# Constructing the Phrase Dictionary and Visualizing Consumer Behaviors in the Food Industry Based on Online Reviews during the COVID-19 Pandemic

Zongwei Li<sup>1</sup>, Xiaoling Tong<sup>1</sup>, Yanhui Zhang<sup>2</sup>\*

<sup>1</sup>School of Economics and Management, Shanghai Institute of Technology, Shanghai, China <sup>2</sup>Business School, East China University of Science and Technology, Shanghai, China \*Corresponding Author.

## Abstract

The COVID-19 pandemic affects food industry largely. In this study, the data of online reviews are collected from dianping.com during the outbreak and stable period of the COVID-19 pandemic in China, and the rules in combination with the statistical methods are adopted to train the dictionary of restaurant phrases. After the K-means algorithm is adopted to cluster the phrases in the dictionary, and the cluster class tags are defined, the co-occurrence analysis and the wordcloud analysis are conducted on the reviews. As indicated from the results, consumers pay attention to the three basic elements (i.e., services, environments and tastes), as well as to the social distance between people; Consumers who are more concerned about the pandemic situation raise higher requirements on environmental health issues than ordinary consumers, and place stress on the acquisition of security. As revealed from the mentioned results, restaurants should primarily take measures to maintain safe social distance between people and raise more rigorous environmental hygiene requirements on the environment. This method is served as a reference for the further online reviews analysis and provides implications for the management of the restaurants in COVID-19 pandemic period.

Keywords: COVID-19, dictionary construction, co-occurrence analysis, wordcloud analysis, food industry

## I. Introduction

The COVID-19 pandemic in 2019, inconsistent with natural disasters, accident disasters, social security and other types of emergencies, has a greater spreading speed and exhibits considerable harmfulness [1]. It imposes a great crisis on the entire restaurant industry, and restaurants in numerous countries and regions were forced to close in the early 2020 [2]. Since May 2020, the pandemic in China has been progressively controlled, and restaurants in various places have begun to reopen. As impacted by the threat of the virus and the strong desire for social communication simultaneously, people's consumption behaviors and attitudes start to shift.

To study the effect of the pandemic on the restaurant industry and consumer behaviors, Juan Palacios et al. [3] used the GPS information of the participants' smart phones to find that the more restaurant consumers around the participants, the greater their willingness to consume. Huang Jusong [4] analyzed the effect of the pandemic situation on restaurant stocks by using the restaurant business data and regression analysis method. Yu et al. [5] studies the main factors that affect the positive income of restaurant enterprises in the pandemic. However, the experiment has the disadvantages of high cost and taking long time, and the researches on the effect of pandemic situation on restaurant enterprises from the perspective of consumers are insufficient. Thus, we obtained massive online review information through a web crawler. The mentioned online reviews can help other consumers get the valuable information, and reflect consumers' reviews emotion the restaurant should be concerned with.

Text mining is capable of discovering potential knowledge, connections and patterns from considerable unstructured text information. On the whole, the text mining of online reviews complies with words [5-7]. However, the mentioned analyses are easily impacted by word frequency. To determine the topic of online reviews, the LDA model [8] and the manual recognition method are primarily used [9]. However, a problem of text sparsity is caused

when the LDA model is being applied in reviews, and a problem of a high subjectivity in the subjects is triggered when identified manually.

To address the mentioned problems, phrases have more complete meaning expression than words, so this study constructs the phrase dictionary by studying the changes and concerns of restaurant consumers in the pandemic period. First, the crawler technology is employed to collect the reviews from dianping.com website, and then the phrase dictionary of online reviews in the pandemic period is built by combining the statistics-based and rule-based methods. Lastly, the content and emotion of consumers are analyzed through the visualization in the pandemic period.

## **II. Related Research**

## 2.1 Text Mining

Text mining, i.e., text knowledge discovery, aims to analyze considerable text data and discover, extract and process the hidden knowledge [10,11]. In the era of big data, text mining is recognized as the vital research object of scholars. For text mining methods, Blei et al. [12] proposed an LDA model capable of recognizing text topics by complying with the Bayesian model and the principle of the probability theory. Mikolov [13] developed the Word2vec model, thereby significantly improving the efficiency of the word vector training. Goswami and Shishodia [14] studied how to exploit the fuzzy logic to aggregate documents in text mining. Specific to application, today's text mining has been extensively used in fields (e.g., biomedicine [15], tourism [16], social media [17], and technology education [18]). However, the existing text mining mostly employs vocabulary as the analysis unit, and the study on the phrase as the analysis unit remains insufficient. Compared with vocabulary, phrases have more complete semantics, and they are not easily affected by word frequency in text mining, thereby adversely affecting conclusions.

