Personalized recommendations based on hotel online review data

Haifei Qin¹, Wanjin Xu¹, Yongchao Zhao¹, Chunlin Wang and Xiaolin Chen¹*

¹Information Department Chuxiong Normal University, ChuXiong, YunNan 675000, China
*Corresponding author’s e-mail: qhfei@edu.cn

Abstract: With the development of network technology, independent travel is more and more popular. This paper constructs Hotel recommendations based on online review data. Firstly, we are starting from data acquisition, perceive and obtain the data that the online overview from the multi-agent-based on JADE. Secondly, we carry out semantic analysis of online comment data, according to the selecting, filtering, identifying, and merging data features, gain the feature of the hotel. Then, classifies online comment data and ranks hotels using combined feature words. Finally, we provide a personalized recommendation to users.

1. Introduction
With the emergence of network technology, the kinds of travel way are increasing and occupying more and more. In the face of the change of customers' travel mode, the hotel industry pays more attention to the shift in consumer behavior, consumption preferences, check-in experience and word of mouth, etc. The online hotel recommendation is born. At present, there are two ways of online guidance. One is to recommend according to the user's score, this recommendation is relatively simple, mainly according to the user of the hotel number score. Another is the semantic analysis of online hotel reviews data, gain the rating of the proposal, positioning types, and specifications of the hotel. When a customer is looking for the hotel, the matching of customer search types and specifications, make a personalized recommendation to the customer.

2. Related work
The choice of accommodation is the primary task for most tourists [1][2][3]. In the decision-making process of consumers, online comments have become a vital source of information [2]. Some people choose the product if the product is recommended by others.

At present, the hotel online review data is still in the various systems, more scattered, not centralized. Group intelligence perception technology is still in the initial implementation stage, but also cannot generally provide fast and robust support for the online hotel review big in data mining. To obtain a large amount of data requires use intelligence acquisition, which is called crawler technology. We apply the technique to get the data of online review data based on JADE.

Online review analysis belongs to the point of mining, but it is not the same as emotion. It is part of mining belong to opinion mining. In 2017, Sun et al. [4] and Li Jianhua et al. [5] made a further summary of viewpoint mining. In 2018, Han Zhongming et al. [6] made a new review on the aspect view mining method of network comments. Hotel online reviews data from different dimensions depict the different characteristics of the hotel and the mixed feelings of the customer. Through
semantic mining, understand the overall features of the hotel and customer experience.

Hotel online reviews find similar groups of customers based on their past accommodation experience and provide personalized recommendations for customers based on similar experiences of related groups. When a tourist wants to choose a hotel on the internet, he will first browse the introduction and online comments of the hotel near the destination, and then make a decision [7]. Consumers are seeking for the peer experience increasingly when preparing to purchase products or services. There is over 60% of consumers consult customers’ feedback before purchase [7]. Hotel reviews determine the hotel’s online image, sales, and future revenue [7]. Personalized hotel recommendation based on customer reviews is a new method of hotel recommendation [8]. According to the comments of use may judge the type of users, location, and provide personalized hotel recommended for users [9].

3. Technology
Personalized recommendation based on the hotel online reviews data, including several stages of perception and acquisition of hotel online review data, semantic analysis, and personalized recommendation.

3.1 Perception and acquisition of online review data
An intelligent perception system of online review data was built on a distributed support platform based on JADE, which facilitates the distributed operation and deployment of the system. Control Agent is the brain of a crawler, which distributes tasks and communicates with the outside world, and monitors the activity of the entire crawler. Collection Agent, according to the collection plan in the configuration file, different ways of information collection can be carried out to meet various information requirements. The crawler is composed of Control Agent, URL Agent, Extraction Agent, and Collection agents.

3.2 Semantic analysis of hotel online review data
Semantic analysis and processing stage includes data cleaning and feature extraction.

3.2.1 Data clean
Data cleaning is a critical problem, a link to ensure the quality of data. It mainly includes data pre-processing (word segmentation, remove the punctuation, remove the stopping words, remove the low-frequency words, and remove the high-frequency non-feature words, etc.). After data cleaning, get the word data suitable for analysis.

