International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n3.20
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Cluster Analysis of Restaurant Evaluation Features Based on UGC Data

Mengen Hu

Inner Mongolia University of Science and Technology Baotou, China
hme820895476@163.com

ABSTRACT

With the advent of the big data era, analyzing using User Generated Content (UGC) data has become popular research. Most of the current research is to use UGC data to analyze user preferences or to analyze the influence of UGC data on other users. This paper, however, is based on the perspective of merchants, using UGC data to analyze the advantageous features of merchants that are recognized by users, and to help merchants improve their brand image. In this paper, the original UGC data is preprocessed through Chinese word segmentation, de-duplication and other techniques to generate a review UGC corpus, and then the Word to Vector (word2vec) model is used to train the UGC corpus to obtain the word vectors, and then the words are clustered into word clusters according to the word vectors, and a number of representative words are picked out from each word cluster to generate a thesaurus of restaurant evaluation features. In this paper, 34 restaurant evaluation keywords are finally extracted and labelled as a total of 14 categories of restaurant evaluation features.

KEYWORDS

Data Digging; User Generated Content (UGC); Word to Vector (word2vec); Clustering Techniques

1. INTRODUCTION

With the rapid popularity of the Internet, the use of user generated content (UGC) data to analyze user behaviors has become a popular field of research, with wide applications in travel decision-making [1], healthcare [2], social media operations [3], e-commerce [4], etc. UGC is the content generated by the users when they freely express their views on people, events and things. UGC is the content generated by users when they freely express their views on people, events and things, which not only can fully reflect users' real thoughts on people, events and things, but also more valuable is that UGC data can also reflect the views that users can accept and understand [5]. Existing research has demonstrated that UGC data is both highly advantageous over traditional methods in extracting customers' demand for product characteristics [6, 7] and can be used to predict the persuasive power of advertisements on user purchases [8].

In recent years, UGC data about the food industry has increased dramatically as a result of the promotion of Volkswagen Dianping. In the past, UGC data mostly appeared in e-commerce platforms to evaluate the goodness of goods or the quality of logistics services. Nowadays, UGC data is also widely used in the evaluation of restaurants by consumers. Therefore, by using the UGC data of Dianping to cluster and extract the features of user evaluation of restaurants, we can obviously get the recognition of the advantageous features of merchants, which can help merchants to strengthen their advantageous features and enhance their competitiveness.

2. THEORETICAL OVERVIEW

User-generated content has become one of the most important sources of big data for business analytics [9]. User-generated content can help companies understand customer needs more fully and deeply so as to 1) improve the design of goods [10]; 2) manage and innovate products [11]; 3) analyze user preferences for product features [12]; 4) analyze the competitiveness of goods [13] and so on. All of these studies have amply demonstrated that UGC is an important source for extracting user requirements. However, since these studies mainly focus on mining the needs of user groups for commodity features, they seldom involve the matching of merchant features with user demand characteristics. In order to achieve the matching between user demands and restaurant features, further research is inevitably needed on how to mine users' demands from UGC data, so as to obtain the features of restaurants that are preferred by users.

Restaurants are closely related to human life, with the development of fast food culture and the acceleration of the pace of life, many consumers are gradually accustomed to dining in restaurants; at the same time, with the improvement of the quality of life, consumers have higher and higher requirements for the quality of restaurants. Therefore, it is particularly important to understand the characteristics of consumers' preference for restaurants and the characteristics of restaurants that attract consumers. Therefore, this paper is dedicated to analyzing and extracting features from UGC data of consumers' evaluation of restaurants to understand the advantageous features of restaurants that attract consumers.

3. RESEARCH DESIGN

3.1. Research Target

In this paper, the public dataset of the popular Dianping website was selected, and after the pre-processing of the data, a total of 147,314 consumer review data of restaurants were selected, which provided effective data support for the cluster analysis of this paper.

