Hindawi
Scientific Programming
Volume 2023, Article ID 9847260, 1 page
https://doi.org/10.1155/2023/9847260

*Retraction*

# Retracted: Construction of Machine Learning Model Based on Text Mining and Ranking of Meituan Merchants

**Scientific Programming**

Received 8 August 2023; Accepted 8 August 2023; Published 9 August 2023



This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

## References

[1] Y. Tang, D. Liao, S. Huang, Q. Fan, and L. Liu, "Construction of Machine Learning Model Based on Text Mining and Ranking of Meituan Merchants," *Scientific Programming*, vol. 2021, Article ID 5165115, 9 pages, 2021.

Hindawi
Scientific Programming
Volume 2021, Article ID 5165115, 9 pages
https://doi.org/10.1155/2021/5165115



*Research Article*

# Construction of Machine Learning Model Based on Text Mining and Ranking of Meituan Merchants

**Yin Tang,[1] Dongxue Liao,[1] Shuqiang Huang,[2,3] Qing Fan,[4] and Liang Liu[5]**

[1]*Management School, Jinan University, Guangzhou 510632, Guangdong, China*
[2]*College of Science and Engineering, Jinan University, Guangzhou 510632, Guangdong, China*
[3]*Guangdong Provincial Key Laboratory of Public Finance and Taxation with Big Data Application, Guangzhou 510320, Guangdong, China*
[4]*School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, Sichuan, China*
[5]*Venture Capital Research Center, South China University of Technology, Guangzhou 510632, Guangdong, China*

Correspondence should be addressed to Liang Liu; dongxuel@stu2018.jnu.edu.cn

Received 27 October 2021; Accepted 29 November 2021; Published 10 December 2021

Academic Editor: Bai Yuan Ding



In the Web 2.0 era, the problem of uneven quality and overload of online reviews is very serious, and the cognitive cost of obtaining valuable content from them is getting higher and higher. This paper explores an effective solution to address comment overload by means of information recommendation in order to improve the utilization of online information and information service quality. This paper proposes a review ranking recommendation scheme that focuses on the information quality of reviews and places more emphasis on satisfying users' personal information need. The paper's approach is used to extract and rank low-frequency keywords that appear only once in the comment set. The more useful the extracted phrases are, the more useful this review will be and the higher the usefulness votes will be, which can reflect the actual situation of this product more objectively and accurately and facilitate better consumption decisions for consumers. The experimental results show that users' satisfaction with the perceived usefulness of the reviews is jointly influenced by the information quality of Meituan's reviews and users' individual information needs; the recommendation strategy achieves the organic integration of the two, and the evaluation results under three different recommendation modes show that compared with "interest recommendation" and "utility recommendation," the satisfaction score of "fusion recommendation" is the highest

## 1. Introduction

With the rapid growth of the Internet and e-commerce platforms in recent years, the usefulness of online reviews has become an important influencing factor in consumer decision making [1]. Online reviews are users' evaluations and experience after experiencing commercial products and services and providing valuable information to other users. Users can learn about merchants' products and services through online reviews, which help them make better consumer decisions and reduce the reference cost of products and services. The famous Jupiter Research company, through years of research and analysis, found that 75% of consumers refer to reviews on the Internet before spending money on dining, travel, and accommodation, purchasing goods, parent-child playgrounds, and many other things. The same is true in China, with platforms such as Taobao, Jingdong, Meituan, and Where to Go [2]. Due to the openness of the Internet, the cost of posting online reviews is very low, and a lot of spam and false information make the quality of information in reviews vary, resulting in a large number of reviews, which is noisy and difficult to distinguish, and there are many ways of reviews and different language expressions, and some reviews do not bring us useful reference value [3].

"Taobao" uses whether there is a picture, whether there is a follow-up review, and the rating of the product as the filtering criteria; "public review network" blocks untrustworthy content based on user feedback; "Douban" and "Amazon" use user votes to sort reviews [4]. These filtering strategies focus on information quality and help users quickly access useful information by placing high-quality reviews at the top. Nevertheless, these filtering strategies do not focus on satisfying individual users' needs [5]. The adoption of information by individuals, besides being influenced by the quality of information, is related to individual information need, and people will care more about whether the information they receive contains content of interest to them. Especially when the amount of information exceeds one's cognitive load, people browse quickly and hope to find the content they are interested in as soon as possible.

In this paper, we propose a low-frequency keyword extraction method for review usefulness voting. The main purpose is to identify low-frequency keywords from the reviews of Meituan and to provide consumers with more choices and decisions through the study of usefulness voting, instead of just looking at the star rating given by users as the judgment index (usually five stars). Therefore, the identification and extraction of low-frequency keywords become a major difficulty for us, which mainly has the following three problems:

(1) The cohesiveness among the parts of low-frequency keywords is weak, and it is impossible to calculate the mutual information among them.

(2) Since the combination of low-frequency keywords is evaluated randomly from the perspective of probability, it is difficult to use machine learning methods by means of labeling.

(3) Low-frequency keywords also have the problem of representation, because of the low number of occurrences and the lack of contextual information. It is difficult to represent them by existing representation methods (e.g., Word2Vector).

Based on the above difficulties, there are still no more studies on the effectiveness of comment voting, which will become a key topic for our research.

