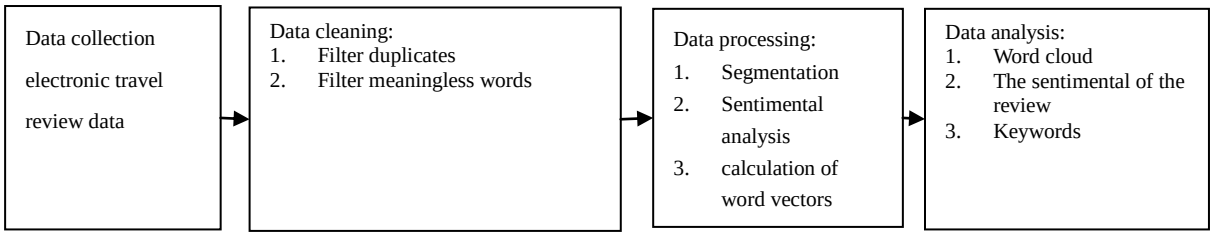


Figure 1. The research process of this study

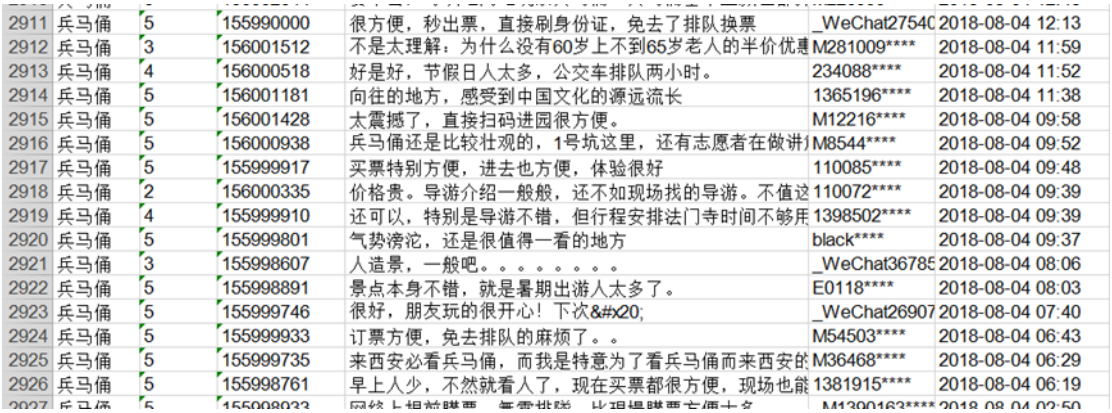
4.1 Data collection

There are three main forms of data: structured data, semi-structured data, and unstructured data. Structured data is mostly database files or tables, while semi-structured data is similar to xml and json format files, but in real life, it is more unstructured data. The webpage we collected for Ctrip's tourists' reviews of Mausoleum Of The First Qin Emperor is unstructured data in html. This kind of data requires us to scrape down the required individual by some means and then process it.

We use the request library in Python. Grab the reviews of each attraction through a certain URL. Due to the anti-crawling mechanism built into some websites, the URL of the review is hidden in the json file, so the original URL requires us to manually go to the source code of the webpage Rummaging. After finding the link, we started program crawling. Due to Ctrip's data protection settings, we were only able to crawl the latest 3051 articles in the past three months, so we crawled each of the attractions twice.

The anti-crawling sensitivity of the website is very high. During the crawling process, we implemented

guerrilla tactics-fighting each other, shooting one, and changing places. After being masked by multiple IPs, we successfully crawled down all the materials used in the topic. The screenshots of some information in this study as shown in Figure 2. The information we obtained in this study includes the five major parts: name of attraction, name of reviewer, review score, review content, and review time.



