A Graphical Approach to Analyze Online Hotel Reviews Based on Text Mining

Li-Ching Ma and Zhi-Xuan Huang*
Department of Information Management, National United University, Taiwan
**Corresponding Author: M0733003@gm.nuu.edu.tw*

Received 25 February 2020; received in revised form 13 May 2020; accepted 3 June 2020

Abstract

People in todays' world are under great pressure due to their busy work, and tourism has become one of the popular ways for people to relax. The demand for travel accommodation has gradually increased. With the popularity of the Internet, customers often browse online customer reviews as a reference before hotel booking. This study takes hotel reviews on Booking.com as an example to analyzes online customer reviews of three well-known tourist hotels in Taiwan including E-da Royal Hotel, Caesar Park Taipei and Howard Beach Resort Kenting. This study adopts text mining to analyze customer reviews of the three hotels and find high-frequency terms. Then, similarity analysis and genetic algorithm are employed to discover the coordinates of the high-frequency terms. All high-frequency terms are displayed on a 2-dimensional graph based on the concept of social network maps. Managers can quickly observe customers' needs and preferences, and then adjust their strategies to enhance their competitive advantages.

Keywords: Text mining, visualization, customer review, booking website, hotel

1. Introduction

Modern people are under great pressure due to their busy work, and tourism has become one of the common ways for people to relax. Therefore, the demand for travel accommodation has gradually increased. According to the statistics of the Tourism Bureau, Republic of China, the total number of hotel stays in Taiwan has grown by nearly 3 million in recent ten years, showing a trend of rapid growth. Due to the increase in travel accommodation demand, the number of travel industry has also grown rapidly. Therefore, how to help the tourism industry to understand customer needs and then find a competitive advantage is a topic worthy of attention.

With the gradual improvement of national income, people are paying more and more attention to leisure activities, and the requirements for travel accommodation are getting higher and higher. In addition, due to the prevalence of the Internet, customers often browse online customer reviews as a reference before booking, and may evaluate the quality of the hotel according to the experience of the predecessors. Common sites with hotel reviews include Google Maps, Booking.com, Agoda, etc. This study takes hotel reviews of Booking.com as example to analyze because Booking.com's reviews include not only the basic information as other booking sites, such as evaluation scores, date and title, etc., but also customer's positive and negative comments; moreover, the number of comments of Booking.com is adequate for analyses.

Comments or evaluations can truly reflect the customer's ideas. Many studies used online comments as a target to analyze. For example, Lin (2015) used the questionnaire survey method to analyze the online food review contents and explored the factors affecting the practicality of the reviews. However, because most of the review contents are highly unstructured, traditional questionnaire methods are not easy to handle text comments. Therefore, some scholars proposed text mining approaches to analyze online reviews. For example, Cao et al. (2011) extracted semantic features from software reviews based on text mining and explored determinants of voting for the "helpfulness" of online user reviews. However, the results of traditional text mining techniques may not be easy for users to read and understand. If the results of data analysis can be presented in a graphical way, readers can quickly understand the meaning of data representation.

This study takes customer reviews of three well-known tourist hotels in Taiwan examples, including E-da Royal Hotel, Caesar Park Taipei and Howard Beach Resort Kenting. This study adopts text mining to analyze customer reviews of the three hotels and find high-frequency terms. Then, similarity analysis (Jaccard, 1901) and genetic algorithm (GA) (Holland,1975; Keshavarz & Abadeh, 2017) are employed to discover the coordinates of the high-frequency terms. All high-frequency terms are displayed on a 2-dimensional graph based on the concept of social network maps (Hu & Li, 2017). Managers can quickly observe customers' needs and preferences, and then adjust their strategies to enhance their competitive advantages.

2. Literature Review

After the rise of Web 2.0, people began to interact and share experiences and feelings on the Internet. Writing online reviews is a way to describe the experience in words. Many previous studies tried to find useful information from online reviews. For instance, Lin (2016) used the comparative opinions in the comments to analyze whether two companies are competitors. This study takes Booking.com as an example to analyze online hotel reviews based on text mining and visualization analysis.