## 2. Related Work

*2.1. A Study of Reviewing Ranking and Recommendation Based on Reviewing Utility.* The essence of the review ranking is to evaluate the utility of reviews and generate a Top N recommendation list based on the utility evaluation. In recent studies, [6] used fuzzy hierarchical analysis and weighted gray correlation analysis to predict the review utility, rank the reviews accordingly, and select the reviews with high information content for final recommendation. Jiang and Mccomas [7] used K-means algorithm to rank the review utility and then optimize the review ranking. Korde [8] calculated the credibility of reviews based on the number of "feature-opinion" pairs in the reviews and then invited users to evaluate the Top N reviews by questionnaire. Wen-Hsiang et al. [9] concluded that the authors' historical reviews reflect the quality of his or her published reviews and they modeled them based on the authors' previous reviews and incorporated them into the review model. It can be seen that the ranking and recommendation of reviews are mainly based on the calculation of evaluation metrics. In these studies, the evaluation metrics focus on a series of elements such as the information and content of the review, the credibility, the level of the writer, and the overall perceived utility of the reading group, which play a crucial role in identifying high-quality reviews.

A recent study, however, points out that the above evaluation indicators reflect only the quality of review information in terms of data reliability and do not emphasize the applicability of review information to the target information users [10]. Researchers argue that the evaluation of the perceived utility of online reviews is a kind of information quality assessment based on the user's perspective, which takes the user's subjective perception as the starting point to explore the utility of information and requires individuals to systematically assess the functional performance of information based on their personal experience [4, 5]. Therefore, user reviews in the online environment should not only be high-quality information that meets the standards but also focus on the degree to which the review information meets the needs and expectations of users and the value it brings to them [11]. There is no shortage of researchers who hold the same view. Hubertrajan and Dhas [12] explores product recommendations, and they argue that the validity of reviews should take consumers' individual preferences into account and look for high-quality reviews that match consumers' personal preferences. Ravi et al. [13] analyzed the quality of cloud service reviews on different online platforms to achieve review recommendations by calculating the similarity between the reviewer's personal information and the background information of the information seekers of the cloud service platform. All these studies take a personalized perspective to study the perceived value of reviews.

*2.2. Research on Review-Based Recommendation Systems.* Recommendation is an effective way to solve information overload, and, by probing users' information needs, recommendation systems can achieve information push oriented to personal interests and alleviate the distress caused by overloaded information [14]. The core of product recommendation system is to build an effective user and product model. Since review information is rich in users' evaluation of products, it has become a hot research topic in recent years to distill users' preferences and build user models from them and introduce them into recommendation systems. Mousavi et al. [15] classified the relevant research into three categories: lexical item recommendation, rating recommendation, and feature recommendation from the perspective of user modeling.

The lexical item-based recommendation is classified as content recommendation, which directly uses the review text

to model users and products. Seker et al. [16] extracted lexical items from users' published reviews and generates a user model with TF-IDF (term frequency-inverse document frequency) as lexical item weights, and the product model is based on the review set of the target product and finally makes recommendations based on the content similarity between the two. The literature recommendation system of [17] models the user based on the literature he has read, characterizes the lexical items with word vectors, and calculates the similarity between the user and the recommendation target (literature) up to the semantic level.

The collaborative recommendation mechanism used in rating recommendation requires the generation of a "user-rating" matrix, but the matrix sparsity problem has been a bottleneck in the performance improvement of collaborative recommendation systems. One of the solutions is to use the text data of reviews to predict users' ratings of products and then improve the "user-rating" matrix to improve the system performance. In [18], sentiment analysis was used to predict users' ratings of products based on their reviews, and a user model was built based on "predicted ratings" for product recommendation. Hiroshi [19] further improved the quality of the model by weighting the user ratings with the product theme information contained in the condensed reviews. Liu et al. [20] proposed a hybrid recommendation algorithm that integrates user ratings, sentiment, and product content and then recommended products by filling in the space "user-rating" matrix.

In summary, online reviews have been emphasized as an important information source for mining users' interests and preferences in recent research on recommendation systems. Collaborative recommendation strategies that use user reviews to generate user models or enhance the quality of the "user-rating" matrix by predicting user ratings of products are commonly adopted. These users and product models obtained from review text learning are characterized as hidden vectors, and probabilistic topic models and deep learning algorithms are widely used to improve modeling quality.

## 3. Model Methodology

In this paper, we discuss the identification and extraction of low-frequency keywords. The comments in the dataset are first segmented into sentences, trained by neural network model, clustered to generate the word structure of keywords, followed by word structure ranking, keyword extraction, and then the low-frequency keywords are ranked in the same phrase pattern according to the topic relevance of Meituan comments to achieve the low-frequency keywords we want to extract [21]. The specific framework is shown in Figure 1.

*3.1. Word Sense Structure Generation.* Word sense structure generation is based on the methods of word clustering or classification in natural language processing. The three following methods are commonly used: The first method is using external knowledge bases (e.g., WorldNet, HowNet, Cyc) to obtain semantic categories of words directly [22]. The disadvantage of this method is that the knowledge base is difficult to build and difficult to update. The second method is using classifiers in machine learning to identify the word classes of words. This method requires a certain number of datasets to be labeled and the classifier to be trained. This method is difficult to apply when there are many classes of words. The third method is using unsupervised clustering method. This method uses a large unlabeled dataset for training and automatically clusters words into different categories using contextual information of word occurrences. The clustering method is relatively weak, but the training data is easy to obtain and the number of word categories can be chosen flexibly.