2911	兵马俑	5	155990000	很方便，秒出票，直接刷身份证，免去了排队换票	_WeChat2754(	2018-08-04 12:13
2912	兵马俑	3	156001512	不是太理解：为什么没有60岁上不到65岁老人的半价优惠	M281009****	2018-08-04 11:59
2913	兵马俑	4	156000518	好是好，节假日人太多，公交车排队两小时。	234088****	2018-08-04 11:52
2914	兵马俑	5	156001181	向往的地方，感受到中国文化的源远流长	1365196****	2018-08-04 11:38
2915	兵马俑	5	156001428	太震撼了，直接扫码进园很方便。	M12216****	2018-08-04 09:58
2916	兵马俑	5	156000938	兵马俑还是比较壮观的，1号坑这里，还有志愿者在做讲	M8544****	2018-08-04 09:52
2917	兵马俑	5	155999917	买票特别方便，进去也方便，体验很好	110085****	2018-08-04 09:48
2918	兵马俑	2	156000335	价格贵。导游介绍一般般，还不如现场找的导游。不值这	110072****	2018-08-04 09:39
2919	兵马俑	4	155999910	还可以，特别是导游不错，但行程安排法门寺时间不够用	1398502****	2018-08-04 09:39
2920	兵马俑	5	155999801	气势磅礴，还是很值得一看的地方	black****	2018-08-04 09:37
2921	兵马俑	3	155998607	人造景，一般吧。。。。。。。。	_WeChat36785	2018-08-04 08:06
2922	兵马俑	5	155998891	景点本身不错，就是暑期出游人太多了。	E0118****	2018-08-04 08:03
2923	兵马俑	5	155999746	很好，朋友玩的很开心！下次	_WeChat26907	2018-08-04 07:40
2924	兵马俑	5	155999933	订票方便，免去排队的麻烦了。。	M54503****	2018-08-04 06:43
2925	兵马俑	5	155999735	来西安必看兵马俑，而我是特意为了看兵马俑而来西安的	M36468****	2018-08-04 06:29
2926	兵马俑	5	155998761	早上人少，不然就看人了，现在买票都很方便，现场也能	1381915****	2018-08-04 06:19

Figure 2. The screenshots of some information in this study

Content of Figure 2 including following 4 elements:

1. Data collection electronic travel review data
2. Data cleaning: 1. Filter duplicates 2. Filter meaningless words
3. Data processing: 1. Segmentation 2. Sentimental analysis 3. calculation of word vectors.
4. Data analysis: 1. Word cloud 2. The sentimental of the review 3. Keywords

4.2 Data cleaning

Tourism websites usually provide a block that supports and is complemented by a corresponding reward mechanism to encourage tourists to write reviews. This caused some problems. Some tourists just wrote some humble reviews just to get rewards, such as' Nice, good, good, good, good. "Booming, sloppy, hip-hop," comments like this, which have no practical meaning, we will remove them when the data is cleaned, so that users can browse the essence.

In the materials we crawled from Ctrip, there are some duplicate contents, garbled characters, etc. We need to remove these parts to facilitate subsequent semantic understanding. The garbled is generated by the emoticons in the comments. This is because the encoding of the emoticons after crawling is inconsistent with our preset. In addition, we also filter repeated and meaningless words.

4.3 Data processing

We imported the data compiled by IBM SPSS 22.0 into Python, and used the Jieba word segmentation tool to perform two round cleaning and filtering to obtain the parts we needed; we performed word frequency statistics on the segmented vocabulary, visualized the frequency of part-of-speech words, and performed sentiment analysis. Discuss the accuracy of the existing data and propose ways to modify it; extract topic keywords from the comments to facilitate user indexing and meet users' various needs.

In the four popular word segmentation modules on the market-Jieba, SnowNLP, Pynlpir, Thulac and other word segmentation tools, For comparison, we chose to use the easier-to-use Jieba module as our word segmentation tool. Jieba has third-party vendor libraries in Python, which can be used by directly pip downloading. Jieba provides three kind of word segmentation modes, namely precise word segmentation mode, full word segmentation mode, and search engine mode, and Jieba provides a custom dictionary function. We can set relevant proper nouns based on the environment of the target text and those that have not appeared in the Jieba corpus. New words, which are helpful for subsequent processing and analysis. SnowNLP is mainly used

for sentiment analysis and comes with word segmentation tools. The word segmentation is not as detailed as jieba (if necessary, it can be corroborated by the picture above). The so-called surgery industry has a specialization. We will use the SnowNLP to perform sentiment analysis in the future.