# 2.2 Phrase Mining

Phrase is considered a continuous lexical sequence in this study. It is a semantic unit formed in a specific context of the text. Compared with words, phrases are the semantic units formed in grammar and semantic structure. In contrast with words, phrases are more complete in grammar and semantic structure, and they cover more explicit semantic information, capable of expressing a specific topic more intuitively and accurately [19, 20].

Existing phrase mining methods largely fall to three categories, i.e., rule-based methods, statistics-based methods and hybrid methods. The first rule-based method predefines part of speech tagging rules and other rule templates through the scientific research of linguistics. However, the mentioned rule templates exhibit the poor applicability and portability [21, 22]. The second method based on statistics refers to extracting phrases by using related algorithms and machine learning [23]. However, such a method requires considerable corpora for training. The hybrid method aims to integrate the methods based on statistics and rules, which is the mainstream method of phrase mining and the method adopted here.

# 2.3 Text Visualization

Text visualization exploits graphic images or animation to express text information that enables people to efficiently acquire the key information covered in the text information and improves the efficiency of information processing [24]. The main study on text visualization involves visualization based on text contents and text relationships. On the whole, the visualization based on text contents addresses the problem of how to enable users to efficiently understand the subject of the text. The most typical form of visualization refers to a tag cloud [25]. After the tag cloud is laid out with beautiful and complex rules [26], it tends to be the present common wordcloud. The wordcloud is capable of setting the font size and color according to the frequency of words appearing. Visualization based on text relationships is adopted to help people understand text contents and discover knowledge. The common form of

visualization is the co-occurrence analysis. Furthermore, the co-occurrence analysis graph is capable of discovering potential knowledge and the correlation characteristics between knowledge in the text [27-29].

## III. The Process of Constructing the Phrase Dictionary and the Visual Analysis

To build a dictionary of restaurant phrases to gain insights into the demands and changes of users, this study divides the main analysis into three aspects, i.e., data collection, phrase mining and visual analysis (Fig. 1). Overall, the data collection module adopts a web crawler to collect online review data. The phrase dictionary building module largely employs the combination of statistics-based and rule-based methods to build a reusable phrase dictionary of online reviews in the pandemic period. The visual analysis module conducts the cluster analysis, the co-occurrence analysis and the wordcloud analysis, and finally draws conclusions to provide the decision support for restaurants.



Fig 1: process of constructing phrase dictionary and visual analysis

## **IV.** Construction of the Phrase Dictionary

After online reviews are obtained, the review collection is mined to build a phrase dictionary. First, the online reviews are grabbed by writing a python crawler. Subsequently, the adjacent phrase mining method is adopted to extract candidate phrases. The combination of the statistics-based and rule-based methods is employed to filter candidate phrases. Since the language of online reviews is not standardized, a more standardized phrase dictionary should be built through the artificial judgment and selection.

### 4.1 Data Source

Dianping.com refers to a leading local life information platform in China and one of the earliest independent third-party consumer comment websites worldwide. It offers users business information, consumer reviews and discounts information. In this study, a web crawler is adopted to scrap the comment information from December 2019 to December 2020 from the public comment website (e.g., 2324 restaurants and 317478 reviews).

After the crawler is used to get online reviews, this study processes the reviews to get the phrase database of restaurant reviews during the pandemic.

### 4.2 Text Preprocessing

Before the restaurant phrase dictionary is built, this study first preprocesses the reviews. Chinese punctuation, facial expressions and numbers are removed, while only Chinese characters are retained. Next, based on the stop words of Harbin Institute of technology, the stop words are removed, and the comment information is segmented. Besides, part of the speech is annotated by Jieba word segmentation tool. After the text preprocessing, the text corpus of the online reviews is obtained.