3.2.2 Feature Extraction
Feature extraction adopts word cloud map, word frequency (TF), document frequency (DF), inverse document frequency (IDF), TF-IDF and other feature selection indexes, and feature selection adopts cluster analysis. According to the Cluster analysis to the index of TF, DF, IDF, and TF-IDF. The primary method of cluster analysis, as follows in Algorithm 1.

Algorithm 1: Feature mining clustering algorithm for online comment data
Step1: Separate the candidate feature words into one category \( \{X_1, X_2, \ldots, X_N\} \);
Step2: Calculate the distance between each class (Quasi-average method, Ward method, maximum distance method and similarity analysis method) to obtain the observed value matrix;
Step3: The two classes with the smallest distance between merged classes are new classes. And recalculate the distance between the new class and each class, update the matrix table, the total number of classes in order to decrease, until 1;
Step4: Draw a cluster tree;
Step5: According to the cluster diagram and expertise determine the number and members of the classification;
3.3 Personalized recommendation of hotel online review data

When entering the query conditions, the system will automatically extract the Spatio-temporal information of the customer, and connect the customer's past accommodation history and comments on the hotel, quickly search the hotel that meets the user's needs in the user's destination and propose intelligent, personalized recommendations for the customer.

4. Experiment and result analysis

4.1 Data acquisition and preliminary analysis
The crawler-based JADE was used to obtain more than 500,000 pieces of online review data of high-end hotels in China from the TripAdvisor website. Through data cleaning, more than 400,000 bits of data that met the requirements. And select more than 170,000 pieces of data of the top 20 hotels, and more than 100,000 pieces of data of the bottom 20 hotels according to the digital score for analysis. The word segmentation system and stuttering word segmentation were used to segment the data set. And obtain more than 4 million pieces and more than 180 items of data of the top 20 hotels.

4.2 Acquisition of feature words
We use the indicators of word frequency (TF), comment frequency (DF), inverse document frequency (IDF), and TF-IDF to reduce the dimension of the candidate feature words based on the word frequency, the paper integrates the methods of word analysis and clustering analysis. The characteristics of the hotel online review data obtained by unsupervised clustering analysis are as follows: service, environment, facilities, overall comfort, and catering. It concluded that the online hotel review of the key is easonable after 20 hotels, check data on the selected keywords. According to analysis, the index of the TF, DF, IDF, TF-IDF, obtain the feature words of online review is Overall comfort, service, facilities, catering, and location. This Context introduces in the therefor paper.

4.3 Hotel segmentation based on feature words
According to the feature words extracted in the paper, subdivide each comment of the first 20 hotels in the dataset one and annotated. The result of the partial subdivision annotation of hotels was in table 1:
Table 1 Classification of comments based on feature words

| Hotel number | Hotel name   | Online comment                                                                 | Feature words |
|--------------|-------------|-------------------------------------------------------------------------------|---------------|
| 1            | ZSWQ_HOTEL | The environment is very good, hot spring is also relatively clean, is a convenient point to drive. | Overall comfort |
| 2            | YX1_HOTEL  | The service is good and the room is clean and tidy. Recommended!             | service       |
| 3            | XM_HOTEL   | The hotel is very good, the service is very good, very careful, the room is very clean. | service       |
| 4            | YX_HOTEL   | Good, but the service for foreign guests needs to be improved                | service       |
| 4            | YX_HOTEL   | Quiet environment, quality service, feel comfortable!                        | service       |
| 5            | ATMD_HOTEL | Very good service attitude.                                                   | service       |
| 5            | ATMD_HOTEL | Close to the office, good service, will also consider check-in.              | service       |
| 5            | HASA_HOTEL | Good value for money, complete facilities. It’s recommended to stay here next time | facilities    |
| 6            | HASA_HOTEL | Hotel facilities is very new, lodging feel good, next time you must stay!   | facilities    |
| 7            | JLSM_HOTEL | Five-star hotel facilities are good, the pool is clean, the festival more children, are playing very happy! | facilities    |
| 8            | ATMD_HOTEL | Very few dishes, there is nothing to eat, taste so-so, too                   | catering      |

For dataset 1 (digital) score in the top 20 hotel online reviews data with statistical analysis, based on keywords to get hotel analysis diagram, as shown in figure 2.