We use a word clustering approach based on natural language processing, which maps individual words in a comment to a semantic vector space. In this space, the Eulerian distances of semantically similar words are also close to each other. The Eulerian distances are then used to cluster words that belong to the same word class and are semantically similar. Each word class is represented by a label, which represents the semantic meaning of the word class in the semantic space. Then, the semantic structure of the keywords is generated by replacing all the words in the candidate keywords with the labels. The specific representation is given by the following equation:

$$y(t) = g(v\, f(U\, w(t))), \tag{1}$$

where $w(t)$ and $y(t)$ denote the input and output layers, respectively, and $s(t) = f$ ($Uw(t$-RRB)) denotes the hidden layer.

*3.2. Lexical Structure Ordering.* In documents, the semantic structure has a high frequency of occurrence compared to low-frequency keywords and can be used to determine whether a semantic structure is valid or not [23]. The semantic structure of a keyword can be obtained by word structure generation, which indicates the usage pattern of the keyword. If the number of word clusters is $k$ and the allowed semantic structure length is $n$, the number of semantic structures of possible parameters is $k_n$.

The number of occurrences of low-frequency keywords is very low in all comments, and the contextual information is sparse. Each low-frequency keyword corresponds to a semantic structure containing many keywords. The ranking of the semantic structures can be done using various ranking methods. We mainly use the number of keywords corresponding to each semantic structure as the evaluation index.

*3.3. Keyword Sorting.* Because the contextual information of low-frequency keywords is sparse, it is difficult to use contextual information to rank different low-frequency keywords under a single lexical structure. We use the contextual information of each word in the document set to rank the low-frequency keywords. For example, in the review of Meituan, "the peanuts in this Meituan are delicious...

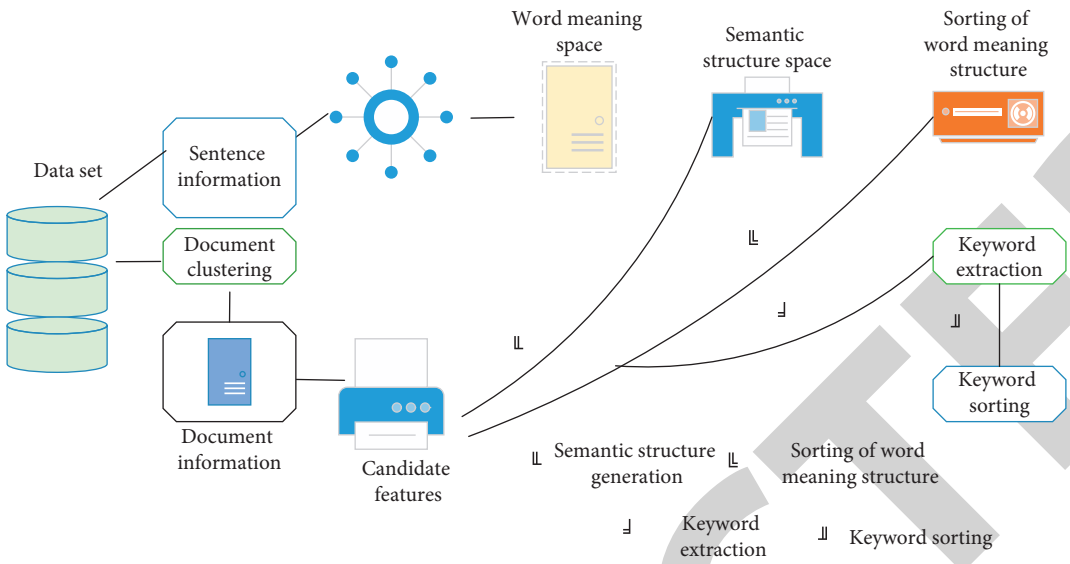


Figure 1: Framework of low-frequency keyword extraction.

and the milk tastes good." If "peanuts and milk" are a low-frequency keyword, the frequency of occurrence is low and the contextual information is sparse. However, the words "peanut" and "milk" appear more frequently in the document. Using contextual information of these words in the entire document set, the words can be ranked according to their relevance to the document topic. In order to rank the low-frequency keywords, we first generate a vector of $V_i$ keywords, which is given by the following equation:

$$V_i = \sum w_i \in P_i \frac{V_{w_i}}{\left\| V_{w_i} \right\|_1}, \tag{2}$$

where $P_i$ denotes the currently ranked keyword, $w_i$ denotes the words that form part of the keyword, and $V_{w_i}$ denotes the vector consisting of the contextual information (word features around which the word occurs multiple times) of word $w_i$ in the document set. Then, the rating of $V_i$ can be given by the following equation:

$$\text{Scoring}\left(V_i, V_t | V_b\right) = \frac{\left|V_i - V_b\right|^* \left|V_t - V_b\right|}{\left\|V_i - V_b\right|_2 \left|V_t - V_b\right\|_2}, \tag{3}$$

where $V_t$ is the word frequency vector produced by the manually selected document clusters after document clustering, indicating the topics related to the usefulness of the USM. $V_b$ shows the background vector generated from the word frequencies in the entire document set. The ranking of low-frequency keywords can be obtained by calculating the score of each keyword on vector $V_i$ separately.

*3.4. Commenting and User Model Building under Theme Space.* In the process of LDA topic modeling [24], the "document - topic" probability distribution matrix is obtained simultaneously, and we denote $\theta$ as Review_MAX$_{i\times k}$, with $i$ corresponding to the number of documents in the comment corpus and $K$ the number of topics. The row vector of Review_MAX$_{i\times k}$ is the description of the probability distribution of comment $r$ in the topic space, as in the following equation:

$$r \cdot \text{topic}_p\text{rofile} = \left[p_1, p_2, \ldots, p_k\right]. \tag{4}$$

The user model is also built on the hidden topic space. For this purpose, a set of product feature words Interest_set is used to describe the user's interest, from which the user selects the word items he/she cares about, and the algorithm maps the sequence of the selected word items to the hidden topic space. The modeling process is divided into 3 steps:

(i) Step 1: set Interest_set to generate user interest descriptions based on feature words selected by users.

