Evaluation and analysis of scenic spots and hotels

Runchen Huang, Xinyi Meng

School of Applied Mathematics, Zhuhai branch of Beijing Normal University, Zhuhai, Guangdong, 519000, China

Keywords: TF-IDF, F-score, Mean square error.

Abstract: To enhance the reputation of scenic spots and hotels and other tourist destinations is the work that the local cultural and tourism authorities and tourism related enterprises attach great importance to and pay close attention to. It will involve how to attract customers and obtain competitive advantages. In this paper, through the construction of a mathematical model, the F-score evaluation method is used to evaluate the service, location, facilities, health and cost performance of scenic spots and hotels, and the mean square error is used to evaluate the model. At the same time, the TF-IDF method is used to calculate the top four words with the highest frequency in all reviews and form a word set. On this basis, a model is built to analyze the effectiveness of online reviews of scenic spots and hotels. Finally, according to the average score of each index of the hotels in the scenic area, we make further specific analysis.

1. Introduction

To enhance the reputation of scenic spots, hotels and other tourism destinations is the work that local cultural and tourism authorities and tourism related enterprises attach great importance to and pay close attention to, which will involve how to attract customers and obtain competitive advantages. Tourist satisfaction is closely related to the reputation of the destination, and there is a positive correlation between them. Therefore, it is extremely important to improve customer satisfaction and improve the reputation of the destination.

2. Model Establishment and Solution

2.1 Scoring Model

2.1.1 Classification

Through manual classification or computer recognition, we can judge whether the words "service, location, facilities, $hygiene$, cost performance" appear in a hotel or scenic spot comment, and classify it into different categories.

2.1.2 Evaluate the Classification Method

We use F-score formula to evaluate whether the above method is suitable. Where p is precision and R is recall. From the calculation formula, it is not difficult to see that there is no direct connection between the two, but in fact, in large-scale data sets, the accuracy rate and recall rate often cannot have both sides and are mutually contained. In general, the higher P is, the lower R is, and the higher R is, the lower P is. The two are inversely proportional. β express the weight in the formula. when β >1, we usually think that the recall rate accounts for more, β<1 We think accuracy is more important. when we think accuracy is as important as recall β=1, so sometimes we often need to make a choice between the operation rate and the accuracy rate. In this problem, we think both P and R are very important, β is assigned to 1, so the common change is as follows:

$$F-score=\frac{1}{n}*\sum_{i=1}^{n}\frac{2*Pi*Ri}{Pi+Ri}$$

Therefore, the better the method is, the higher the F-score value is, and the worse the method is, the lower the F-score value is.

2.1.3 Preliminary Scoring

In each category of different comments that have been obtained, each comment needs to be scored. If words such as "very good", "like", "very good", "not bad" appear, we can give this comment 5 points. If "OK", "just so so" and other similar words appear, the comment can be given 4 points. If words such as "general" and "make do" appear, three points can be given. If there are words such as "not good" and "bad", 2 points can be given. If there are words such as "very poor" and "poor comment", 1 point can be given.

2.1.4 Calculate the Final Score

(1) After getting the scores of different reviews of different categories of different hotels or scenic spots, we need to calculate the arithmetic average score of different categories of reviews of each hotel or scenic spot, assuming that there are n comments, and the corresponding score of each i comment is N_i. Then the arithmetic mean score of this class $\overline{N}$ is as follows:

$$\overline{N} = \frac{\sum_{i=1}^{n} N_i}{n}$$

(2) In the above process, the arithmetic average score based on comments of a certain hotel or scenic area has been obtained. Because the experts' scores are known, the final score can be obtained by calculating the weighted average score based on comments and the scores given by experts. Let the weight of the obtained score based on comments be r, the weight of the expert score be s, and the expert score be m. The weighted average score A is expressed as:

$$A = rN + sM$$

The formula can modify the weight according to the need. If the expert score and the score based on comments are considered equally important, it can be expressed as follows:

$$A = 0.5 * N + 0.5 * M$$

2.1.5 Using the Mean Square Error to Evaluate

From the previous step, we have got the new score of each hotel or scenic area. We need to analyze the mean square error between the new score and the expert score to judge whether the model is effective. Suppose that the weighted average score of the k-th hotel or scenic spot's p-th category is A_{jkp}

$$J = \begin{cases} 1, & hotel \\ 2, & scenic\ spot \end{cases}$$

$$k = \begin{cases} 1, & 01 \\ \cdots\cdots\cdots & \\ 50, & 50 \end{cases}$$

$$p = \begin{cases} 1, & \text{service} \\ 2, & location \\ 3, & \text{facilities} \\ 4, & hygiene \\ 5, & cost\ performance \end{cases}$$

$$MSE = \frac{1}{J * k * p} * \sum_{j=1}^{2} \sum_{k=1}^{50} \sum_{p=1}^{5} (A_{jkp} - M_{jkp})^2$$

The smaller the MSE value is, the more reliable the model is. Through the above weighted scoring method, we can get new scores for hotels and scenic spots. The MSE is 0.00298. Because the numerical value is very small, it shows that the model fits well.