4.3 Candidate Phrase Extraction ISSN: 0010-8189 © CONVERTER 2020 www.converter-magazine.info To extract candidate phrases, this study selects the adjacent phrase mining method. According to segmentation sequence  $T = \{t_0, t_1, t_2, ..., t_n\}$ , the phrase  $P_0$  is generated by integrating  $t_0$  and  $t_1$ , and phrase  $P_1$  is generated by combining  $t_0$ ,  $t_1$  and  $t_2$ . Likewise, this study creates  $P_2$ ,  $P_3$  to  $P_n$ . When the phrase set  $D = \{P_0, P_1, P_2, ..., P_n\}$  generated by the document M is being formed, the number of occurrences of the generated phrase  $P_1$  is  $C_1$  and others, the number of occurrences of  $P_n$  is  $C_{tn}$ . Next, the score of Eq. (1) acts as the measure of the optimal match [30]. Eq. (1) expresses a generalized form of statistical method:

$$\mathbf{E} = (1 - \mu)\mathbf{C}_{pn} + \frac{\sigma \mathbf{C}_{pn}}{(1 - \mu)\mathbf{C}_{tn}}$$
(1)

Among Eq. (1),  $\mu$  and  $\sigma$  are manually defined parameters, which are phrases of words and phrases to set offset weights. Whether the phrase P is a high-frequency phrase is determined by the phrase importance *E*. when *E* exceeds 1 (*E* > 1), it is added to the candidate dictionary *D*. The algorithm is built as the phrase is generally composed of adjacent words, and it is a regression model. Accordingly, in the form of left to right combination, all possible phrase sets can be enumerated in a sentence. Besides, in the task of building a dictionary, the selection of high-frequency words acts as a vital factor. By Eq. (1), the correctness and collocation of phrase combination can be calculated.

#### 4.4 Candidate Phrase Filtering Based on Rules

After phrases are extracted by mining adjacent phrases, there are numerous inferior phrases. Accordingly, this study filters the part of speech combination of phrases before constructing the phrase dictionary. The following eight rules of phrase part of speech combinations are adopted:

- 1. adjective + noun
- 2. adnoun + noun
- 3. verb + noun
- 4. noun + verb
- 5. locative noun + noun
- 6. noun + noun
- 7. noun + noun + noun
- 8. noun + gerund + noun

#### 4.5 Candidate Phrase Filtering Based on Statistics

Information entropy is adopted to express the concept of information uncertainty. The higher the information uncertainty, the greater the information entropy will be. The left and right information entropy of the target phrase can examine the uncertainty of the left and right boundaries of the target phrase. The larger the left and right information entropy of the target phrase, the more uncertain its left and right boundaries will be, and it is more likely a complete phrase [31]. The equations of left information entropy and right information entropy are expressed as:

$$LE(s) = -\sum_{l \in L} P(ls|s) \log_2 P(ls|s)$$
(2)

$$LE(s) = -\sum_{l \in L} P(ls|s) \log_2 P(ls|s)$$
(3)

Where S denotes the target phrase; L and R represent the left and right adjacents of the target phrase, respectively; L is the set of all the left adjacents of S; R expresses the set of all the right adjacents of S; P(ls|s) is the conditional probability of co-occurrence of L and s when s appears; P(rs|s) is the conditional probability of co-occurrence of R and s if s appears. The larger the values of LE(s) and RE(s), the more uncertain the left adjacents of the candidate phrase will be, and the more probably s will be a complete phrase.

Next, the C-value of the phrase to filter is calculated [32]. Before a phrase is included into a dictionary, whether the phrase is in a longer phrase should be considered. If so, which phrase to keep is considered. The calculation of C-value largely complies with the following logic, i.e., the C-value of a candidate phrase is directly proportional to

its word frequency in the domain corpus, and inversely proportional to its length. If the candidate phrase is nested, its weight will decrease. The calculation equation of C-value is expressed as:

$$C - \text{value}(A) = \begin{cases} \log_2 |A| * f(A), A \text{ is not nested} \\ \log_2 |A| * f(A) - \frac{1}{C(A)} \sum_{i=1}^{C(A)} f(b_i), \text{ others} \end{cases}$$
(4)

A denotes the target phrase; |A| represents the length of the candidate phrase; f(A) expresses the frequency of the candidate phrase in the corpus;  $b_i$  is the candidate phrase nested A; C(A) represents the number of candidate phrases nested A. C-value is easy to use, exhibiting a high applicability and a significant phrase screening effect.

