Big Data Research 24 (2021) 100193

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

# Information Needs Mining of COVID-19 in Chinese Online Health Communities ☆

Jie Wang [a,b,*], Lei Wang [a], Jing Xu [c], Yan Peng [a]

[a] School of management, Capital Normal University, Beijing 100056, China
[b] State key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[c] Shenzhen International Graduate School, Tsinghua University, Shenzhen, Guangdong, 518055, China

ARTICLE INFO

*Article history:*
Received 27 May 2020
Received in revised form 30 October 2020
Accepted 31 December 2020
Available online 8 January 2021

*Keywords:*
Information needs
Topic mining
Online health communities
Lexical meaning co-occurrence
COVID-19

ABSTRACT

This study explores the information needs for the novel coronavirus pneumonia (COVID-19) in Chinese online health communities (OHCs). Based on the question and answer data about COVID-19 in six Chinese OHCs, topic mining and data analysis were conducted. We propose a CL-LDA topic model (Latent Dirichlet Allocation Model with co-occurrence of lexical meaning) based on lexical meaning co-occurrence analysis and LDA topic model. Four main information need topics and their proportion are found in this study, including symptom (45.50%), prevention (36.11%), inspection (10.97%), and treatment (7.42%). We also discover that men are most concerned about symptom information while women are most concerned about prevention information; young users have the largest proportion of information needs, and they are most concerned about prevention information. Experiment results show that the CL-LDA model can well adapt to the topic mining task of short text which is semantic sparse and lacking co-occurrence information in OHCs. The research results are helpful for OHCs to provide accurate information assistance and improve service quality.



## 1. Introduction

The outbreak of novel coronavirus pneumonia (COVID-19) has seriously affected people's health and normal life [1]. On January 30, 2020, the outbreak was recognized as a public health emergency of international concern (PHEIC) by the World Health Organization (WHO). In the process of prevention and control of the epidemic, the Chinese online health communities (OHCs) have played the unique advantages of online and remote consultation, reduced the risk of patients' infection, alleviated the situation of medical resources shortage in some areas, solved the plight of patients without medical treatment caused by the closure of nonemergency outpatient clinics in some hospitals. Compared with the same period, the number of the diagnosis and consultation volume of some third-party Internet medical service platforms increased more than 20 times, and the prescriptions increased nearly ten times [2]. However, the information provided by OHCs is mixed, and it isn't easy to obtain the required high-quality health information quickly and accurately. Therefore, it is very important to determine the health information needs of OHCs users and promote OHCs to provide more high-quality services.

In this study, topic mining and data analysis are conducted to discover the information needs of COVID-19 in Chinese OHCs. The main contributions of this paper are summarized as follows.

- To deal with the short, poorly prescriptive and domain-specific question and answer (Q&A) data from OHCs, a novel CL-LDA topic model is proposed. The CL-LDA model takes full account of the influence of synonyms on topic generation in the corpus, merging synonyms into one lexical meaning element, and replaces the original document word matrix of LDA with the co-occurrence matrix as the feature vector of the corpus. This model can be applied to other topic extraction tasks based on short texts from Internet.
- Based on real data from six Chinese OHCs, the information needs topic distribution, trend changes and impact factors of COVID-19 in Chinese OHCs are discovered. The research findings are of great practical value for OHCs to provide personalized information service and improve the service quality.

☆ This work was supported by Open Foundation of State key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications) (SKLNST-2020-1-19), Natural Science Foundation of Beijing Municipality (1202020) and Social Science Foundation of Beijing Municipal Education Commission (SM201910028017).

* Corresponding author at: School of management, Capital Normal University, Beijing 100089, China.
*E-mail address:* wangjie@cnu.edu.cn (J. Wang).

https://doi.org/10.1016/j.bdr.2021.100193
2214-5796/

The rest of this paper is organized as follows. Section 2 presents the related works and theoretical foundation. Section 3 discusses the research method of this study, including our research framework and key steps. Section 4 presents the experiment and results analysis. Section 5 discusses the principal findings. Finally, Section 6 concludes the paper.

## 2. Related literature and theoretical foundation

### 2.1. Research on information needs in OHCs

Health information generally refers to the information related to people's physical and mental health, disease, health preservation, etc., which can guide clinical and healthy behavior [3]. OHCs have become one of the most important sources for searching and exchanging health-related information, experiences, advice, support and opinions [4]. Scholars use interviews, questionnaires, content analysis and text mining to study the health information needs of users in OHCs from different perspectives.

Prabha et al. [5] used text mining technology to analyze the pregnancy data of MedHelp in OHCs, and classified the adopted and unused answers with SVM-RBF algorithm. Liu et al. [6] combined the LDA model and sentiment analysis to study the differences in topics and emotions of patients with physiological and psychological diseases in OHCs. Xu et al. [7] explored the health information needs of middle-aged and elderly users in OHCs by qualitative and quantitative analysis method. Jin et al. [8] studied the health information needs of elderly users in Yahoo Answers, the content of diabetes-related Q&A data was encoded to extract text features, and then the text was clustered through multi-dimensional scale analysis. Similarly, Oh et al. [9] collected 81,434 cancer questions from Yahoo Answers, and identified health-related topics by text mining to find the users' multi-dimensional cancer information needs. Zhou et al. [10–12] have studied health related data of OHCs from many aspects. Based on real data crawled from OHC, an integrated deep neural network (DNN)-based learning model [10] was proposed to analyze and describe the latent behavioral influence hidden across multiple modalities. A semi-supervised learning framework [11] was introduced to incorporate massive amount of unlabeled data with small portion of labeled data to enhance health activity recognition. Besides, an integrated CNN-RNN framework [12] was designed to model and analysis patient-physician-generated data, and an intelligent recommendation mechanism was then developed to provide patients with automatic clinic guide and pre-diagnosis suggestions in a data-driven way. Using k-means algorithm to cluster the hypertension related data, Tang et al. [13] found that the information needs topics of hypertension in OHCs included diagnosis, treatment and complications, while life treatment was the most concerned. Lu et al. [14] used the latent semantic index model and MapReduce distributed text clustering technology to mine the user's information demand of the tumor in OHCs, and mined five information needs topics and their proportions. However, most previous studies had limited considerations of the information need mining of sudden outbreak infectious diseases in OHCs.