(ii) Based on the LDA clustering results and the classification of cell phone features by e-commerce platform, the feature words describing the performance of cell phones are divided into 8 topics, namely, "screen effect, network signal, appearance design, photography, audio and video entertainment, operation performance, cost performance, and battery life," from which users select the features they are interested in. For example, if user $u$ is concerned about the "appearance" and "battery performance" of the cell phone, he selects a topic descriptor from the corresponding topic to characterize $u$, with u.feature_profile = { battery, battery life, appearance, appearance, screen, body, size, ...}. The canonical expression is in equation (5), where Topic($f$) corresponds to the set of topic words under the user's topic of interest, with mapping u.feature_profile to the LDA hidden topic space.

$$u.\text{feature}_p\text{rofile} = \left\{t_i | t_i \in \text{Topic}\left(f\right), f \in \text{Interest\_set}, i = 1, 2, \ldots, m\right\}. \tag{5}$$

(iii) Step 2: word vector representation of user interest.

(iv) A word vector is a distributed representation of words obtained based on shallow neural network learning by representing words as an N-dimensional high-density real vector, where the word items correspond to a point in the N-dimensional space and the spacing of the points reflects the potential semantic relationships between the word items. Before mapping user interests based on feature words to the topic space, the study introduces word vectors by first converting u.feature_profile into a word vector matrix $u \cdot \mathrm{vec}_M AX_{m\times v}$ for word vector dimensionality. The user interest model based on word vector description can convey the semantic meaning and improve the recommendation accuracy. The $u \cdot \mathrm{vec}_M AX_{m\times v}$ matrix representation also facilitates the mapping of the user model to the topic space, where the user interest and review models are based on the same topic space; that is, they can be regarded as two points in the space, and their correlation is directly calculated by the distance formula. The word vector introduced in the study is an open-source Chinese pretraining model of Beijing Normal University [25]. The training corpus of this word vector is "Baidu Encyclopedia" with a corpus size of 4.1 G and a vector space dimension of 300.

(v) Step 3: user interest model in topic space. Topic $t$ is expressed by the probability distribution of "topic - lexical items" generated by LDA clustering, as shown in the following equation:

$$t \cdot \mathrm{feature}_p\mathrm{rofile} = \{\langle f_i, w_i\rangle, \quad i = 1, 2, \ldots, n\}, \tag{6}$$

where $f_i$ are the feature words describing topic $t$, $w_i$ are the weight of $f_i$, and $n$ is the number of feature words. Correspondingly, the word vector matrix of topic $t$ is established as $t \cdot \mathrm{vecMAX}_{m\times v}$. Under the word vector space, the interest matrix of $u$ is multiplied with the transpose matrix of topic $t$, while incorporating the topic feature word weight matrix = $W_{n\times v} = [w_1, w_2, \ldots, w_n]^T$, and finally the maximum value of the matrix operation is taken as the semantic relevance of $u$ and $t$. The correlation of user $u$ with $K$ topics is calculated according to equation (7), and the user interest model under topic space is generated as shown in equation (8):

$$\mathrm{Sim}_1 = \mathrm{Max}\left(u \cdot \mathrm{vec\_MAV}_{m\times v} \times t \cdot \mathrm{vec\_MAX}_{n\times v}{}^T \times W_{n\times l}\right), \tag{7}$$

$$u \cdot \mathrm{topic}_p\mathrm{rofile\$} = [\mathrm{Sim}_1, \mathrm{Sim}_2, \ldots, \mathrm{Sim}_K]. \tag{8}$$

## 4. Experiment and Conclusion

*4.1. Experimental Data.* In this experiment, we extract data from Meituan, the largest merchant review site in China, which includes 23 areas such as restaurants, shopping centers, hotels, and travel [26]. The Meituan data contain 984,502 Meituan reviews and 584,762 non-Meituan reviews. We focus on the reviews related to Meituan in the Meituan dataset and classify them into two categories based on their usefulness: first, useful reviews, of which 449,437 reviews have a usefulness value > 0; second, useless reviews, of which 535,065 reviews have a usefulness value = 0.

*4.2. Experimental Procedure.* In this paper, we focus on three aspects: candidate word generation, phrase filtering, and phrase scoring. Finally, we verify the effectiveness of our experiments by determining the percentage of usefulness of the extracted low-frequency keywords in the comments and whether they are useful for users' selection and decision making. The following is a detailed introduction in three parts.

*4.3. Candidate Word Generation.* In modern generative linguistics, it is difficult to separate function words from content-related words. Our main work is to use function words as boundaries to form candidate words. The steps are as follows:

(1) In the document, each comment is first separated by a punctuation mark, such as { ,.; ! ? : } .

(2) The LIWC2015 dictionary contains 19,281 discontinued words, and we use the LIWC2015 dictionary to check for separating comments, and if they are in the dictionary, we use them as boundaries to generate candidate phrases [27].

(3) Generated candidate phrases are exported to obtain the candidate phrases of the whole corpus. In order to reduce the noise and complexity of the experiment, we check whether the above problems occur by using the lexicon dictionary (the word list of lexicon dictionary contains 67,725 words) and discard the candidate phrases directly if they are not in this list [28]. By using the above two screening steps, we end up with 1,078,414 phrases in the Meituan dataset, with 31,093,419 occurrences. The distribution of phrase types is shown in Figure 2.