4.4 Sentiment analysis

Sentiment analysis or opinion mining is people's opinions, emotions, and assessments of attitudes to entities such as products, services, and organizations. The development and rapid start of this field are due to the rapid development of social media on the Internet, such as product reviews, forum discussions, Weibo, and WeChat, because this is the first time in human history that there has been such a huge digital record. Since the beginning of 2000, sentiment analysis has grown into one of the most active research areas in natural language processing (NLP). Extensive research in data extraction, web mining, text mining and information retrieval.

According to the different nuances of processing text, sentiment analysis can be roughly divided into three research levels: word level, sentence level, and text level. Chapter-level sentiment classification specifies an overall sentiment direction / polarity, which determines whether the article (for example, a full online review) conveys overall positive or negative opinions. In this context, this is a binary classification task. It can also be a regression task, for example, an overall score inferred from a review of 1 to 5 stars. It can also be considered as a 5-level classification task.

The sentiment analysis of a sentence is inseparable from the sentiment of the words that make up the sentence. The sentiment analysis methods of words can be summarized into three categories: (1) knowledge-based analysis methods; (2) web-based analysis methods; (3) corpus-based analysis methods. The sentiment of words is the basis of sentiment analysis at the sentence or discourse level. Early text sentiment analysis mainly focused on judging the positive and negative polarity of the text. The corpus of sentiment analysis includes two types of vocabulary packages: positive vocabulary and negative vocabulary. According to the meaning scoring method, each comment is scored from words to sentences.

word2vec is also called word embeddings, Chinese name "word vector", it can convert words in natural language into dense vectors (Dense Vector) that the computer can understand. Before the advent of word2vec, natural language processing often turned words into discrete individual symbols, namely One-Hot Encoder.

1. Hangzhou [0,0,0,0,0,0,0,1,0, ..., 0,0,0,0,0,0,0]
2. Shanghai [0,0,0,0,1,0,0,0,0, ..., 0,0,0,0,0,0,0]
3. Ningbo [0,0,0,1,0,0,0,0,0, ..., 0,0,0,0,0,0,0]
4. Beijing [0,0,0,0,0,0,0,0,0, ..., 1,0,0,0,0,0,0]

For example, in the above example, in the corpus, Hangzhou, Shanghai, Ningbo, and Beijing each correspond to a vector. Only one of the vectors has a value of 1 and the rest are 0. However, using One-Hot Encoder has the following problems. On the one hand, the city codes are random, the vectors are independent of each other, and there is no relationship between the cities. Second, the size of the vector dimension depends on how many words are in the corpus. If the vectors corresponding to the names of all cities in the world are combined into a matrix, then this matrix is too sparse and will cause dimensional disaster.

Using Vector Representations can effectively solve this problem. Word2Vec can convert One-Hot Encoder into low-dimensional continuous values, that is, dense vectors, and words with similar meanings will be mapped to similar positions in the vector space. If the vectors corresponding to the names of all cities in the world are combined into a matrix, then this matrix is too sparse and will cause dimensional disaster. Using Vector Representations can effectively solve this problem. Word2Vec can convert One-Hot Encoder into low-dimensional continuous values, that is, dense vectors, and words with similar meanings will be mapped to similar positions in the vector space. If the city vector after embedding is visualized through PCA dimensionality reduction, this is what it looks like.