3.2. Research Framework

In this paper, from the perspective of restaurant merchants, online user review data is obtained from the public dataset of the popular Dianping platform as a basis for mining the advantageous features of restaurants that attract consumers. Based on the characteristics of the restaurant and the active reviews of customers, the UGC data is analyzed by clustering using natural language processing methods such as python to find the restaurant features that attract consumers. The main process of this paper is to extract the keywords describing the features of the restaurant from the UGC data and classify them according to the categories of the features they describe, and the specific implementation steps are as follows: 1) Massive user review UGC data of the restaurant will be collected through the web public dataset platform, which contains the textual information of the user's review of the restaurant; 2) The original UGC text will be preprocessed through Chinese word splitting, de-duplication and other techniques to Generate the restaurant review UGC corpus [14]; 3) Use the word2vec (word to vector) model in natural language processing technology to get the language model trained from the restaurant review UGC corpus, in order to project the words into a spatial vector according to the semantics, which is called the word vector [15]; 4) Use the clustering technology in machine learning to cluster the words into several word clusters based on the word vectors; and pick a number of word clusters from each word cluster. words into a number of word clusters; from each word cluster, a number of representative words are selected; 5) The representative words are manually identified as keywords describing the characteristics of the restaurant and labelled with the category to which they belong (e.g., the keyword ‘fresh’ is labelled as ‘taste’); 6) The corresponding keyword categories are retained to generate a library of restaurant evaluation feature

words. (e.g., the keyword ‘delicious’ is labelled as the category ‘taste’); 6) retain the corresponding keywords to generate a thesaurus of restaurant evaluation features. The specific framework flow is shown in “Fig. 1”.

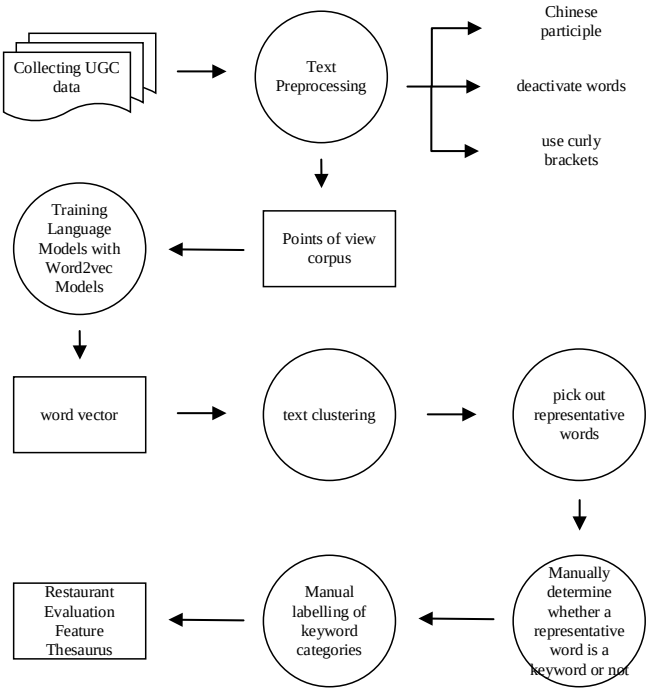

Figure 1. Research framework

3.3. Research Methodology

According to the research direction of this paper, the research methods to be adopted in this paper are as follows: 1) Since this paper needs to go through the literature analysis to understand the current status of the existing research on feature clustering using UGC, and to determine the direction and focus of the research, this paper needs to use the literature analysis method; 2) This paper will use a small amount of human to identify the clustered feature keywords whether it is a restaurant evaluation feature thesaurus or not, which is more accurate than simply using the computer technology has a higher accuracy rate, at the same time, the clustering method uses machine learning related technology, which is more efficient compared to manual classification, so this paper needs to use the human-computer combination method; 3) Since this paper needs to deal with large-scale text-based data, it needs to transform the text data into word vectors to facilitate computer processing, so this paper needs to use the natural language processing technology; 4) Since this paper needs to mine the feature keywords from user generated text data to mine feature keywords, so this paper needs to use data mining methods.

4. DATA DIGGING

4.1. Data Collection and Cleaning

In this paper, python data analysis tool is used to perform cluster analysis of restaurant evaluation features. The first step of data mining in this paper is data preparation. The dataset for this paper comes from the public dataset of the Dianping website. The initial dataset contains 240,000 merchants, which include restaurants, attractions, hotels and so on; the evaluation comes from 540,000 users, which contains 4.4 million reviews and ratings data. The second step is data preprocessing. Data preprocessing is an essential preliminary work for data mining, in the massive amount of raw data, there are a large number of repetitive, vacant and dirty data, which seriously affects the validity and

correctness of the data. Therefore, the collected raw data must be preprocessed before data mining to improve the efficiency, accuracy, and performance of data mining. Data preprocessing mainly includes data cleaning, data integration, data transformation and so on.