A represents the whole corpus, B represents the useful data comments of Meituan, and C represents useless data comments of Meituan. The percentages of candidate phrases with more than 9 occurrences are 6.27%, 6.98%, and 7.49%, respectively, while the percentages of only 1 occurrence are 71.7%, 71.12%, and 70.01%, respectively. This shows that removing low-frequency phrases will lose a lot of useful information, which is not conducive to better text information extraction and the evaluation of the usefulness of the Mission's comments.

*4.4. Phrase Filter.* This experiment focuses on the usefulness of the reviews of Meituan. In order to verify that low-frequency keywords contain a lot of important information and great research significance, the three following processes will be used to filter the candidate phrases [29, 30]. (1) High-frequency words can increase the accuracy of the

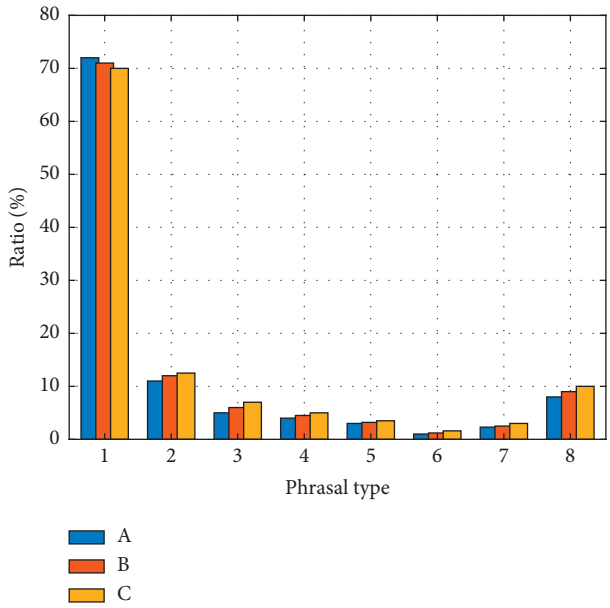


Figure 2: Distribution of phrase types.

representation. Therefore, in order to support word grouping, phrases with less than $N = 300$ word occurrences are removed. (2) In the experiment, to simplify the discussion, only the filtered comments containing phrases consisting of two words will be studied. (3) Since the goal of the experiment is to study low-frequency keywords, only phrases that occur once are discussed.

Through the above phrase filtering, there are 327, 345, 120, 828, 78, and 247 phrases left in A, B, and C datasets, respectively, and their percentages are 30.35%, 25.61%, and 23.58% respectively. The final filtering results are shown in Figure 3.

*4.5. Phrase Rating.* The phrase score is very important for the whole keyword extraction. Through the above phrase filtering, we finally obtained 199,075 Meituan phrases that only appeared once in the text and contained only two words [31]. The whole Meituan phrase database is represented by a distribution of trained words, and K-means clustering is performed; that is, according to the similarity principle, data objects with high similarity are classified into the same class clusters, and data objects with high dissimilarity are classified into different class clusters, where K represents the number of class clusters and means represents the mean value of data objects in the class clusters. The clusters are divided into 200 groups, and each group is identified by the label range of "C000–C199." In order to reduce the noise, reduce the processing difficulty, and achieve better classification effect, 20,277 useful phrases and 16,362 useless phrases of Meituan were generated by replacing the extracted keywords with word labels. Since we mainly focus on the usefulness of Meituan reviews, here, we only list the usefulness categories. The details are shown in Table 1.

C15 for fruit, C155 for sweets, C51 for flavor phrases, C63 for meat or cereals, C125 for emotional adverbs, C152 for price or affect adjectives, and C149 mostly for words that describe the environment.

In this paper, we collect 2013–2014 USG usefulness reviews, and, in order to rank low-frequency words with the same phrase pattern, we define a target vector $V_t$, which represents the textual topic relevance of the dataset, and the identification algorithm about low-frequency keywords is shown in Table 2.

*4.6. Experimental Conclusions.* From the experiment, we can get the distribution of usefulness comments of Meituan, so we can see that the usefulness votes with 5 or more occurrences only account for 6.08% of the whole Meituan comments, while those with 1 occurrences account for 52.78% of the whole usefulness votes. The low-frequency words are mostly words that objectively express the dining experience, such as "quite affordable, unforgettable, and very cold." The higher the "usefulness" vote is, the more valuable the review is and the more useful the phrases it contains; the high-frequency words are mostly words about Meituan entities, such as "steak salad, Meituan seats, cheese bread." The lower the "usefulness" vote, the lower the value of the comment and the more useless the phrases included. The distribution of "usefulness" votes is shown in Table 3.

This experiment not only shows that ignoring low-frequency keywords will lose a lot of important information but also verifies that our proposed method has made great progress in dealing with low-frequency keywords and has achieved good results in the restaurant usefulness poll, providing consumers with accurate and useful information in a more objective way.

Model parameter setting: for the LDA model, the value of the subject number K, which is related to $\alpha$ and $\beta$ of the model, is critical. K is used as the optimization parameter and the value is determined experimentally. Figure 4 shows the clustering effects of the three modeling schemes with different K values. Overall, with increasing K, Avg_similarity tends to decrease, indicating that the intertopic similarity decreases and the stability of the clustering structure increases. On the contrary, KL dispersion increases gradually, indicating that the intertopic differences are widened and the internal cohesion is increased. With increasing K, the two metrics gradually converge. Specifically for the three modeling schemes, both sets of indicators show that the clustering effect of "synonymous feature word normalization" is significantly better than that of "noun + verb" and "feature word." Therefore, the topic clustering scheme of "synonymous feature word normalization" was adopted in the subsequent experiments. According to the experimental results (see Figure 4), the clustering model is the best. KL scatter = 8.267, Avg_similarity = 0.05, and finally K = 13.