To obtain high-quality phrases and improve the reusability of the phrase database of online reviews, the candidate phrases should be screened again through the artificial judgment. The irrelevant phrases and wrong phrases were removed to form a reusable phrase library of food and beverage reviews in the pandemic period. Lastly, a dictionary of 3433 phrases is generated.

## V. Visual Analysis

### 5.1 Cluster Analysis

The established phrase lexicon is taken as the word segmentation dictionary to process all the reviews again; then the word vector model is used to train the segmentation results, so a word can be represented by a vector. Next, the word vectors corresponding to phrases are clustered by using K-means algorithm. The number of K-means clustering clusters is defined as 10 by calculating the sum of square error; after clustering, each class is labeled according to its phrase description. The sample phrases in the respective cluster are listed in Table 1:

	Cluster label	Examples of phrases in clusters	
1	Taste and evaluation of dishes	dishes and ingredients, the seasoning is good, full weight, clear grains, good taste	
2	Service attitude and measures	good service, self-service, service attitude, friendly service, reception at the door	
3	Customer experience	feeling good, very bad, good impression, disappointment, surprise	
4	Environmental location	not hard to find, the atmosphere is good, the environment is good, the traffic is convenient, the air conditioner is on	
5	Queuing	no need to line up, queue ended fast, very popular, the queue is very strong, making an appointment and waiting in line	
6	Business Booking	picking up the number, book in advance, no need to make an appointment, business is very good, overcrowded	
7	promotion tactics	half price coupons, super affordable, package coupons, package discounts, discount activities	
8	Taking out	packing and going back home, taking out platform, ordering takeout, packing box, taking away	
9	cost performance	cheap and affordable, low cost performance, average cost performance, especially expensive, consumption per person	
10	Pandemic prevention and control	wearing a mask, taking preventive measures, sitting apart, hand gel, body temperature registration	

Table 1	phrase of	clustering	results
---------	-----------	------------	---------

As indicated from Table 1, the main types of food comment phrases in the pandemic period consist of "dish taste and evaluation", "service attitude and measures", "customer experience", "environmental location", "queuing situation", "business reservation", "promotion tactics", "taking out", "cost performance", as well as "pandemic prevention and control". The mentioned phrases describe the specific situation of a certain type of tags, and contain more complete

semantic information than a single word, which can improve the readability and accuracy of the visual analysis results.

## 5.2 Co-occurrence Analysis

After the synonyms of phrases in phrase knowledge base are merged, python language is used to write code to create phrase co-occurrence matrix (top 500 co-occurrence frequency), and social network analysis tool Gephi is adopted to mine and visualize phrases. The graph of co-occurrence reveals the co-occurrence network relationship of high-frequency phrases: the larger the node, the higher the degree of centrality will be, thereby indicating that the more frequent it appears in online reviews, the more words it will be connected with. The connecting lines of nodes indicate the co-occurrence relationship between phrases; the thicker the lines are, the closer the feature relationship will be. The knowledge element with large nodes and the thick lines refer to the main reflection of online users' perception after consumption in offline restaurants, which directly reflects the changes of consumers' concerns and attitudes in the pandemic period. The results of the co-occurrence analysis are illustrated in Fig. 2:



Fig 2: visualization of phrase co-occurrence analysis

Through the analysis of the network structure of the picture, it shows that:

1. As indicated from the high centrality of "good service", "service attitude", "fresh and delicious", "cost-effective", "good environment", "serving speed" and others, users will largely concentrate on services, tastes, environments and prices when consuming food in the pandemic period. Moreover, the mentioned words are closely connected.

2. "Pandemic period" and "no queuing" are closely connected, which reflects the effect of the pandemic on the business of restaurant stores, while revealing consumers' reluctance to have considerable contact with people in the queuing process in the pandemic period.

3. Pandemic period" is connected with phrases (e.g., "packing and going back home", "wearing a mask", "siting apart" and "hand sanitizer gel"), which plays a guiding role in the service mode of businesses in the pandemic period. To be specific, the co-occurrence analysis of phrases in "pandemic prevention measures" is presented in Fig. 3. It includes "during the pandemic period", "after the pandemic", "pandemic prevention measures", "wearing masks", "separate seats", "showing health code", "body temperature registration", "hand sanitizer gel" and "chopsticks and

spoon". To be specific, "during the pandemic" and "wearing masks" are closely linked, and there is a tight link between "hand gel" and "body temperature registration".