In this paper, as the original dataset has been preliminarily cleaned, only the information such as reviews and ratings in the original dataset are retained and other useless information is removed. Therefore, we only filter the data, firstly, we use excel table to filter the data, select the restaurant review information in it, and remove the related review information of attractions and hotels. Then because we need to mine the characteristics of restaurant reviews generated by user reviews, we only save the user ID and user-generated review content information, which greatly reduces the scope of the dataset and improves the accuracy of our data mining. In the end, a total of 147,314 consumer review data of restaurants were selected.

4.2. Text Clustering

After data cleaning, pycharm data analysis tool is selected to start our experimental part of this paper. In order to make the data understandable by the computer, first of all we have to preprocess the data. In this paper, we firstly use the jieba split word in gensim library to process the text data, in which we carry out the Chinese split word and de-duplication and other algorithms, and finally form the text data as shown in “Fig. 2”, that is, the corpus of the review that we need.

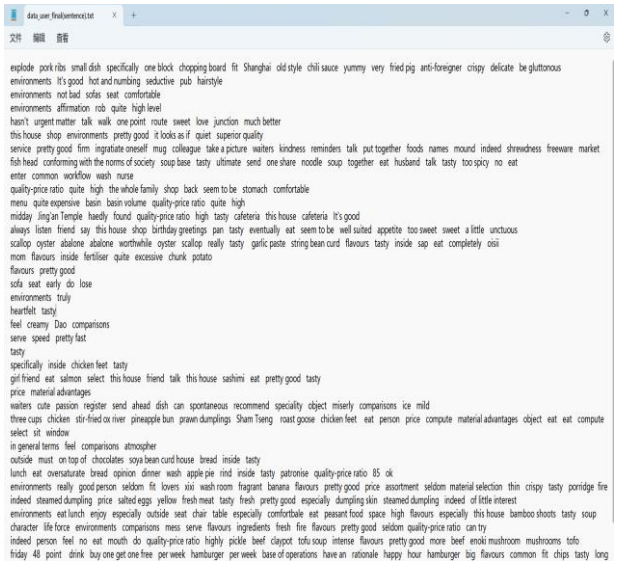

Figure 2. Partial data after disambiguation

After the participle processing, in order the data to be processed by computers, we need to transform the text data into vectors that are easily understood by computers. Therefore, we use word2vec to train a model on a corpus of reviews that have been segmented in order to generate word vectors that can be processed by a computer. After generating word vectors, they are then clustered using clustering techniques in machine learning to generate representative words. Since some of the representative words generated by clustering do not match the characteristics of the restaurant, the feature keywords are first selected manually, as shown in Table 1.

Table 1. Featured Keywords

Serial Number	Featured Keywords
1	saturation
2	portion size
3	have a nice cooked texture
4	salty
5	spicy
6	cold/hot
7	greasy
8	thickness
9	crispiness
10	sour-sweet
11	air conditioning
12	product category
13	crowded and noisy
14	queuing time
15	character
16	long queuing times
17	service attitude
18	deceive consumers
19	service level
20	service speed
21	environmental cleanliness level
22	environmental hygiene
23	food hygiene
24	public evaluation
25	brand impact
26	overall impression
27	decoration environment
28	style
29	price level
30	promotional activities
31	atmospheres
32	convenience of location
33	likes or dislikes
34	size of space

As can be seen from Table 1 above, after manual selection, this paper has selected 34 feature keywords, which encompass most of the characteristics of the user's evaluation of the restaurant, so this paper will be artificially tagged with the 34 feature keywords tagged as 14 restaurant evaluation feature classes. The first category is ‘portion size’, which includes saturation size, reflecting whether the restaurant makes consumers feel full and the size of the portion; the second category is ‘taste’, which includes cooked, salty, spicy, cold/hot, greasy, thick, thin, crispy, sweet and sour, reflecting consumers' recognition of the taste of the restaurant's dishes. The second category is ‘taste’, which includes the taste of cooked, light, spicy, cold/hot, greasy, thick/thin, crispy, sweet and sour, reflecting the degree of consumers' recognition of the taste of the restaurant's food. The fifth category is ‘people flow’ contains a lot of noise, queuing time, popularity, queuing time is long, mainly reflecting the popularity of the restaurant and the lack of staffing arrangements; the sixth category is ‘service’ contains service attitude, deception of consumers, service level, service speed, mainly is the consumer's opinion of the restaurant's service attitude, deception of consumers, service level, service