### 2.2. LDA topic model

After adding the prior distribution of Dirichlet based on the probabilistic latent semantic analysis (PLSA), BLEI et al. [15] proposed the Latent Dirichlet allocation (LDA) model, which is very effective for identifying the potential topic information in large-scale documents or corpus. LDA is a three-layer Bayesian probability model based on "document, topic and word". In LDA, each document is a mixed Dirichlet distribution of potential topics, and each potential topic is a probability distribution of words. The model is shown in Fig. 1, and the meaning of each symbol is shown in Table 1.

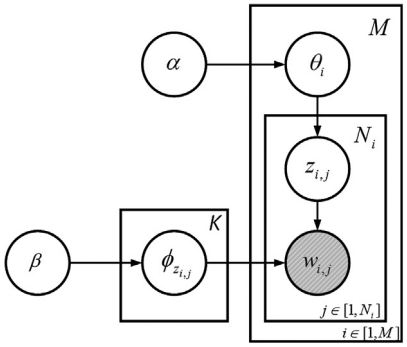


**Fig. 1.** Graph representation of LDA model.

**Table 1**
Meaning of symbols in LDA model.

| Parameter | describe |
|---|---|
| $M$ | Total documents |
| $N_i$ | Total number of words in the documents $i$ |
| $K$ | Number of topics |
| $\alpha$ | Prior distribution of topic distribution |
| $\beta$ | Prior distribution of word distribution |
| $\theta_i$ | Topic probability distribution of documents $i$ |
| $\varphi_{z_{i,j}}$ | Topic - probability distribution of words |
| $z_i, j$ | Topic $j$ of the first word of the document $i$ |
| $w_i, j$ | The $j$ word of the document $i$ |

In the LDA topic model, a document is generated as follows: firstly, the topic distribution of document $i$ is generated by sampling from the Dirichlet prior distribution $\alpha$ of the topic distribution of document $i$; secondly, the topic $z_i, j$ of the $j$ word in a document $i$ is generated by sampling from the polynomial distribution $\theta_i$, and then the corresponding word distribution $\varphi_{z_{i,j}}$ of topic $z_i, j$ is generated by sampling from the Dirichlet prior distribution $\beta$; finally, the word $w_i, j$ is generated from the polynomial distribution $\varphi_{z_{i,j}}$.

The joint distribution formula of all variables in LDA is:

$$p(w_i, z_i|\alpha, \beta) \sum_{j=1}^{N_i} p(w_{i,j}|z_{i,j}, \beta) p(z_{i,j}|\alpha) \tag{1}$$

Where, $p(w_{i,j}|z_{i,j})$ is the probability of sampling words under the topic, and the probability distribution formula of words in the document $i$ is:

$$p(w_{i,j}) = \sum_{j=1}^{N} p(w_{i,j}|z_{i,j}) p(w_{w_{i,j}}) \tag{2}$$

## 3. Methods

### 3.1. Research framework

Our research framework has four main steps, as shown in Fig. 2. Firstly, we collect the Q&A data related to COVID-19 in OHCs and preprocess the data, including Chinese word segmentation, part of speech tagging and removing stop words. Secondly, the OHCs synonym list based on "Expanded Version of Synonymy Thesaurus" [16] is built, and the standard topic lexical meaning corpus is produced. Thirdly, we generate the lexical meaning co-occurrence matrix of the lexical meaning corpus and carry out topic mining based on the CL-LDA model of lexical meaning co-occurrence, the text topic clustering results are obtained. In the last step, after formulating the topic coding rules, the top 10 keywords under each topic

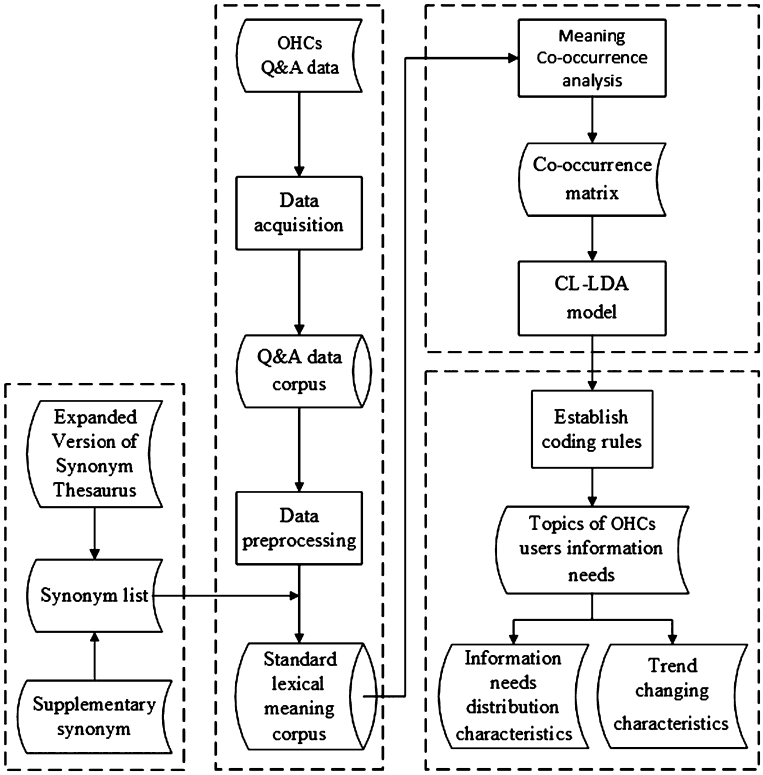


**Fig. 2.** Research framework.

category are extracted according to the topic clustering results, and then the information needs distribution and trend changing characteristics are output.

### 3.2. The generation of synonym list

#### 3.2.1. The influence of synonyms on topic lexical meaning aggregation

In LDA model, the co-occurrence probability between words is the only criterion to measure the words' similarity. LDA model does not consider the lexical meaning and grammatical relations between words, which will bring some problems for mining the Q&A data of OHCs. In the Q&A data corpus of OHCs, there are words with the same or similar meanings. However, due to the differences in the prevalence of words, users' health literacy level and word usage habits, there may be multiple ways for different users to express the same word meaning, which will results in large differences in the word and lexical meaning probability space of the corpus. For example, in topic $T$, the lexical meaning $M$ has two words expressions (word $m_1$ and $m_2$), while N has one word expression (word $n$). In topic T, the weight of $n$ is 0.045, while the weight of $m_1$ and $m_2$ is 0.025 each. In this case, words $m_1$ and $m_2$ have higher probability to be ignored than word $n$ because of their back weight order in the word space. However, in the lexical meaning space of topic $T$, the weight of $M$ should be the sum of word $m_1$ and $m_2$, which is 0.05 and greater than $N$ in topic $T$, the lexical meaning of $M$ is more able to represent the content of the topic.