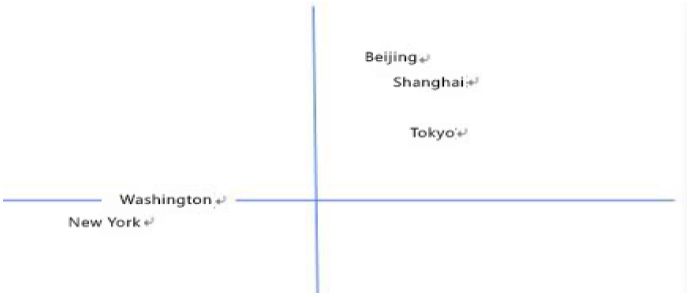


Figure 3. City vector after embedding is visualized through PCA dimensionality reduction

We can find that Washington and New York come together, Beijing and Shanghai come together, and the distance from Beijing to Shanghai is like Washington to New York. In other words, the model learns the geographical location of the city and the relationship between the city's status.

4.5 Analysis of travel reviews

This research crawls the tourist attractions and tourist reviews of some of the attractions on the Ctrip website. We extract the high-frequency words in the reviews and present them in a more interesting and intuitive way. Scoring makes a rough comparison, researches and finds factual problems, and proposes solutions to improve them. The comments are upgraded with word vectors to find related words on a topic, and to implement keyword query and other functions.

5. RESULTS

5.1 Word cloud of Mausoleum Of The First Qin Emperor travel reviews

We perform word frequency statistics and label classification on the review data of words. The larger the font appears in the picture; the same nature of words are marked with the same color, different colors represent different nature of words, and the colors of the nature of words are randomly assigned.

Figure 4. Word Cloud of Mausoleum Of The First Qin Emperor

5.2 Comparison of the sentimental of Mausoleum Of The First Qin Emperor travel reviews

The scores of scenic spots on travel websites are generally obtained by averaging the review scores. However, you will find that the content of many reviews and scores are inconsistent. We compared the review scores and review sentiments of Mausoleum of The First Qin Emperor attractions to prove that this phenomenon is more common.

According to SnowNLP's own corpus, this study use the Python third-party library SnowNLP to score sentiment on Mausoleum Of The First Qin Emperor.

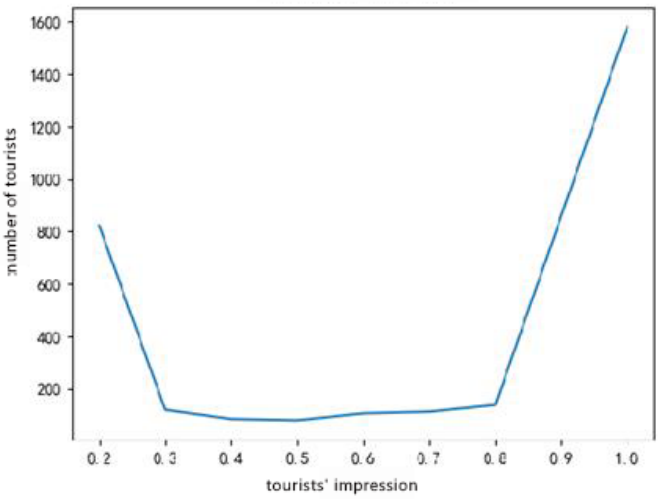


Figure 5. Score on Mausoleum of The First Qin Emperor. (x: tourists' impression; y: number of tourists)

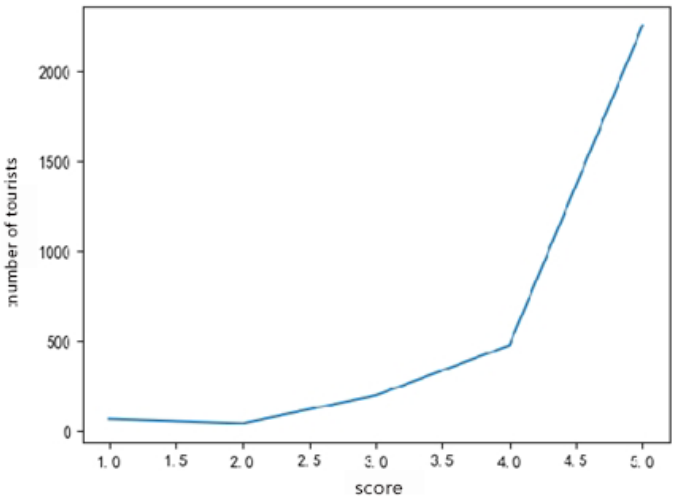


Figure 6. Score on Mausoleum of The First Qin Emperor. (x: score; y: number of tourists)