Most of data we encounter every day are presented in unstructured text format. In order to analyze data with text format and find meaningful rules or patterns, many studies utilized text mining techniques for analyses. Text mining is the process of obtaining useful information from text. The main processing steps include data collection, preprocessing, data cleansing, data mining, modeling, and model evaluation (Zuo, 2018). In recent years, text mining methods have been widely employed to various fields. For example, Wang et al. (2017) applied text mining technology to fault diagnosis of railway systems that could effectively classify fault categories. Scandariato et al. (2014) adopted text mining to predict which parts of the Android application were more vulnerable, and constructed reliable prediction models. Because the online hotel reviews on Booking.com are composed of a large number of unstructured texts, the text mining technique is also adopted in this study.

In terms of visual analysis, graphical displays can reduce complicated information into simple patterns which readers are easier to observe and understand implicit meaning of data set. Many scholars have applied visualization methods to their research. For example, Tsai (2018) conducted an experimental study of elementary school children, and showed that visual programming could help reduce students' fear for conventional textual programming. Hu and Li (2017) proposed navigation graph models based on the concept of social network graphs to assist readers in searching TED Talks videos effectively. This study tries to develop a graphic approach for visualizing online hotel reviews based on the concept of social network graphs.

3. The Proposed Approach

3.1 Research Structure

The research structure of this study is shown in Figure 1. It can be divided into three parts, including data collection, data pre-processing and text mining, and graphical analysis. The data source of this research is from the customer reviews in Booking.com. The required data have been selected for subsequent analysis. The second part uses the data mining to perform word segmentation, and find out the high-frequency terms. The third part first calculates the similarity between terms and then employs the genetic algorithm to find the coordinate of each term. Finally, the results can be displayed on a 2-dimensional plane.

3.2 Data collection and Data Mining

In this study, we first selected the top ten international tourist hotels with the most number of rooms from the public information of the Tourism Bureau, Ministry of Transportation and Communications, Taiwan. The top ten international tourist hotels are listed in Table 1. Then, we searched for the number of reviews on Booking.com for these ten hotels. Three hotels with more than 300 reviews were chosen as research object, including E-da Royal Hotel, Caesar Park Taipei and Howard Beach Resort Kenting.

Table 1: Top 10 International Tourist Hotels with the Number of Rooms and Comments

Area	Hotel name	2019 /1 number of room	2018/05/01-2019/04/30 number of comments
Taipei	Grand Hyatt Taipei	866	< 200
Taipei	Sheraton Grand Taipei Hotel	692	< 100
Kaohsiung	E-Da Royal Hotel	656	468
Kaohsiung	85 Sky Tower Hotel	592	207
Taipei	Regent Hotels & Resorts	569	< 100
Kaohsiung	Hotel Grand Hi Lai	540	203
Taoyuan	Novotel Taipei Taoyuan International Airport	519	< 200
Taipei	Caesar Park Hotel Taipei	478	337
Pingtung	Howard Beach Resort Kenting	458	798
Kaohsiung	Ambassador Hotel	457	294