Therefore, in our process of document modeling, words with the same meaning are aggregated into one lexical meaning element and expressed by word with the largest weight. Therefore, the topic word space of the original model is transformed into the topic lexical meaning space so that the weight of each element can fully reflect its weight of the lexical meaning in the topic space, which is more consistent with the law of language generation and can greatly reduce the dimension of co-occurrence matrix.

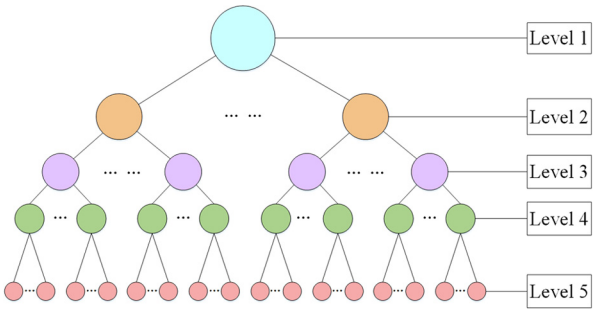


**Fig. 3.** Synonyms structure.

#### 3.2.2. Generation of synonym list based on synonymy thesaurus

The key step to aggregating words with the same meaning into one lexical meaning element in documents is to generate a proper synonym list. At present, there are two common methods, one is based on HowNet, and another is synonym thesaurus [17,18]. We choose synonym thesaurus, for it takes both the similarity of lexical meaning and the relevance of words into account.

The "Expanded Version of Synonymy Thesaurus" is a synonym thesaurus with a large Chinese word list compiled by the HIT-CIR based on "Synonymy Thesaurus" [19], which integrates a large number of word-related resources. It classifies all entries into five levels: large, medium, small, word group, and atomic word group, and organizes them together with a three-level system. The structure is shown in Fig. 3, and the coding method is shown in Table 2. Take "Da22B03= symptom, syndrome, disease, morbid appearance, pathology" as an example, "Da22B03=" is the code, "symptom, syndrome, disease, morbid appearance, pathology" represents the words of the class.

Based on the experiments of filtering and merging the synonyms in Lu's research [20], the fifth level was the best. Therefore, based on the fifth level synonymy thesaurus, we also add some synonyms which do not appear in the thesaurus, but appear in the

**Table 2**
Synonyms coding.

| Coding bit | Symbols | Semiotic properties | level |
|---|---|---|---|
| 1 | D | Large class | Level 1 |
| 2 | a | Middle class | Level 2 |
| 3 | 2 | Subclass | Level 3 |
| 4 | 2 | | |
| 5 | B | Word group | Level 4 |
| 6 | 0 | Atomic word group | Level 5 |
| 7 | 3 | | |
| | = | Synonym | |
| 8 | # | Same kind | |
| | @ | Independent | |

corpus frequently, such as synonym $S$ = {3M, M6200, M9001, ⋯, M9021}, which are synonymous expressions of "3M mask". Finally, a synonym list of indefinite dimensions for OHCs Q & A data is formed, and parts are shown in Table 3.

### 3.3. CL-LDA topic model based on lexical meaning co-occurrence

LDA topic model finds topics by mapping the co-occurrence probability relationship of words to semantic space at the level of available data. However, in OHCs, the length of Q&A data is short, the features are sparse, and the co-occurrence information contained in available data is insufficient, so the LDA topic model is difficult to generate high-quality topics [21]. Given LDA's limitations in short text topic mining, Gao [22] proposed a CO-LDA model combining the co-occurrence analysis of words with LDA topic model, which can improve the quality of topic generation. Based on the CO-LDA model and the lexical meaning information, we propose a CL-LDA topic model based on the lexical meaning co-occurrence.

Word co-occurrence model is one of the important models in the natural language processing field based on statistical methods [23]. Its basic idea is, if two words appear in the same window unit of the corpus, the two words are semantically related to each other, and the higher the frequency of co-occurrence means the closer relationship between them. The co-occurrence matrix is the matrix expression of the relationship between all words in the corpus, and it is the quantitative form of the co-occurrence model of words. By analyzing the co-occurrence of lexical meaning in the corpus, we can replace the original document word matrix with the co-occurrence matrix as the feature vector of the corpus, which can retain the co-occurrence information of words on the whole corpus level, overcome the sparsity of short text, and improve the quality of LDA model. In our CL-LDA topic model, the steps to generate the co-occurrence matrix are as follows.

1) Generate the lexical meaning corpus vocabulary. The definition of the standard lexical meaning corpus with $M$ comments is $D = \{d_1, d_2, \cdots, d_M\}$, by scanning every Q&A data in $D$, we can extract all mutually exclusive lexical meanings and count their word frequency, then add the lexical meanings which are greater than $\delta$ to the vocabulary in turn, and finally get the lexical meaning corpus vocabulary $W = [w_1, w_i, \cdots, w_N]$. Among them, $\delta$ is the word frequency threshold because the low-frequency words have a very low degree of topic identification [24], if it is added to the lexical meaning co-occurrence matrix, it will increase the noise of short text and reduce the accuracy of topic extraction. Therefore, by setting the word frequency threshold in the CL-LDA model, we only retain the lexical meaning $p(w_i)$ higher than the threshold $\delta$, and filters out the low-frequency features of noise and redundancy.
2) Generate the co-occurrence matrix of lexical meaning. For each lexical meaning $w_i$ in the vocabulary, scan each Q&A data $d_i$ in corpus $D$ in turn, and count the co-occurrence times $c_{ij}$ between $w_i$ and each lexical meaning $w_j$ in vocabulary $W$, as shown in formula (3); finally, generate a symmetric co-occurrence matrix $C$ whose row and column are $N \times N$, as shown in formula (4).

$$c_{ij} = \sum_{d=d_1}^{d_M} f(w_i, w_j), f(w_i, w_j) = \begin{cases} 1 & w_i, w_j \text{ co-occur in } d_a \text{ and } NPMI(w_i, w_j) \geq \epsilon \\ 0 & w_i, w_j \text{ not co-occur in } d_a \text{ or } NPMI(w_i, w_j) \leq \epsilon \end{cases} \tag{3}$$