speed, mainly is the consumer's opinion of the restaurant's service. The sixth category is ‘service’, which includes service attitude, cheating consumers, service level, service speed, which is mainly the evaluation of consumers' attitude towards the service personnel of the restaurant; The seventh category is ‘cleanliness’, which includes the level of environmental cleanliness, environmental hygiene and food hygiene, reflecting consumers' evaluation of the overall hygiene environment and food safety of the restaurant; the eighth category is “evaluation”, which includes the public evaluation, brand impact and overall impression, which is mainly the word-of-mouth of consumers and the brand effect of the restaurant formed over time; the ninth category is “decoration”, which includes the decoration environment and style. The eighth category is ‘evaluation’, which includes public evaluation, brand influence and overall impression, mainly because of the word of mouth of consumers and the brand effect formed by the restaurant for a long time; the ninth category is ‘decoration’, which includes the decoration environment and style, and a good decoration and popular style of a restaurant will bring more consumers and repeat customers; the tenth category is ‘price’, which includes price level and preferential activities, and the performance of the restaurant is the best choice. Containing the price level and preferential activities, cost-effective restaurants can often attract more customers, and in the cost-effective basis to carry out preferential activities can improve the reputation of the restaurant; eleventh category is the ‘environment’, which is mainly embodied in the atmosphere, a good dining atmosphere can bring great benefits to the restaurant; twelfth category is the ‘location’. The twelfth category is ‘location’, which is mainly about the convenience of the location, such as parking, traffic and so on, a good location can attract more consumers; the thirteenth category is ‘preference’, the degree of personal preference for the restaurant also determines whether the restaurant meets the consumer's taste. The last category is ‘space size’, which is mainly the size of the seat space, consumers want to have a comfortable space when consuming in the restaurant, which determines the consumer attitude. The specific clustering results are shown in Table 2.

Table 2. Clustering Results

Clustering Category	Category Contains Keywords
component	saturation, portion size
taste	have a nice cooked texture, salty, spicy, cold/hot, greasy, thickness, crispiness, sour-sweet
facilities	air conditioning
merchandise Type	product category
people flow	crowded and noisy, queuing time, character, long queuing times
service	service attitude, deceiving consumers, service level, service speed
cleanliness	environmental cleanliness level, environmental hygiene, food hygiene
evaluation	public evaluation, brand impact, overall impression
decoration	decoration environment, style
price	price level, promotional activities
environment	atmosphere
location	convenience of location
preferences	likes or dislikes
size of space	size of space

4.3. Category Sorting

According to the clustering results in Table 2, we have classified the feature keywords into 14 categories, and in order to be able to understand more clearly how each category affects consumers in these categories, we sort them into categories. It can be seen that, regarding the degree of consumer care for evaluation features, it can be judged from the number of keywords in the text clustering of this paper, the higher the number of keywords, proving that the more UGC data consumers comment

on this clustered feature, then the higher the degree of care for this clustered feature. The specific sorting results are shown in Table 3.

Table 3. Category Sorting

Clustering category	Category Sorting
Taste	1
People flow, service	2
Cleanliness, reviews	3
Portion size, decor, price,	4
Amenities, variety of products, environment, location, preferences, size of space	5

Table 3 can be seen, the most consumers are interested in no surprise is the taste of the restaurant, which proves that consumers no matter when to go to the restaurant to consume, whether the food to meet the consumer's taste is the most important; followed by the flow of people in the restaurant and the service, the restaurant's service can be a major feature of the restaurant to attract consumers, such as the current Hai Di Lao hot pot, which is famous for its service, and the flow of people has always been a relatively important part of the various restaurants. Focus on one, if too many people in a reasonable arrangement of seats and let consumers reduce the waiting time and waiting for the sense of irritability is more important; clean and evaluation of the third, nowadays due to the rapid development of the Internet, the popularity of major review sites such as Meituan, the evaluation of other consumers has become a consumer choice of one of the necessary conditions of the restaurant, and at the same time, due to the enhancement of the awareness of the citizens of health, cleanliness and hygiene has also become a reference for consumers, and the restaurant has become an important part of the restaurant's service. Cleanliness and hygiene have also become one of the reference conditions for consumers. Portion size and price have always been the characteristics of restaurants that consumers care more about, and decoration is a plus for restaurants. The other features at the end of the list are the ones that restaurants can add to their own goodwill and attract consumers to come back for more.