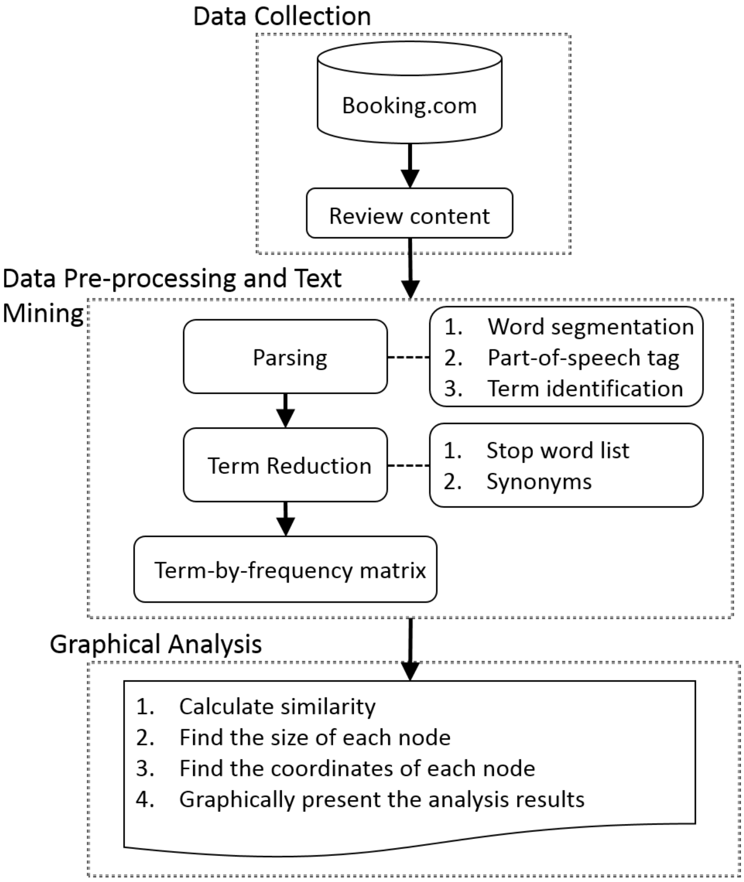

Figure 1: Research Structure

The three hotels are located in the north and south of Taiwan respectively, which can further explore the impact of cultural differences between North and South of Taiwan. The data collection period was from May 1, 2018 to April 30, 2019, and only comments written in Chinese were taken into account. A crawler program is developed to get required data from hotel reviews of Booking.com. The collected data include date of the comment, date of stay, score, title, positive comments, and negative comments.

After screening, 1,603 valid comments are selected, including 468 comments for E-Da Royal Hotel, 337 comments for Caesar Park Hotel Taipei, and 798 comments for Howard Beach Resort Kenting. Because the comments are in text format, we employ the data mining technique to perform parsing, reduce terms and generate a term-by-frequency matrix. In the parsing step, we adopt Jieba segmentation system and corpus to perform word segmentation which splitting a comment paragraph into individual words. Part-of-speech tagging identifies the part-of-speech of each word and classifies each word as a noun, verb, adjective, etc. Only nouns are identified as terms for further analyses in this study.

This study collected 1,603 comments from online reviews of three hotels in one year. After word segmentation, 43,924 words are recognized. After term identification and reduction, 1,763 terms are identified. Because the average frequency of terms in all reviews is 5.21, the terms with frequency higher than or equal to 6 are chosen as high-frequency terms. Finally, 202 high-frequency terms are selected for analyses.

Based on these high-frequency terms, a term-by-frequency matrix is built. If a term appears in the comment, the corresponding element in the matrix is assigned to 1; otherwise, it is assigned to 0.

3.3 Visualization

Finally, the concept of social network map is used to represent the relationship between the high-frequency terms and their relationships in the comments. The size of the term is used to indicate the frequency of the term appearing in the comment. Jaccard similarity (Jaccard, 1901) is used to measure similarity between two terms. A similarity matrix and distance matrix are then built. The distance between two terms indicates the similarity.

The genetic algorithm is employed to find the coordinates of each node. GA is based on the Darwin's principle of "Survival of the fittest".

The fitness function is to calculate the sum of absolute differences between the Euclidean distance and the dissimilarity, and the goal is to minimize the fitness value. In this study, the R language is adopted to find the coordinates of each term. The relevant parameters are set as the default value of R language. The initial group is 30 groups, the mating rate is 1.0, and the mutation rate is 0.01. The number of iterations is set to 1,000.

Finally, the results can be display on a 2-dimensional plane. This study uses the NodeXL graphical tool to present the high-frequency terms and their relationships. The size of a node is used to indicate the frequency of a term appearing in comments. The distance between two terms represents similarity. The higher the similarity, the shorter the distance yields; that is, the shorter distance indicates the higher the frequency of two terms appearing in the same comment.

4. Results

This study analyzes 1,603 online reviews of three well-known hotels in Taiwan. The top tree high-frequency terms include “room” with 613 times, “breakfast” with 463 times, and “employee” with 353 times. Top ten high-frequency nouns for three hotels are listed in Table 2. It can be seen that the customers’ overall comments on the three hotels focus on breakfast, rooms and staffs. Terms "breakfast", "room", "employee", "Facilities" and "Services" appear in the top ten lists of the three hotels. For Caesar Park Taipei, "Location", "Traffic" and "Station" are unique terms for this hotel. Caesar Park Taipei is popular for convenient transportation because its location is close to Taipei railway station. Regarding to Howard Beach Resort Kenting, "beach" is the unique term for this hotel. It may imply that people prefer this hotel because of its proximity to the beach.

Table 2: Top 10 High-Frequency Nouns for Three Hotels

Ranking	E-da Royal Hotel	Caesar Park Taipei	Howard Beach Resort Kenting
1	breakfast	location	room
2	room	room	breakfast
3	employee	employee	facilities
4	facilities	traffic	employee
5	bed	facilities	beach
6	children	Taipei	location
7	services	breakfast	swim pool
8	swim pool	bathroom	bed
9	bathroom	services	services
10	attitude	station	children

This study introduces a graphical approach to present the high-frequency terms and their relationships in the comments. The size of the node represents the frequency of the terms, and the distance between two nodes indicates similarity. Take E-da Royal Hotel for example. The graphical representation of E-da Royal Hotel is shown in Figure 2. From the size of nodes, it can be seen that the "breakfast", "room" and "employee" nodes are the top three high-frequency terms, among which “breakfast” is the largest one.