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N1} & c_{N2} & \cdots & c_{NN} \end{bmatrix} \tag{4}$$

$$NPMI(w_i, w_j) = -\frac{\log_2(\frac{p(w_i, w_j)}{p(w_i)p(w_j)})}{\log_2 p(w_i, w_j)} \tag{5}$$

Where $f(w_i, w_j)$ is the co-occurrence frequency of the lexical meaning $w_i$ and $w_j$; NPMI is the standardized point mutual information, and the calculation method is shown in formula (5); $\epsilon$ is the co-occurrence threshold. In the process of co-occurrence analysis, there may be some accidental co-occurrence pairs of words, which are not related in the actual semantic space, so it is unreasonable to include them in the co-occurrence matrix. Standardized point-wise mutual information (NPMI) has a solid foundation in probability and information theory and is widely used in semantic similarity calculation and feature selection, so we use NPMI value between co-occurrence pairs to determine whether co-occurrence is reasonable. If the value of NPMI$(w_i, w_j)$ is greater than or equal to co-occurrence threshold, co-occurrence of $w_i$ and $w_j$ is considered reasonable, and co-occurrence times of $w_i$ and $w_j$ are retained; otherwise, co-occurrence times of $w_i$ and $w_j$ are considered unreasonable, and co-occurrence times of $w_i$ and $w_j$ are set to 0 to filter unreasonable co-occurrence word pairs.
3) Feature selection of lexical meaning co-occurrence matrix. In the traditional LDA model, corpus features are represented by topic word matrix; in CL-LDA model based on lexical meaning, corpus features are represented by co-occurrence matrix $C$. The row of co-occurrence matrix $C$ is a Q&A data composed of multiple feature words, and the column is a feature word representing the feature dimension. Because only a small part of the meaning in the vocabulary $W$ is shared in the corpus $D$, the generated co-occurrence matrix $C$ is a highly sparse symmetric matrix. To avoid the interference of the sparsity of the matrix on the topic mining, we select the TF-IDF value as the lexical meaning contribution degree to select the features of the meaning matrix.

The lexical meaning contribution degree is used to describe the contribution degree of lexical meaning to the whole lexical meaning co-occurrence matrix, and its calculation is shown in formula (6). Among them, $w_i$ means line feature word, $w_j$ stands for column characteristic word, $c_i$ is the frequency of $w_i$, $c_{ij}$ is the co-occurrence times of $w_i$ and $w_j$, $N$ is the total row number of co-occurrence matrix, and $n_j$ is the number of rows in the matrix containing $w_j$.

$$\begin{cases} TC(w_j) = \sum_{i=1}^{N} TF(w_j) \cdot IDF(w_j) \\ TF = \frac{c_{ij}}{c_i} \\ IDF = \lg \frac{N}{n_j} \end{cases} \tag{6}$$

**Table 3**
Synonym List for OHCs Q&A data.

| Synonym 1 | Synonym 2 | Synonym 3 | Synonym 4 | ··· |
|---|---|---|---|---|
| hospital- specialized hospital | diarrhea-watery diarrhea | effects-utility | station-bus station | |
| hospital-health center | diarrhea-washy | effects-function | station-railway station | |
| hospital-cottage hospital | diarrhea-run belly | effects-efficiency | station-departure | ··· |
| hospital-infirmary | diarrhea-tummy trouble | effects-validity | station-origin station | ··· |
| hospital-medical station | diarrhea-suffer from diarrhea | effects-achievement | station-terminus | ··· |
| ··· | ··· | ··· | ··· | ··· |

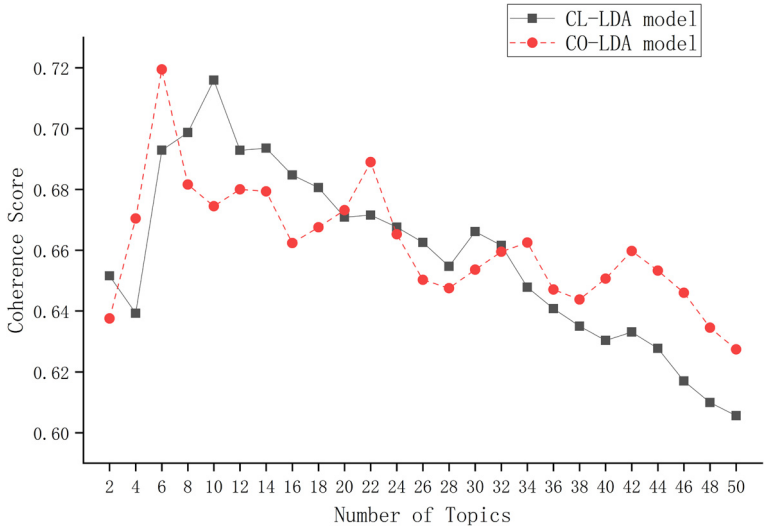


**Fig. 4.** Topics coherence score.

After calculating each lexical meaning contribution degree of $N$ feature words in turn, the lexical meaning contribution degree table is output in descending order according to the percentage weight of a single lexical meaning contribution degree in the full degree, and the first $M$ lexical meaning is extracted as the feature word table, whose total lexical meaning contribution degree weight is equal to the contribution degree threshold $\lambda$, and the columns of non-feature words in the lexical meaning co-occurrence matrix are deleted according to the feature word table. Finally, we get the co-occurrence matrix of $N$-row and $M$-column $C'$.

$$C' = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1M} \\ c_{21} & c_{22} & \cdots & c_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N1} & c_{N2} & \cdots & c_{NM} \end{bmatrix} \tag{7}$$

## 4. Experiments and results

### 4.1. Data sources and experiment environment

Because COVID-19 is a new infectious disease, most Chinese OHCs have not set up a specialized consultation column for it now. So we set “COVID-19”, “coronavirus virus”, “novel coronavirus” and “coronavirus pneumonia” as keywords and searched the Q&A data from six Chinese OHCs, including the 39 health website, good doctor online, answer website, doctors and medicine seeking website, micro medicine website, and spring rain doctors website. We collected Q&A data related to COVID-19 from January 20, 2020 to March 31, 2020. Each data includes the user’s gender, age, question time, question title, question description and doctor’s reply. After data cleaning, 8178 experiment data were finally obtained.