Fig 3: visualization of "epidemic prevention measures" cluster

### 5.3 Wordcloud Analysis

Furthermore, consumers fall to two groups: the people more concerned about the pandemic situation, as well as the general population. The criterion for distinguishing the two groups is presented below. If a user's reviews involve phrases in the pandemic prevention and control category, the user is considered more concerned with the effect of the pandemic during the meal. Subsequently, the differences between the two types of user reviews are compared to gain insights into the key points that restaurant should stress in the pandemic period. Fig. 4 illustrates the wordcloud of reviews that are more concerned about the pandemic situation than general user reviews.



Fig 4: cloud chart of epidemic situation concerned Vs ordinary user reviews

The most common phrases in the wordcloud of general user reviews consist of "delicious", "service attitude", "cost-effective", "no need to queue up", "decoration style", "instant entrance" and "will come again". For consumers more concerned with the COVID-19 pandemic situation, the wordcloud covers considerable phrases (e.g., "no waiting", "clean and tidy", "order takeout", "retaliatory consumption", "don't worry", "cost-effective", "every other table" and "retaliatory consumption"). Consistent with the general users, the users more concerned with the pandemic also highly care for the service attitude and cost performance. However, consumers with the concern of the pandemic situation pay more attention to the following: focus on the hygiene of restaurant environment; require the pandemic prevention measures being taken by restaurant enterprises in the pandemic period to gain a sense of security; keep the social distance between people.

# VI. Conclusion

The mining and analysis of online reviews on food and beverage in pandemic period can help understand consumers' consumption. Given the defects of conventional text mining based on the vocabulary granularity and the LDA model, this study selects the phrase granularity to build the phrase database in the pandemic period. A reusable phrase dictionary is built by combining statistics-based and rule-based methods. Subsequently, the cluster analysis is conducted on the phrase dictionary. Thus, the co-occurrence analysis and the wordcloud analysis are conducted to lays a decision-making basis for restaurants.

As revealed from the visualization results, in the pandemic period, the public concentrate on the restaurant service environment, as well as the social distance between people. Compared with other consumers, consumers more concerned with the pandemic situation will care about the sanitation of the restaurant environment, and highlight the pandemic prevention measures of restaurant enterprises to obtain a sense of security.

The subsequent research will primarily address the following two problems. Before online reviews act as the corpus to train the phrase dictionary, false reviews can be distinguished and removed to obtain a higher accuracy. The dictionary training method is optimized to reduce the workflow of the manual judgment, which can underpin the extensive establishment and application of the phrase dictionary in a wide range of fields.

### Acknowledgements

The authors acknowledge the National Social Science Fund of China (Grant: 18BGL093), the Shanghai Pujiang Program (Grant: 2019PJC096).

#### References

- [1] S. Gautam, L. Hens, "COVID-19: Impact by and on the environment, health and economy," In: Springer. 2020
- [2] K. Severson, D. Yaffe-Bellany, "Independent restaurants brace for the unknown," The New York Times, pp. 21, 2020.
- [3] Palacios, Juan, et al., "Encouraging the resumption of economic activity after COVID-19: Evidence from a large scale field experiment in China," 2021.
- [4] Song, Hyoung Ju, Jihwan Yeon, Seoki Lee, "Impact of the COVID-19 pandemic: Evidence from the US restaurant industry," International Journal of Hospitality Management, 2021
- [5] B. Yu, J. Zhou, Y. Zhang, et al., "Identifying restaurant features via sentiment analysis on yelp reviews," arXiv preprint arXiv:1709. 08698, 2017.
- [6] A. Salinca, "Business reviews classification using sentiment analysis," 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), IEEE, vol. 2015, pp. 247-250.
- [7] M. Fan, M. Khademi, "Predicting a business star in yelp from its reviews text alone," arXiv preprint arXiv: 1401, 0864, 2014.
- [8] T., G. M., G.P. Marcelo, "The impact of COVID-19 on the Brazilian food service industry: topic modelling of online reviews," CLAV, 2020.
- [9] Y. Tausczik, K. Faasse, J.W. Pennebaker, et al., "Public anxiety and information seeking following the H1N1 outbreak: blogs, newspaper articles, and Wikipedia visits," Health communication, vol. 27, no. 2, pp. 179-185, 2012.
- [10] Y. Kodratoff, "Knowledge discovery in texts: a definition, and applications. International Symposium on Methodologies for Intelligent Systems," Springer, Berlin, Heidelberg, vol. 1999, pp. 16-29.
- [11] M. A. Hearst, "Untangling text data mining," Proceedings of the 37th Annual meeting of the Association for Computational Linguistics, vol. 1999, pp. 3-10.