The distance between two terms represents their similarity. The genetic algorithm is adopted to find the coordinate of each node. The higher the similarity, the shorter the distance represents; that is, the shorter distance indicates the higher the frequency of two terms appearing in the same comment. From Figure 2, it can be observed that the distance between "room" and "facility" is relatively shorter than the distance between "room" and "children ". In addition, terms "employee", "service" and "attitude" often appear together in reviews, which may imply that customers pay more attention to service and attitude of the employee of this hotel. Customers with "children" pay close attention to "facilities". Regarding to "Environment", it is often discussed together with "Landscape" and "Restaurant".

For Caesar Park Taipei, terms "Taipei", "Station" and "Location" are often mentioned together indicating the hotel is close to Taipei Station and the location is convenient. "Elevator", "MRT" and "Exit" are often discussed in the comments at the same time, which also shows the convenience of transportation. For Howard Beach Resort Kenting, terms "Beach", "Swimming pool", "Facility" and "Location" often appear together representing that the hotel has swimming pool and related facilities, and is very close to the beach. In addition, "Space" and "Bed" are often mentioned together because this hotel offers large rooms with comfortable beds.

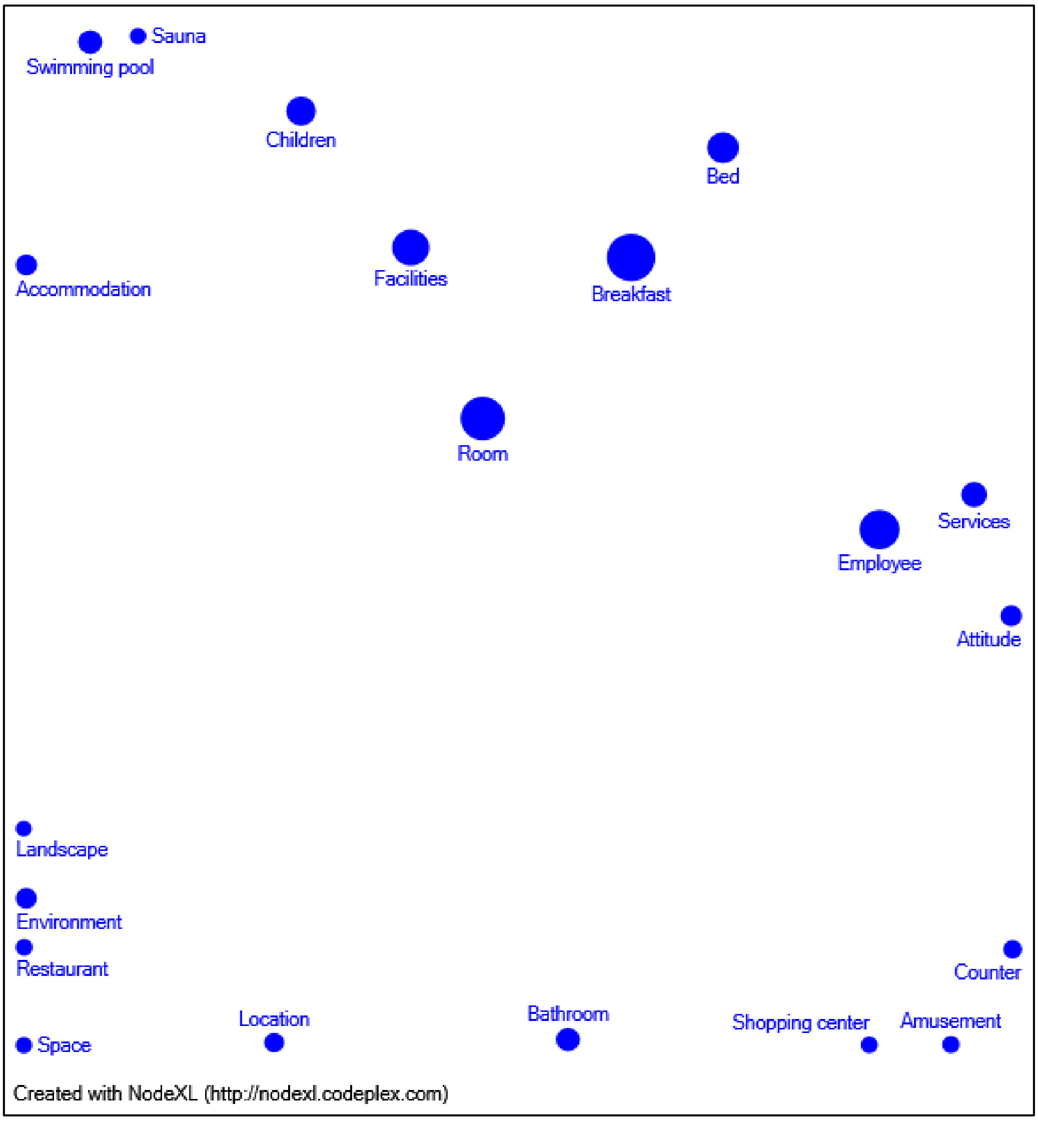

Figure 2: Graphical Representation of E-da Royal Hotel

5. Conclusion

This study proposed a graphical approach to analyze online hotel reviews based on the text mining. By incorporating the concept of similarity analysis, genetic algorithm, and social network map, online hotel reviews can be displayed on a 2-dimensional plane graphically. The high-frequency terms and relationships among those terms can be observed directly. The results show that, for E-da Royal Hotel, the top three high-frequency terms which often appear in the hotel reviews are "breakfast", "room" and "employee". The terms "employee", "service" and "attitude" often appear together in reviews, which may imply that customers pay more attention to service and attitude of the employee of this hotel. Customers with "children" pay close attention to "facilities". For "Environment", it is often discussed together with "Landscape" and "Restaurant". The results can assist hotel managers in quickly observing customers' needs and preferences, and then adjust their strategies to enhance their competitive advantages.