Under Windows 10 Operating system, we used the Anaconda as a development environment and Python 3.6 as the development language.

### 4.2. Determination of the optimal number of topics

According to the experience and experiment results of previous studies, we set 3 as the word frequency threshold $\delta$, 0.15 as the co-occurrence threshold $\epsilon$, 0.8 as the contribution degree threshold $\lambda$, set model super parameter value $\alpha = 50/K$, $\beta = 0.01$, and set the number of iterations of Gibbs sampling as 1000. The CV method is used to calculate the topic consistency score within the range of 2-50, and the results are shown in Fig. 4. The horizontal axis represents the topic number and the vertical axis represents the topic consistency score. It can be seen that when the number of topics is 6, the topic consistency score of CO-LDA is the highest. When the number of topics is 10, the topic consistency score of CL-LDA is the highest. However, for CL-LDA, when the topic number is set to 8, the topic consistency score is close to that of 10, and while the topic number is set to 10, the obtained meanings of topic 5 and topic 6, topic 7 with topic 8 and topic 9 are overlap, and topic 10 is meaningless noise topic. Therefore, in order to integrate the topic consistency score and topic quality, we choose to set the CO-LDA optimal topic quantity to 6, and the CL-LDA optimal topic quantity to 8. Some experimental results are shown in Table 4, Table 5 and Table 6.

In Table 6, topic 1 to 4 and topic 1′ to 4′ are the first four topics of the CL-LDA and CO-LDA model respectively. It can be seen that the results of topic 1 to 3 in these two models are similar, which are related to the disinfection method, disease treatment and immunity enhancement respectively. However, it can be found that compared with the CO-LDA model based on words, the CL-LDA model based on lexical meaning has more centralized topic semantics, no or less semantic repetition, which can express the

**Table 4**
The topic features of CL-LDA model with 8 topics.

| Topic 1-4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 |
|---|---|---|---|---|
| These topics are shown in Table 6 | mask | propagation | symptom | cough |
| | washing hands | infection | cough | symptom |
| | ventilate | defend | feverish | hospital |
| | travel | mask | patient | feverish |
| | defend | contact | manifestation | temperature |
| | protect | aerosol | breathe | infection |
| | contact | patient | stuffy nose | influenza |
| | outing | protect | runny nose | contact |
| | wear | method | constipation | manifestation |
| | infection | contagion | influenza | sense |
| | medical masks | wear | acute | inspection |
| | dense | crowd | treatment | treatment |
| | aerosol | quarantine | case | patient |
| | patient | outing | syndrome | constipation |
| | contagion | channel | sense | throat |

**Table 5**
The topic features of CL-LDA model with 10 topics.

| Topic 1-4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 | Topic 9 | Topic 10 |
|---|---|---|---|---|---|---|
| These topics are shown in Table 6 | mask | propagation | symptom | cough | cough | food |
| | defend | infection | cough | symptom | feverish | rest |
| | contact | contact | feverish | hospital | symptom | enhance |
| | wear | disease | patient | feverish | treatment | symptom |
| | ventilate | patient | breathe | temperature | temperature | treatment |
| | hand washing | quarantine | manifestation | throat | manifestation | cough |
| | protect | defend | runny nose | influenza | lung | infection |
| | outing | protect | contact | contact | infection | sleep |
| | travel | treatment | influenza | feverish | influenza | body |
| | infection | contagion | treatment | infection | runny nose | influenza |
| | medical masks | aerosol | stuffy nose | inspection | contact | nutrient |
| | propagation | crowd | constipation | treatment | constipation | exhaustion |
| | aerosol | mask | drug | drug | sense | aerosol |
| | method | method | syndrome | anxious | stuffy nose | poisoned |
| | contagion | channel | sense | return | hospital | outing |

**Table 6**
Comparison of CL-LDA and CO-LDA topic features.

| Topic 1 | Topic 1′ | Topic 2 | Topic 2′ | Topic 3 | Topic 3′ | Topic 4 | Topic 4′ |
|---|---|---|---|---|---|---|---|
| alcohol | disinfection | treatment | treatment | enhanced | rest | test | |
| disinfecting water | alcohol | medicine | medicine | nutrition | washing hands | nucleic acid | |
| disinfection | disinfectant | antiviral | antiviral | body | food | diagnosis | |
| ultraviolet rays | ultraviolet rays | effective | infection | immunity | nutrition | test | |
| containing chlorine | containing chlorine | nucleic acid | effect | going out | infection | treatment | |
| ether | disinfectant | patient | examination | do exercise | diet | symptoms | This topic is not recognized by the CO-LDA model based on words |
| peracetic acid | ether | methods | patient | washing hands | symptoms | fever | |
| kill virus | peracetic acid | disease | symptom | effect | mask | CT | |
| effect | kill virus | symptom | condition | treatment | body | cough | |
| inactivation | inactivation | infection | body | infection | treatment | patient | |
| destroying bacteria | destroying bacteria | medicine | fever | medicine | go out | lungs | |
| surface | surface | isolation | action | antiviral | fever | hospital | |
| survival | chloroform | hospital | testing | ventilation | population | infection | |
| environmental | ethanol | interferon | diagnosis | mask | enhance | blood routine | |
| temperature | temperature | recovery | dconfirm the diagnosis | illness | exercise | positive | |

topic characteristics more intuitive and multi-level. In topic discovery, the inspection topic is not recognized by the CO-LDA model, and the CL-LDA model is more sensitive to different topics and can accurately identify them. In terms of mining efficiency, the CL-LDA reduces the co-occurrence matrix from the original $3487 \times 3487$ dimension to the $3156 \times 1203$ dimension, which greatly reduced the time complexity and improves the efficiency of topic mining.