Clustering results: Figure 5 shows the clustering results generated by pyLDAvis for K = 13. On the whole, the themes are well distributed, and most of them are clearly distinguished, with a few overlapping (themes 4 and 5, themes 1 and 2). For this reason, the following treatment was performed: for each topic clustered, the topic words were ranked in descending order of probability, and the top 8

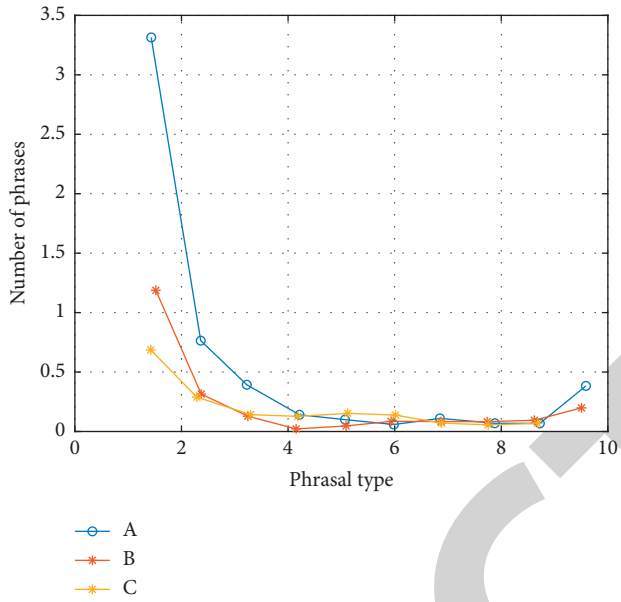


Figure 3: Phrase filtering distribution.

Table 1: Example of phrase grouping.

| Phrase | Give an example |
|---|---|
| C15–C155 | Cucumber frozen cake, grape mousse, cherry milkshake, peanut milk, peach crisp, almond milk |
| C155–C155 | Cream cookies, dessert custard, cotton mousse cheese pudding, cream sundae, walnut biscuits |
| C15–C51 | Pumpkin mustard, peanut seasoning, blackberry jam, fruity butter, cherry jam, strawberry jam |
| C63–C63 | Pork sausage, sausage Tujia, diced chicken with vegetables, beef fried rice, pineapple corn, sausage cheese |
| C129–C152 | Very cheap, very attractive, very bad, absolutely bad, full of taste, ridiculous |
| C129–C149 | Very quiet, comfortable, elegant, slightly high-grade, energetic, super luxurious and very cold |

Table 2: Model corresponding algorithm.

| |
|---|
| Input: a group of low-frequency phrases in the phrase pattern, all comments of the whole corpus |
| Output: low-frequency keyword sorting list: $L_0$ |
| 1) Divide the comments into restaurants and backgrounds |
| 2) Divide restaurant comments into usefulness and uselessness |
| 3) Generate target vector $V_t$ and background vector $V_h$ |
| 4) Perform the algorithm and calculate the scoring value |
| 5) Arrange $L$ in ascending order $L_0$ |

Table 3: Distribution of "usefulness" votes.

| Number of comments | Comment on "usefulness" low-frequency words/item | Number of comments | Comment on "usefulness" low-frequency words/item |
|---|---|---|---|
| 1 | 237225 | 6 | 7859 |
| 2 | 104578 | 7 | 5247 |
| 3 | 46587 | 8 | 3567 |
| 4 | 23458 | 9 | 2549 |
| 5 | 13256 | 10 | 1478 |

words were used to describe the topic semantics. If a word appears in more than one topic at the same time, it will be assigned to the topic with the highest weight value. For example, "battery capacity" appears in both topic 4 and topic 12, but the weight value under topic 12 (0.052) is higher than that under topic 4 (0.019), so it is placed under topic 12. Clustered subject terms were adjusted to better clarify the meaning of the topics. According to the list of topic words of each topic, the 13 topics were assigned to 9 feature categories of "operation performance, screen effect, network signal,