2.2 A model for judging the validity of reviews

2.2.1 TF-IDF method

Today, there are tens of thousands of comments, of which the content related accounts for the vast majority, while the content irrelevant or simple copy and paste and no effective content still account for a small number. So, we can first use TF-IDF method to calculate the top words with the highest frequency in all comments and form a word set.

Calculate word frequency: word frequency (Q_i) = the number of times a word appears in the text. It can be expressed as:

$$\sum_{i=1}^{n} Q_i$$

Suppose that the total number of words in all comments is R, and the word frequency (TF) = (the number of times a word appears in the text) / (the total number of words in the text), which can be expressed as:

$$TF = \frac{\sum_{i=1}^{n} Q_i}{R}$$

Through the above calculation, several phrases with the highest frequency can be calculated. It is summarized into a large word set.

2.2.2 Correlation Judgment

Each comment is matched with a large word set. If the number of overlapped words is more, the relevance of the comment is higher. If the number of overlapping words is less or even zero, it can be preliminarily judged that the content of the comment is irrelevant.

2.2.3 Simple Copy Modification Judgment

Simple copy and modification of the comments often have a high degree of coincidence, we can use the machine to judge the coincidence degree between one comment and the rest of the comments. If it exceeds a certain duplicate check value, it can be considered that there must be a comment between them, which is a simple copy modification. All the comments with high coincidence degree can be deleted together, and only one comment can be retained.

2.3 Characteristic Analysis

2.3.1 Hierarchical

The hotels and scenic spots are divided into three levels: high, middle, and low, and then the average score of each category is calculated for each level of hotels and scenic spots. Through the calculation, we can see that the average score of each level of scenic spots and each category is as follows.