According to the above ranking and elaboration, this paper can finally focus on the advantageous characteristics of the restaurant that consumers care about, and put forward improvement suggestions for the restaurant: 1) the restaurant needs to improve its own taste of food to meet the specific taste needs of different customers; 2) the restaurant is trying to improve the flow of people, improve their own waiter service level, and make preparatory measures for the excessive flow of people; 3) the restaurant needs to ensure that the restaurant environment hygiene and food safety, and strive to get every consumer's positive evaluation. Strive to get every consumer's positive evaluation, but also through related activities to promote consumers in the review site for positive comments; 4) restaurants can be based on their own appropriate profitability, increase the amount of food, preferential food prices, and can be appropriate to launch promotional activities to enhance the restaurant's cost-effective; 5) in other such as the environment, geographic location and other plus points, the restaurant can be appropriate to improve.

5. CONCLUSIONS OF THE STUDY

The work of this paper mainly lies in the use of UGC data for clustering restaurant evaluation features, on the basis of clustering manually classified, selected the most important consumer attention of the restaurant can attract consumers' advantageous features, and ultimately these features are sorted to select the order of the features that consumers care about, and based on these analyses for the restaurant to improve the improvement of the views. From the research of this paper, it can be seen that this paper finally categorized a total of 14 categories, and the categories will be sorted to get the

consumers care most about the taste of the restaurant, which is in line with reality, proving that the research has practical significance.

REFERENCES

[1] J. M S, Sierra R. The role of user-generated content in tourism decision-making: an exemplary study of Andalusia, Spain [J]. Management Decision, 2024, 62(7):2292-2328.

[2] Yu X, Wang H, Chen Z. The Role of User-Generated Content in the Sustainable Development of Online Healthcare Communities: Exploring the Moderating Influence of Signals [J]. Sustainability, 2024, 16(9):

[3] Bernd W. Wirtz, Vincent Göttel, Paul F. Langer, Marc-Julian Thomas. Antecedents and consequences of public administration's social media website attractiveness [J]. International Review of Administrative Sciences, 2018, 86(1): 38-61.

[4] Meng L, Kou S, Duan S , et al. The impact of content characteristics of Short-Form video ads on consumer purchase Behavior: Evidence from TikTok [J]. Journal of Business Research, 2024, 183114874-114874.

[5] Bin Fang, Qiang Ye, Deniz Kucukusta, Rob Law. Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics [J]. Tourism Management, 2016, 52: 498-506.

[6] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, Wenwu Zhu. Learning Disentangled Representations for Recommendation [A]; Neural Information Processing Systems [C], 2019.

[7] Artem Timoshenko, John R. Hauser. Identifying Customer Needs from User-Generated Content [J]. Marketing Science, 2019, 38(1): 1-192.

[8] Xiao Fang, Paul J. Hu. Top Persuader Prediction for Social Networks [J]. MIS Quarterly, 2018, 42(1): 63-82.

[9] Erick Kauffmann, Jesús Peral, David Gil, Antonio Ferrández, Ricardo Sellers, Higinio Mora. Managing Marketing Decision-Making with Sentiment Analysis: An Evaluation of the Main Product Features Using Text Data Mining [J]. Sustainability, 2019, 11(15): 4235-4253.

[10] Young Kwark, Jianqing Chen, Srinivasan Raghunathan. User-Generated Content and Competing Firms' Product Design [J]. Management Science, 2018, 64(10): 4471-4965.

[11] Deng Yuan. Intelligent innovative knowledge management integration method based on user generated content [J]. Cluster Computing, 2019, 22: 4793–4803.

[12] Anning Wang, Qiang Zhang, Shuangyao Zhao, Xiaonong Lu, Zhanglin Peng. A review-driven customer preference measurement model for product improvement: sentiment-based importance–performance analysis [J]. Information Systems and e-Business Management, 2020.

[13] Thi T T N, Shurong T. The impact of user-generated content on intention to select a travel destination [J]. Journal of Marketing Analytics, 2022, 11(3):443-457.

[14] Zhang S, Hu Z, Zhu G, et al. Sentiment classification model for Chinese micro-blog comments based on key sentences extraction [J]. Soft Computing, 2020, 25(prepublish):1-14.

[15] Machine Learning; Findings from National University of Defence Science and Technology Broaden Understanding of Machine Learning (Paragraph Vector Representation Based on Word to Vector and CNN Learning) [J]. Computer Weekly News, 2018, 446-.