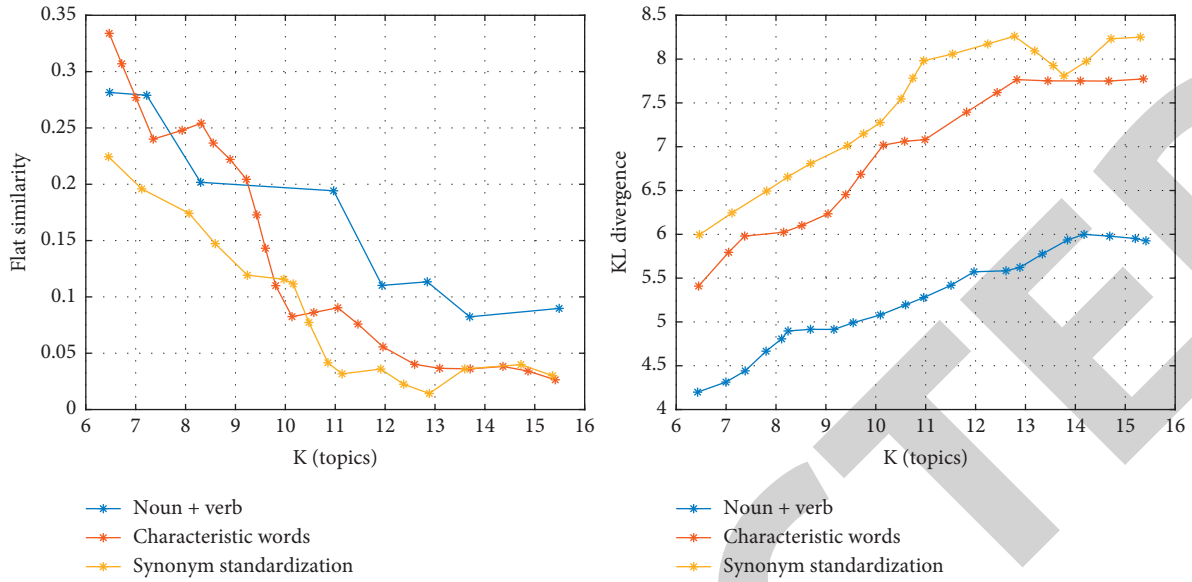


Figure 4: LDA topic clustering results with 3 feature modeling schemes (left: Avg_similarity; right: KL scatter in vertical coordinate).

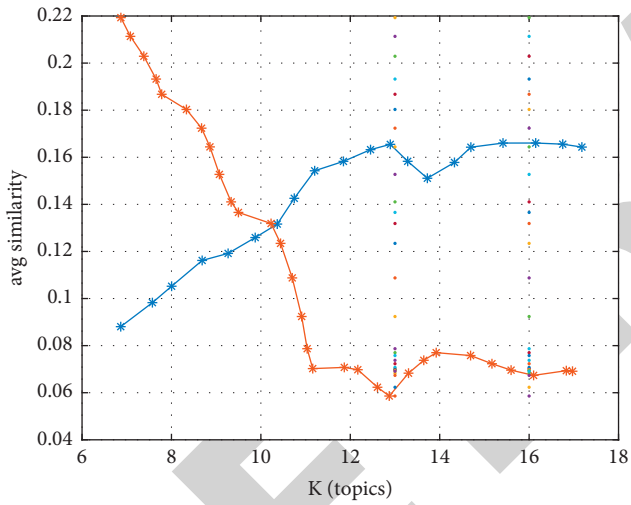


Figure 5: Effect of LDA topic clustering with different K values (Avg_similarity and KL scatter).

appearance design, photography, audio and video entertainment, cost performance, battery life, and others" by referring to the settings of cell phone feature indexes in digital websites, and the feature word set of user interest selection was generated accordingly, Interest_set, used for user modeling.

## 5. Conclusions

The study uses a probabilistic topic model to construct a user interest model in the topic space and incorporate it into the review perceived value calculation model, based on which a review recommendation strategy that integrates user interest and review utility is proposed, and the effectiveness of the recommendation strategy is tested by an online evaluation system. For the user model, the feature words characterizing user interest are treated with equal weights, but, during the testing process, it is found that users focus on product performance, and subsequent research can set weights for the feature words describing user interest to build a more refined user interest model.

The follow-up research is also prepared to introduce deep learning algorithms to explore user modeling in depth, extract user features from user comments, and improve the personalized recommendation algorithm.

## Data Availability

The datasets used in this paper are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest regarding this work.

## Acknowledgments

This work was supported by the Educational Reform Project of the Education Department of Guangdong Province named "Based on the securities crowdfunding circulation platform to promote the reform and practice of student entrepreneurship" under grant no. 55611173 and Open Foundation of Guangdong Provincial Key Laboratory of Public Finance and Taxation with Big Data Application under grant no. 202101.

## References

[1] S. Hido, S. Suzuki, R. Nishiyama et al., "Modeling patent quality: a system for large-scale patentability analysis using text mining," *Journal of Information Processing*, vol. 20, no. 3, pp. 655–666, 2012.

[2] A. C. Sönnichsen, N. Donner-Banzhoff, and E. Baum, "Chi-square-based scoring function for categorization of MEDLINE citations," *Methods of Information in Medicine*, vol. 49, no. 4, pp. 371–378, 2010.

[3] G. Kaur, M. Kumar, and M. Kumar, "A review on sentiment analysis of social media data using text mining and machine learning," *International Journal of Advanced Research*, vol. 4, no. 5, pp. 772–775, 2016.

[4] C. Zhang, T. Xie, K. Yang et al., "Positioning optimisation based on particle quality prediction in wireless sensor networks," *IET Networks*, vol. 8, no. 2, pp. 107–113, 2019.

[5] C. H. Cao, Y. N. Tang, D. Y. Huang, W. M. Gan, and C. J. Zhang, "IIBE: An Improved Identity-Based Encryption Algorithm for WSN Security," *Security and Communication Networks*, vol. 2021, Article ID 8527068, 8 pages, 2021.

[6] S. Zhou and T. Zhang, "Research on the construction of flipped classroom model for English teaching based on SPOC," *Revista de la Facultad de Ingenieria*, vol. 32, no. 14, pp. 267–273, 2017.

[7] F. Jiang and W. F. Mccomas, "Analysis of nature of science included in recent popular writing using text mining techniques," *Science & Education*, vol. 23, no. 9, pp. 1785–1809, 2014.

[8] V. Korde, "Information extraction for personalised services based on conference alerts," *International Journal of Data Mining, Modelling and Management*, vol. 8, no. 1, pp. 93–105, 2016.