![Hotel analysis based on feature words](image)

We can see from the figure 2 "ZSWQ" stand out in the overall comfort, in 10253 comments, there are 6678 comments mentioned overall comfort-related content, 62.13%. Can be seen from the figure 2 many hotels and customers are pay attention to the "services", such as "YX1_HOTEL" referred to in the 6686 comments service 4936, accounting for 73.83% of the comments, "XM_HOTEL" article 16227 in the 25189 comments data, accounting for 64.42% of the comments, and similar "YX_HOTEL", "LYHY_HOTEL", "LBJX_HOTEL" is more than 50%. To finding distinguish similar hotels, the ward clustering method was adopted in this paper to cluster 20 hotels. The clustering results are in figure 3:
According to figure 3 hotel clustering figure, we can see that a reasonable hotel divided into four class, class 1 is "WSY_HOTEL", "ZSWQ_HOTEL" in clustering. Furtherly, we can see that this type of hotel in the comments is most obtain "overall comfort" and "location" to illustrate this hotel pay attention to "the overall comfort level" and "location" and also get the recognition of customers. Class 2 is "YX_HOTEL", "XM_hotel", "YX1_HOTEL", most comments in "service", "CQGM_HOTEL", "LJWY_HOTEL", "JLSM_HOTEL" comments, at least in terms of "service" to illustrate this hotel pay attention to service or lacking in terms of service, needs to be improved. Class 3 is "SHDS_HOTEL," the most comments on "facilities," "GMLH_HOTEL" least in terms of "facilities" "comment". It illustrates these hotel notice facilities or to lack in terms of facilities, needs to be improved. The rest of the hotel enjoys a relatively balanced in all aspects. It will be a specific classification furtherly and further subdivided.

4.4 Personalized hotel recommendation
According to the method above, split each hotel's key stored in the system. When a customer arrives at a specific place, want wants to book a hotel. He only needs to enter conditions of the query, and the system will automatically extract keywords, time, and space information. And associated with the customer's history and comments on the hotel accommodation, quickly search out by user requirements in the user's destination hotel, puts forward the intelligent, personalized recommendation for customers.

5. Conclusion
This paper analyzes the hotel online review data and studies it through data perception acquisition, semantic analysis, and personalized recommendation. This paper only uses part of the hotel data, which is not comprehensive enough. However, this paper only verifies the feasibility of the study from the technical point of view, and the actual operability must adopt more in-depth data. In the next, we will study all the data in some areas, so that the operability will be stronger.

Reference
[1] Li, G., Law, R., Vu, H.Q., Rong, J., 2013. Discovering the hotel selection preferences of Hong Kong inbound travelers using the Choquet Integral. Tour. Manag. 36 (3), 321–330.
[2] Raguseo, E., Neirotti, P., Paolucci, E., 2017. How small hotels can drive value their way in infomediation. The case of ‘Italian hotels vs. OTAs and TripAdvisor’. Inform. Manag. 54 (6), 745–756.
[3] Hong-gang Peng, Hong-yu Zhang, Jian-qiang Wang Cloud decision support model for selecting hotels on TripAdvisor.com with probabilistic linguistic information, International Journal of Hospitality Management 68 (2018) 124-138
[7] Sun S, Luo C, Chen J. A review of natural language processing techniques for opinion mining systems[J]. Information Fusion, 2017, 36:10−25. [doi: 10.1016/j.inffus.2016.10.004]

[8] Xie, K.L., Zhang, Z. and Zhang, Z., The business value of online consumer reviews and management response to hotel performance, International Journal of Hospitality Management, 43(2014), 1-12.

[9] D. Gavilan et al. The influence of online ratings and reviews on hotel booking consideration Tourism Management 66 (2018) 53-e61

[9] Jian-Qiang Wang et al. Hotel recommendation approach based on the online consumer reviews using interval neutrosophic linguistic numbers, Journal of Intelligent & Fuzzy Systems 34 (2018) 381–394