### 4.3. The clustering results of CL-LDA topic model

Because the Q & A data in Chinese OHCs involves a lot of knowledge in medical and health, it can't guarantee the accuracy of topic division completely relying on the CL-LDA model, so it is necessary to manually identify and merge the CL-LDA clustering results in combination with relevant research. Referring to

**Table 7**
Topics and key words list of information needs of COVID-19 in Chinese OHCs.

| Topic | Subclass | Concrete content | Key word |
|---|---|---|---|
| Symptom | - | Describe the symptoms of COVID-19 directly and ask if being infected with COVID-19 | Symptoms, cough, fever, body temperature, throat Cold, patient, sensation, diarrhea, hospital, etc. |
| Prevention | Disinfection method | Ask how to eliminate the virus or whether the disinfection method is effective | Alcohol, disinfectant, disinfection, UV, chlorine, Ether, peracetic acid, killing, effect, inactivation, etc. |
| | Transmission routes | Ask about the route and mode of transmission of the virus; describe the contact history to ask if it will be infected; ask about the selection and use of masks | Transmission, mask, prevention, infection, infection Protection, going out, flying foam, contact, medical mask, etc. |
| | Strengthen immunity | Ask about ways to improve immunity against viral infection, including nutrition, physical exercise and lifestyle | Strengthening, nutrition, body, immunity, going out Exercise, wash hands, effect, treatment, infection, etc. |
| inspection | - | Describe the relevant examination indicators, consult whether there is infection or how to make self-diagnosis | Examination, nucleic acid, diagnosis, detection, treatment Fever, CT, cough, patient, blood routine, etc. |
| Treatment | - | Ask about treatment methods and precautions for suspicious symptoms and diseases | Treatment, drugs, antiviral, efficacy, patients Methods, diseases, conditions, infections, drugs, etc. |

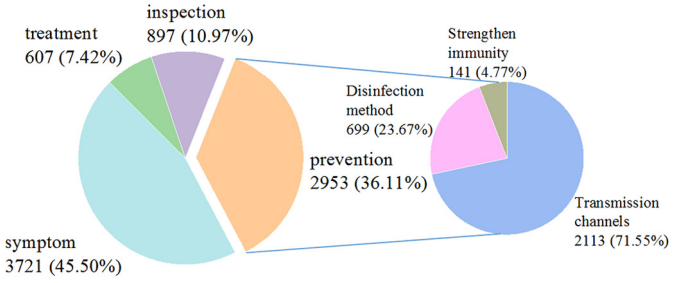


**Fig. 5.** Questions distribution of COVID-19 in Chinese OHCs under different topics.

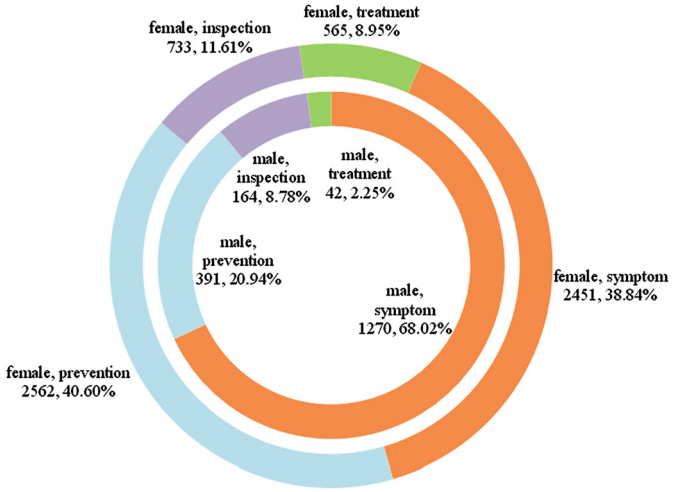


**Fig. 6.** Distribution of health information needs of different gender users.

the research of Chinese Center for Disease Control and Prevention [25], we combined the topic clusters of disinfection mode, transmission route and immunity enhancement into prevention topic, and the topic clusters describing suspicious symptoms from different clinical manifestations into symptom topic. For the combined topic, TF-IDF method is used to calculate the weight of each lexical meaning in the topic and extract the top 10 words meanings of the weight ranking in the topic as the keywords to represent the topic.

The topics and keywords of users' information needs of COVID-19 in Chinese OHCs are shown in Table 7. The number distribution of the problems under each topic is shown in Fig. 5. Some related nouns, such as COVID-19, novel coronavirus, novel coronavirus pneumonia and etc., and other none practical meaning words such as "only", "approval" are eliminated.

COVID-19 related information needs topics are mainly focused on symptoms, prevention, inspection and treatment in four aspects, as shown in Fig. 5, and these topics are closely related to the characteristics of COVID-19. COVID-19 cases are reported to be highly infectious, but most of the patients are mild [26]. Clinical presentations [27] are characterized by fever, dry cough and fatigue. A few of the cases are accompanied by diarrhea, runny nose and expectoration. Because the COVID-19 is highly infectious and the initial symptoms are similar to those of the upper respiratory tract infection and influenza, which is frequent in winter and spring, the users' information needs for symptoms, prevention and examination are most urgent, and occupy a large proportion in the Q&A. In the topic of prevention, COVID-19 is mainly focused on the subcategories of transmission routes and disinfection methods, while the information needs for improving immunity are few. This means that the main way to prevent the COVID-19 is to cut off spread and eliminate the virus. Since currently, there is no COVID-19 specific drug, and once the patient is identified as suspected or confirmed, the Chinese government will provide free and comprehensive medical treatment, so the users' information needs for treatment are few and mostly for the treatment of existing suspicious symptoms.

## 5. Discussion

On the basis of the Q&A data and results of topic mining experiments, we analyze the information needs trend and distribution, furthermore discern several important findings.

### *5.1. The gender differences of information need distribution*

The gender distribution of COVID-19 health information needs is shown in Fig. 6. On the whole, the number of female users' questions is far more than that of male users, the proportion is about 3.38 (6311): 1 (1867). According to the statistical data [26],