[9] L. U. Wen-Hsiang, L. F. Chien, and H. J. Lee, "Anchor text mining for translation of Web queries: a transitive translation approach," *ACM Transactions on Information Systems*, vol. 22, no. 2, pp. 242–269, 2004.

[10] K. B. Cohen, L. E. Hunter, and F. Lewitter, "Chapter 16: text mining for translational bioinformatics," *PLoS Computational Biology*, vol. 9, no. 4, Article ID 1003044, 2013.

[11] D. Wu, C. Zhang, L. Ji, R. Ran, H. Wu, and Y. Xu, "Forest fire recognition based on feature extraction from multi-view images," *Traitement du Signal*, vol. 38, no. 3, pp. 775–783, 2021.

[12] R. Hubertrajan and J. P. M. Dhas, "A method for classification based on association rules using ontology in Web data," *International Journal of Computer Application*, vol. 49, no. 8, pp. 13–17, 2012.

[13] S. V. Ravi, B. D. Kumar, S. P. J. G. Raj, and A. Divyacharan, "Text mining methods for online topics and reviews using machine learning," *Journal of Physics: Conference Series*, vol. 1916, no. 1, Article ID 012215, 2021.

[14] M. Verma, "Lexical analysis of religious texts using text mining and machine learning tools," *International Journal of Computer Application*, vol. 168, no. 8, pp. 39–45, 2017.

[15] S. M. Mousavi, A. H. Alavi, A. H. Gandomi, and A. Mollahasani, "Nonlinear genetic-based simulation of soil shear strength parameters," *Journal of Earth System Science*, vol. 120, no. 6, pp. 1001–1022, 2011.

[16] S. E. Seker, O. Altun, U. Ayan, and C. Mert, "A novel string distance function based on most frequent K characters," *International Journal of Machine Learning and Computing*, vol. 4, no. 2, pp. 177–182, 2014.

[17] J. W. Son and S. B. Park, "Learning word sense disambiguation in biomedical text with difference between training and test distributions," *International Journal of Data Mining and Bioinformatics*, vol. 6, no. 2, p. 216, 2012.

[18] D. Yongping, J. Liu, L. Jingxuan, and G. Xuemei, "Hierarchy construction and text classification based on the relaxation strategy and least information model," *Expert Systems with Applications*, vol. 100, pp. 157–164, 2018.

[19] I. Hiroshi, "A designing model for the construction of e-based learning," *Computer Education*, vol. 10, pp. 14–20, 2001.

[20] Y. Liu, S. Chen, and B. Guan, "Layout optimization of oil-gas gathering and transportation system in constrained three-dimensional space," *Chinese Science Bulletin*, vol. 65, no. 9, pp. 834–846, 2020.

[21] Y. Liu, S. Chen, B. Guan, and P. Xu, "Layout optimization of large-scale oil-gas gathering system based on combined optimization strategy," *Neurocomputing*, vol. 332, no. 7, pp. 159–183, 2019.

[22] B. Guan, S. Chen, and Y. Liu, "Wave patterns of (2+1)-dimensional nonlinear Heisenberg ferromagnetic spin chains in the semiclassical limit," *Results in Physics*, vol. 16, Article ID 102834, 2019.

[23] T. Xie, C. Zhang, Z. Zhang, and K. Yang, "Utilizing active sensor nodes in smart environments for optimal communication coverage," *IEEE Access*, vol. 7, Article ID 11338, 2018.

[24] Z. Zhang, C. Zhang, M. Li, and T. Xie, "Target positioning based on particle centroid drift in large-scale WSNs," *IEEE Access*, vol. 8, Article ID 127709, 2020.

[25] L. Do-Yeop, Y. Cheol-Hwan, and P. Chan-Sik, "Development and application of failure-based learning conceptual model for construction education," *Journal of Construction Engineering and Project Management*, vol. 1, no. 2, pp. 11–17, 2011.

[26] A. Kumar, G. Vashishtha, C. P. Gandhi, H. Tang, and J. Xiang, "Tacho-less sparse CNN to detect defects in rotor-bearing systems at varying speed," *Engineering Applications of Artificial Intelligence*, vol. 104, Article ID 104401, 2021.

[27] J. Wei, X. Lu, and J. Dang, "A model-based learning process for modeling coarticulation of human speech(knowledge, information and creativity support system)," *–Transactions on Info and Systems*, vol. 90, no. 10, pp. 1582–1591, 2007.

[28] L. Wang, C. Zhang, Q. Chen et al., "A Communication Strategy of Proactive Nodes Based on Loop Theorem in Wireless Sensor Networks," in *Proceedings of the 2018 9th International Conference on Intelligent Control and Information Processing (ICICIP)*, pp. 160–167, IEEE, Wanzhou, China, November, 2018.

[29] H. Li, D. Zeng, L. Chen, Q. Chen, M. Wang, and C. Zhang, "Immune Multipath Reliable Transmission with Fault Tolerance in Wireless Sensor Networks," in *Proceedings of the International Conference on Bio-Inspired Computing: Theories and Applications*, pp. 513–517, Springer, Singapore, Octobar, 2016.

[30] G. Lang, Q. Li, M. Cai, and T. Yang, "Incremental approaches to knowledge reduction based on characteristic matrices," *International Journal of Machine Learning & Cybernetics*, vol. 8, no. 1, pp. 1–20, 2014.

[31] N. Z. Lima and R. C. Mesquita, "Point interpolation methods based on weakened-weak formulations," *Journal of Microwaves, Optoelectronics and Electromagnetic Applications*, vol. 12, no. 2, pp. 506–523, 2013.