For the scenic spots：

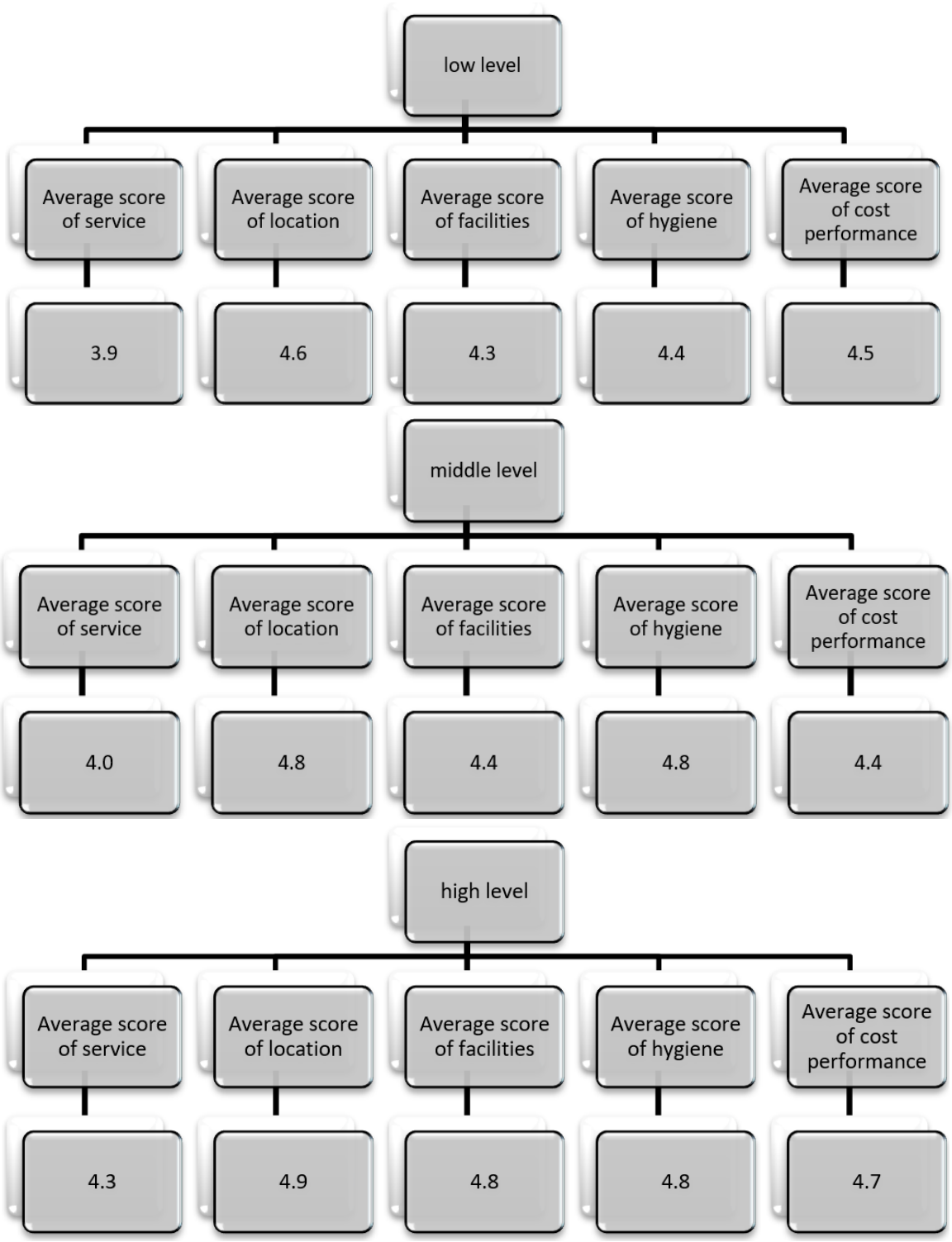

Figure 1. Average scores of scenic spots at different levels

For the hotel:

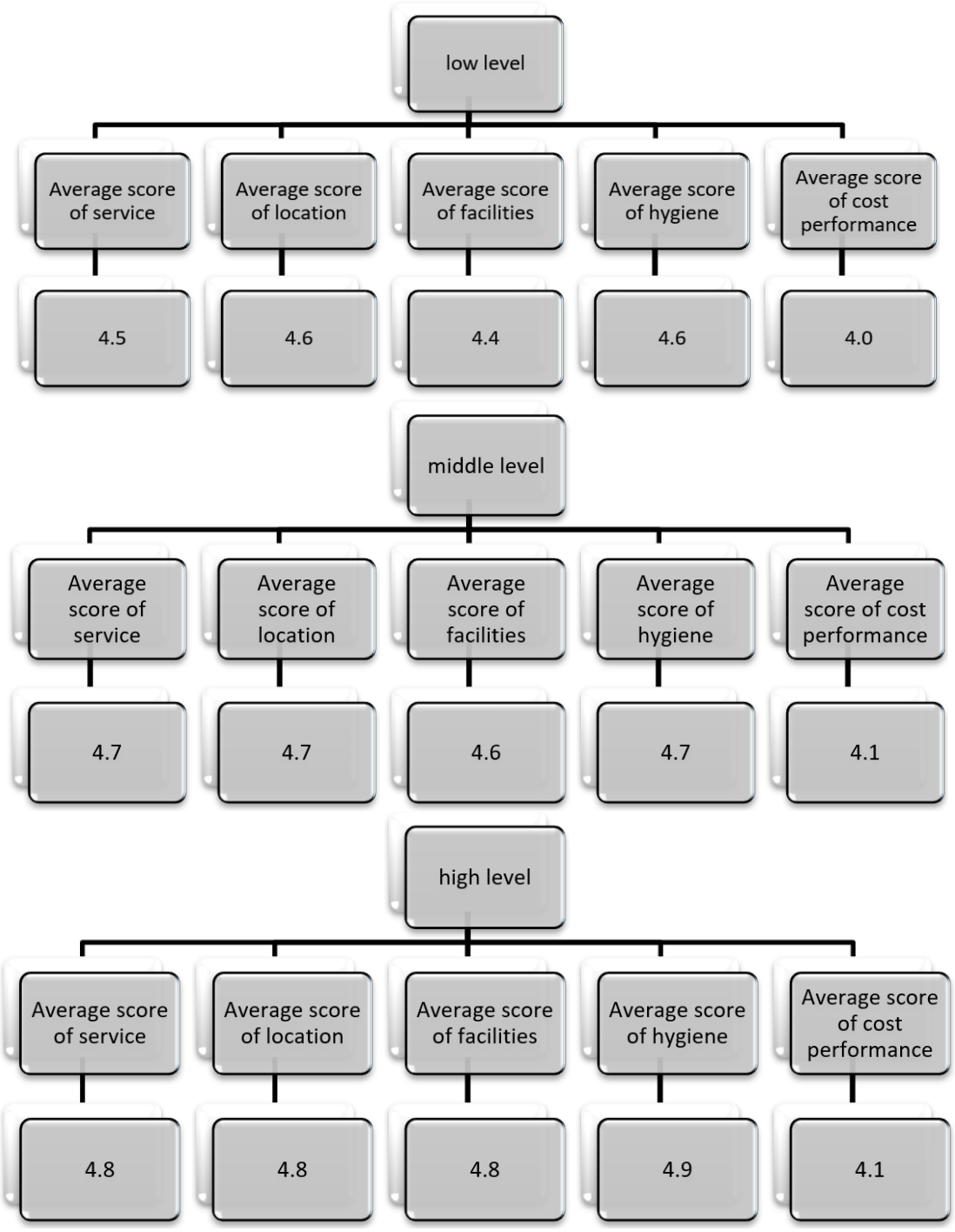

Figure 2. Average score of hotels at different levels

From the above average score, it can be preliminarily judged that the weak point of the scenic spot is the service category, and the average score of the service category in the high, middle, and low levels is far lower than that of the other major categories. It can also be judged that the weakness of hotels is cost performance. The average score of cost performance in high, middle, and low levels is far lower than that in other categories, and the score of cost performance in different levels is almost the same. Therefore, when the total score is almost the same and the scores of other major categories are almost the same, customers may first consider the service score and service comments of the scenic spot, as well as the cost performance score and cost performance comments of the hotel. Next, we need to analyze the reviews of hotels and scenic spots. We need to randomly select three hotels in each level of scenic spots and hotels.

2.3.2 Analysis and comment on the scenic spot to get the results

(1) From the low-level randomly selected to A24, A10, A42 three scenic spots, through the analysis of their comments, we can draw the conclusion: A24 main advantage is location and cost-effective, of which the cost-effective advantage is the most prominent; The main advantage of A10 is health; The main advantages of A42 are location and facilities, of which facilities are the most prominent.

(2) Three scenic spots, A13, a16 and A17, are randomly selected from the middle level. Through the analysis of their comments, we can draw the conclusion that the main advantages of A13 are service and location, among which service is the most prominent; The main advantages of a16 are location hygiene and cost performance, among which the cost performance is the most prominent; The main advantages of A17 are location, facilities, and sanitation, among which facilities are the most important.

(3) We randomly selected A21, A35 and A38 scenic spots from the high-level, through the analysis of their comments, we can draw the conclusion: the main advantages of A21 are cost performance and health; The main advantage of A35 is facilities; The main advantages of A38 are health service and cost performance, among which the service is the most prominent.

2.3.3 Analyze and comment on the hotel and get the results

(1) H18, H29 and H47 hotels were randomly selected from the low level. The main advantage of h18 is health; The main advantages of H29 are service and location, of which location is the most prominent; The advantages of H47 are health and cost performance, and the mid-term cost performance is the most prominent.

(2) Three hotels H05, H09 and A17 were randomly selected from the middle level. The advantages of H05 and H05 are service, location, facilities, and sanitation; The advantages of H09 are service, location and cost performance; The advantages of H31 are service, location, hygiene, and cost performance, among which hygiene is the most obvious.

(3) H04, H06 and H23 hotels were randomly selected from the high level. The advantages of H04 are service, location, facilities, and sanitation, among which facilities are the most obvious; The advantages of H06 are service, location, facilities, sanitation, and cost performance, among which the cost performance is the most obvious. The advantage of H23 is cost performance.

3. Model Evaluation

3.1 Model Advantages

(1) The use of F-score can accurately judge the advantages and disadvantages of the classification method.

(2) Through the establishment of the model and the use of Excel, the amount of calculation is greatly reduced.

3.2 Model Disadvantage

(1) In the process of scoring the comments, the attitude words in the comments cannot be fully identified.

(2) There will still be large errors and omission of important information in the message by replacing the message with more frequent word set.

References

[1] Sheng Ju. Probability theory and mathematical statistics [Fourth Edition]. Zhejiang University.

[2] Jia Junping. Statistics [7th Edition]. Renmin University of China.

[3] Zhang Xianliang, Zhang Yousai. Design of hierarchical search engine based on TF-IDF algorithm [J]. Computer and digital engineering, 2021, 49 (03): 456 - 461.

[4] Chen Liping, Liu Xiaobing. Analysis of hotel service quality evaluation system at home and abroad [J]. Contemporary tourism, 2021, 19 (15): 37 - 40.