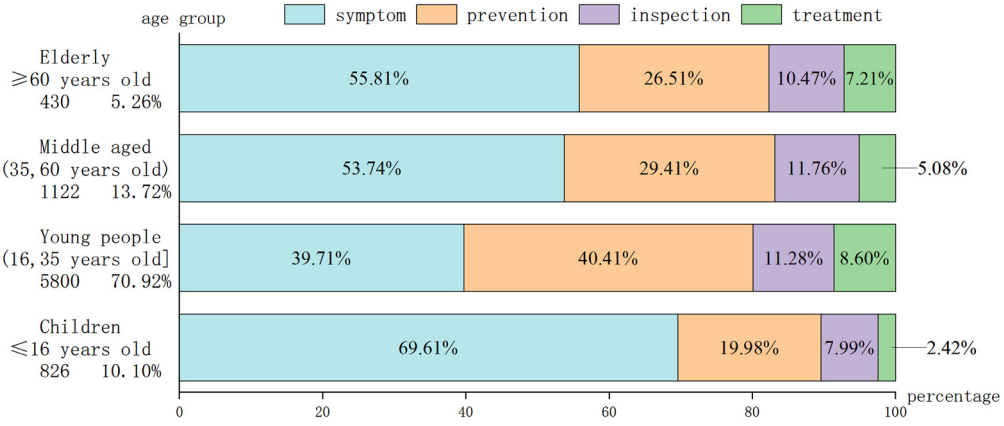


**Fig. 7.** Distribution of health information needs of users at different age stages.

the proportion of women and men in the national confirmed cases is about 1:1.06. Although the prevalence of men is slightly higher than that of women, the willingness of women users to obtain relevant information from OHCs is still greater than that of men. The Chi-square test result of gender distribution and information needs topics is $\chi^2 = 521.596$, $P < 0.01$, which demonstrates that there is a significant difference in the health needs of COVID-19 between female and male users. We discover that male users have the most urgent needs for diagnosis of suspicious symptoms, followed by prevention needs, while female users are more concerned about how to prevent virus infection, followed by symptom needs. Thus it can be seen that male users pay more attention to the recognition and judgment of existing suspicious symptoms, while female users incline to prevent the disease in advance.

### 5.2. The age difference of information needs distribution

Based on the study of Lu [14], we divide users into four different age stages: children ($\leq$ 16) years of age, young people between (16, 35] years old, middle-aged people between (35,60) years old and elderly people ($\geq$60) years old. The users' age distribution of information needs for COVID-19 is shown in Fig. 7. Overall, the number of questions from young people is the largest, accounting for 70.92% of the total, which is closely related to the age structure of Internet users in China [28]. Different from previous studies, the proportion of children's questions (10.10%) exceeds that of the elderly (5.26%) and is close to that of the middle aged (13.72%). The increase of the number of children's questions may be due to the increased frequency of children's use of electronic devices and the Internet during holidays. The number of elderly users' questions only accounts for 5.26% of the total. On the one hand, it is related to elderly users' preferences for access to health information. Wang's study [29] shown that people over 60 years old mainly obtained health information through WeChat forwarding or friends circle (22.99%) and search engine (14.81%), only 5.09% of the elderly people get health information through interactive OHCs. On the other hand, the elderly users accept new things a little more slowly, and their sensitivity of new infectious diseases such as COVID-19 is relatively low, which also limited the number of elderly users' questions.

The Chi-square test result of information needs distribution from different age stages is $\chi^2 = 344.329$, $P < 0.01$, which demonstrates that there are significant differences among users of different ages. Among them, children pay more attention to the topic of symptoms (69.61%), but less attention to the topic of prevention (19.98%), which indicates that children have a low awareness of the importance of prevention. Young users' needs for prevention topic (40.41%) is higher than that for symptom topic (39.71%), which is different from other age groups' focus on symptom topic, indicating that young users have a higher awareness of virus prevention.

### 5.3. The information needs changing trend over time

The information needs changing trend of COVID-19 over time is shown in Fig. 8. It can be seen that from January 20, 2020, to March 29, 2020, the user's overall information needs related to COVID-19 tends to be consistent with the change of various topics over time, showing a trend of firstly rising, then fluctuating, and then declining and becoming stable. On January 20, 2020, the National Health Commission of the People's Republic [30] released the No.1 announcement in 2020, and incorporated the COVID-19 into the class B infectious diseases stipulated in the infectious disease prevention act of the People's Republic of China and adopted the prevention and control measures of class A infectious diseases, then the information about COVID-19 began to appear in Chinese OHCs and showed a rapid growth trend. As shown in Fig. 9, on January 27, 2020, the number of newly confirmed cases increased significantly in China, which increased slowly after reaching the first epidemic peak [26]; on January 31, 2020, the user's needs for health information about COVID-19 also reached the first peak (lag period of 4 days), and then decreased slowly. The main reason for the decline is that with the increase of the government's propaganda and prevention and control efforts, people's knowledge of COVID-19 has been deepened. On February 4, the newly confirmed cases in China reached the second peak; on February 7, the user's needs for health information in OHCs reached the second peak (lag period of 3 days). On February 12, due to the change of the diagnosis standard of the confirmed cases, the new confirmed cases reached the third peak after including the 13332 clinical diagnosis cases in Hubei Province, and the user's demand for health information in OHCs also reached the third peak (lag period of 4 days) from February 16 to February 19. Then the information needs of COVID-19 decreased gradually and stabilized with the continuous reduction of newly diagnosed cases and the increasing public awareness of the virus.

Based on the above analysis, we can see that the information needs of COVID-19 in Chinese OHCs were affected by the development characteristics of the epidemic. The overall trend of change was consistent with the trend of new confirmed cases in the country, but there was a 3-4 day lag in the emergence of the peak.

## 6. Conclusion

In order to discover the information needs of COVID-19 in Chinese OHCs, we propose a CL-LDA model based on lexical meaning