This research uses only nouns to analyze the key terms that are often discussed in specific hotels, and finds out the relationship between them. In further studies, adjectives, adverbs, positive and negative comments can be included and analyzed by sentiment analyses. Visualization techniques can also be applied to provide consumers an overview of the hotels at a glance.

References

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach. *Decision support systems*, *50(2)*, 511-521.

Holland, J. H. (1975). *Adaptation in natural and artificial systems.* Ann Arbor, MI: University of Michigan Press.

Hu, C.-C., & Li, H.-L. (2017). Developing navigation graphs for TED talk. *Computers in human behavior*, *66*, 26-41.

Jaccard, P. (1901). Étude comparative de la distribuition florale dans une portion des alpes et des jura. *Bull soc vaudoise sci nat*, *37*, 547-579.

Keshavarz, H., & Abadeh, M. S., (2017). ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. *Knowledge-based systems*, *122*, 1-16.

Lin, C.-H. (2015). *Antecedents of online review helpfulness in the gourmet food*. National Taipei University, Department of Business Administration Master Thesis.

Lin, Y.-H. (2016). *Competitor mining using customer reviews*. National Sun Yat-sen University, Department of Information Management Master Thesis.

Scandariato, R., Walden, J., Hovsepyan, A., & Joosen, W. (2014). Predicting vulnerable software components via text mining. *IEEE transactions on software engineering*, *40*(10), 993-1006.

Tsai, S.-C. (2018). *The eEffects of visual programming on primary school children's learning of text-based programming*. National Taipei University of Technology, Graduate Institute of Technological & Vocational Education Master Thesis.

Wang, F., Xu, T., Tang, T., Zhou, M.-C., & Wang, H. (2017). Bilevel feature extraction-based text mining for fault diagnosis of railway systems. *IEEE transactions on intelligent transportation systems*, *18*(1), 49-58.

Zuo, C. (2018). Defense of computer network viruses based on data mining technology. *International journal of network security*, *20*(4), 805-810.

About Authors

Li-Ching Ma is a Professor in the Department of Information Management at National United University, Taiwan. She received her PhD degree in Information Management from National Chiao Tung University, Taiwan. Her research interests include decision-making, optimization, data mining and visualization. Her articles have appeared in Decision Support Systems, OMEGA, Computers & Operations Research, European Journal of Operational Research, Asia-Pacific Journal of Operational Research, Computers & Industrial Engineering, International Journal of Information Technology and Decision Making, etc.…

Zhi-Xuan Huang is a graduate student in the Department of Information Management at National United University, Taiwan.