Figure 5 shows the user reviews sentiment and the number of users of The First Qin Emperor attractions. Figure 6 shows the user ratings and number of users of The First Qin Emperor attractions. It can be seen from the two figures that the good rating of content is about 60%, and the good rating accounts for about 90%. The conclusions reached by the two dimensions are not consistent, indicating that there is a personal subjective view of tourists when scoring or habitually giving full marks.

Gensim is an open source third-party Python toolkit for unsupervised learning from the original unstructured text to the topic vector representation of the text hidden layer. It supports a variety of topic model algorithms including TF-IDF, LSA, LDA, and word2vec. We use Gensim to implement the word2vec model.

Extract the keywords of the travel reviews of The First Qin Emperor and calculation, then remove the meaningless contents. This study summarized six keywords of The First Qin Emperor: children, crowds, convenient ticket access, scenery, tour guide, history. However, in order to facilitate user indexing, we have defined these 6 keywords as babies, popular attractions, quick entry to parks, places of interest, guided tours, and cultural heritage. Such daily expressions are simple, easy to understand and closer to the needs of users.

6. CONCLUSIONS

We aggregate the output results of the above three parts on one image. This presentation method is interesting and easy to read. From the high-frequency vocabulary counted by word cloud, it is found that many people in the Mausoleum of The First Qin Emperor need a tour guide when they visit. And based on the result of two figures of sentimental analysis, it is found that there are different curves between reviews and ratings, which represent the scoring doesn't exactly reflect tourists' real feeling about this visit. Therefore, this study concludes that scores do not represent the true views of tourists. Therefore, it is recommended that future tourism managers, tourists and researchers should no longer rely on biased quantitative scores but must retrospectively review the true original review data to listen tourists in order to get real opinions and get real public opinion tendencies, and further to create a higher quality tourism environment for tourists.

REFERENCES

[1] Jagadish, H., et al., Big data and its technical challenges. Communications of the ACM, 2014. 57(7): p. 86-94.

[2] Murdoch, T.B. and A.S. Detsky, The inevitable application of big data to health care. Jama, 2013. 309(13): p. 1351-1352.

[3] Card, S., J. Mackinlay, and B. Shneiderman, Information visualization. Human-computer interaction: Design issues, solutions, and applications, 2009. 181.

[4] Carney, R.N. and J.R. Levin, Pictorial illustrations still improve students' learning from text. Educational psychology review, 2002. 14(1): p. 5-26.

[5] Krum, R., Cool infographics: Effective communication with data visualization and design. 2013: John Wiley & Sons.

[6] Playfair, W., The commercial and political atlas: representing, by means of stained copper-plate charts, the progress of the commerce, revenues, expenditure and debts of england during the whole of the eighteenth century. 1801: T. Burton.

[7] Viégas, F.B. and M. Wattenberg, Timelines tag clouds and the case for vernacular visualization. interactions, 2008. 15(4): p. 49-52.

[8] Stepchenkova, S., Y. Chen, and A.M. Morrison. China and Russia: a comparative analysis of organic destination images. in The 11th APTA Conference Proceedings. 2005.

[9] Hunter, W.C., The social construction of tourism online destination image: A comparative semiotic analysis of the visual representation of Seoul. Tourism Management, 2016. 54: p. 221-229.

[10] Zhang, D., et al., Chinese comments sentiment classification based on word2vec and SVMperf. Expert Systems with Applications, 2015. 42(4): p. 1857-1863.