- [12] D.M Blei, A.Y Ng, M.I Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.
- [13] T. Mikolov, K. Chen, G. Corrado, et al., "Efficient estimation of word representations in vector space," arXiv preprint arXiv: 1301, 3781, 2013.
- [14] S. Goswami, M.S. Shishodia, "A fuzzy based approach to text mining and document clustering," arXiv preprint arXiv: 1306. 4633, 2013.
- [15] M. Gou, Z. Wenlong, "Knowledge Mapping Analysis on Text Mining Research of Medicine Related Fields in Different Regions," Cross-Cultural Communication, vol. 13, no. 9, pp.1-9, 2017.
- [16] M. Maghrebi, A. Abbasi, T.H. Rashidi, et al., "Complementing travel diary surveys with twitter data: application of text mining techniques on activity location, type and time," 2015 IEEE 18th international conference on intelligent transportation systems, IEEE, vol. 2015, pp. 208-213.
- [17] S.A. Salloum, M. Al-Emran, K. Shaalan, "Mining social media text: extracting knowledge from Facebook," International Journal of Computing and Digital Systems, vol. 6, no. 02, pp. 73-81, 2017.
- [18] D. Shin, J. Shim, "A systematic review on data mining for mathematics and science education," International Journal of Science and Mathematics Education, vol. 19, pp. 639-659, 2021.
- [19] K.W. Church, "A stochastic parts program and noun phrase parser for unrestricted text," International Conference on Acoustics, Speech, and Signal Processing, IEEE, vol. 1989, pp. 695-698.
- [20] A. Nandi, H.V. Jagadish, "Effective phrase prediction," Proceedings of the 33rd international conference on Very large data bases, vol. 2007, pp. 219-230.
- [21] T. Koo, X. Carreras, M. Collins, "Simple semi-supervised dependency parsing," Proceedings of ACL-08: HLT, vol. 2008, pp. 595-603.
- [22] S. Quiniou, P. Cellier, T. Charnois, et al., "What about sequential data mining techniques to identify linguistic patterns for stylistics?" International Conference on Intelligent Text Processing and Computational Linguistics, Springer, Berlin, Heidelberg, vol. 2012, pp. 166-177.
- [23] Y. Lu, M. Castellanos, U. Dayal, et al., "Automatic construction of a context-aware sentiment lexicon: an optimization approach," Proceedings of the 20th international conference on World wide web, vol. 2011, pp. 347-356.
- [24] Cao, N., Cui, W., "Introduction to text visualization," 2016
- [25] M. Wattenberg, F. Viegas, "Tag clouds and the case for vernacular visualization," Interactions, vol. 15, pp. 49-52, 2008.
- [26] F.B. Viegas, M. Wattenberg, J. Feinberg, "Participatory visualization with wordle," IEEE transactions on visualization and computer graphics, vol. 15, no. 6, pp. 1137-1144, 2009.
- [27] S.C. Weller, "Shared knowledge, intracultural variation, and knowledge aggregation," American Behavioral Scientist, vol. 31, no. 2, pp. 178-193, 1987.
- [28] Z. Xueqin, "Research Status of Knowledge Aggregation and Service and Recommendations for Future Research," Information Studies: Theory & Application, vol. 2015, pp. 02.
- [29] C. List, "Group knowledge and group rationality: a judgment aggregation perspective," Episteme, vol. 2, no. 1, pp. 25-38, 2005.
- [30] K. Church, W. Gale, P. Hanks, et al., "Using statistics in lexical analysis," Lexical acquisition: exploiting on-line resources to build a lexicon. Psychology Press, vol. 2021, pp. 115-164.
- [31] H. Jin, K.F. Wong, "A chinese dictionary construction algorithm for information retrieval," ACM Transactions on Asian Language Information Processing (TALIP), vol. 1, no. 4, pp. 281-296, 2002.
- [32] K. Frantzi, S. Ananiadou, H. Mima, "Automatic recognition of multi-word terms: the c-value/nc-value method," International journal on digital libraries, vol. 3, no. 2, pp. 115-130, 2000.