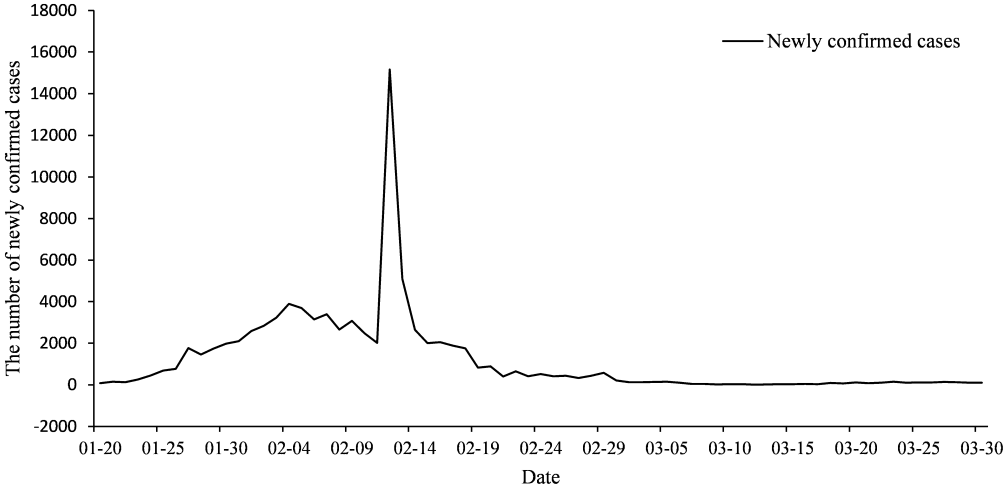


**Fig. 8.** Changing trend of different topics over time.

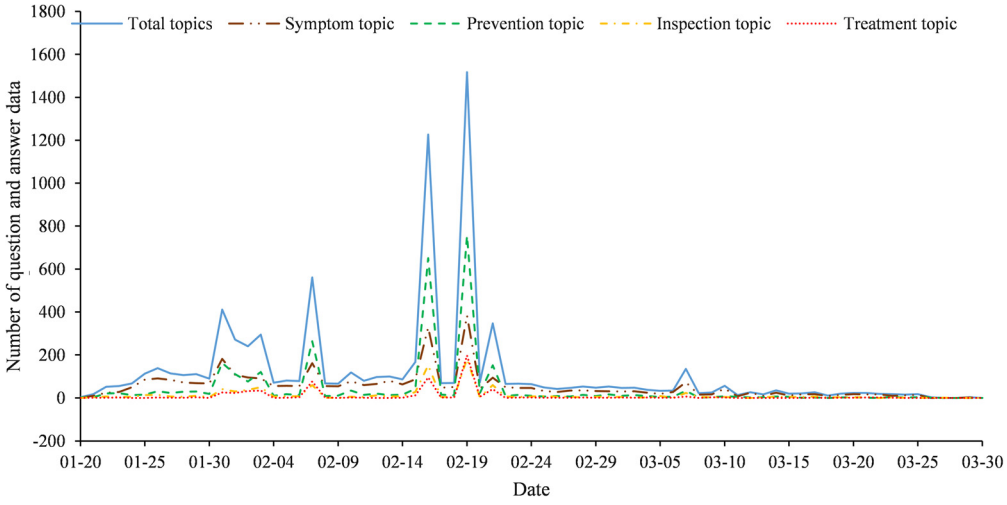


**Fig. 9.** Trend of newly confirmed cases in China over time.

co-occurrence analysis and LDA topic model in this study. While comparing with CO-LDA model, the experiment results showed that CL-LDA had more centralized topic semantics, expressed the topic characteristics more intuitive and multi-level, and greatly reduced the dimension of the co-occurrence matrix and improved the efficiency of topic mining. The important findings of our study are concluded as follows.

- The information needs of COVID-19 in Chinese OHCs were affected by the development characteristics of the epidemic, and the overall changing trend was consistent with the trend of new confirmed cases in the country. Therefore, OHCs should allocate resources reasonably, and give more resources preference when the information needs reaches the peak, so as to alleviate the pressure of social medical resources shortage.
- Children pay the least attention to the topic of prevention, suggesting that the government should strengthen education on the prevention of infectious diseases among children and raise their health awareness with the support of schools.
- There were significant gender and age differences in users' information needs of COVID-19 in Chinese OHCs, and the OHCs should provide personalized information services for different user groups to improve service levels and service quality.

In the following work, the topic emotion and evolution analysis will be carried out for the mined topics to help the Chinese OHCs improve the quality of service and provide decision support for information and education on infectious diseases.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] H. Jiang, D. Yang, C. Guo, Impact of novel coronavirus pneumonia on agricultural development in China and its countermeasures, Reform 003 (2020) 5–13.
[2] Publicity Department of Health Committee of the People's Republic of China, Transcript of press conference on March 20, 2020, http://www.nhc.gov.cn/xcs/s3574/202003/4ad24ab68e2441668b569757b147c100.shtml, 2020.
[3] J. Shi, S. Li, Y. Qian, L. Zhou, B. Zhang, Information needs of domestic and international HCQA users an empirical analysis, in: Data Analysis and Knowledge Discovery, vol. 003, 2019, pp. 1–10.
[4] A.C. Johnston, J.L. Worrell, P.M.D. Gangi, M. Wasko, Online health communities: an assessment of the influence of participation on patient empowerment outcomes, Inf. Technol. People 26 (2013) 213–235.
[5] M.S. Prabha, B. Sarojini, Online healthcare information adoption assessment using text mining techniques, Mob. Netw. Appl. 24 (2019) 1160–1165.
[6] J. Liu, J. Kong, Z. Xin, Study on differences between patients with physiological and psychological diseases in online health communities: topic analysis and sentiment analysis, Int. J. Environ. Res. Public Health 17 (2020) 1508.

[7] X. Xu, Y. Zhao, Q. Zhu, An empirical study on health information needs of elderly users in online health communities, Libr. Inf. Serv. 63 (2019) 87–96.
[8] B. Jin, X. Xu, Health information needs of diabetics in social Q & A community, Chin. J. Med. Libr. Inf. Sci. 23 (2014) 37–42.
[9] S. Oh, Y. Zhang, M.S. Park, Cancer information seeking in social question and answer services: identifying health-related topics in cancer questions on Yahoo!Answers, Inf. Res. 21 (2016) 718.
[10] X. Zhou, W. Liang, K.I. Wang, S. Shimizu, Multi-modality behavioral influence analysis for personalized recommendations in health social media environment, IEEE Trans. Comput. Soc. Syst. 6 (2019) 888–897.
[11] X. Zhou, W. Liang, K.I. Wang, H. Wang, L.T. Yang, Q. Jin, Deep-learning-enhanced human activity recognition for Internet of healthcare things, IEEE Int. Things J. 7 (2020) 6429–6438.
[12] X. Zhou, Y. Li, W. Liang, CNN-RNN based intelligent recommendation for online medical pre-diagnosis support, IEEE/ACM Trans. Comput. Biol. Bioinform. (2020), PP(99):1–1.
[13] X. Tang, J. Li, Analysis on the topic and sentiment of information needs in online health community, in: Digital Library Forum, 2019, pp. 12–17.
[14] Q. Lu, A. Zhu, J. Zhang, J. Chen, Research on user information requirement in Chinese network health community: taking tumor-forum data of Qiuyi as an example, in: Data Analysis and Knowledge Discovery, vol. 3, 2019, pp. 22–32.
[15] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[16] W. Che, Z. Li, T. Liu, LTP: a Chinese language technology platform, in: Coling 2010: Demonstrations, Coling 2010 Organizing Committee, Beijing, China, 2010, pp. 13–16.
[17] S. Huang, Y. Zhou, Semantic orientation computing based on HowNet&Cilin, Software 34 (2013) 73–74 + 94.
[18] C. Zeng, D. Zhang, Phrase subject extraction based on synonyms and HowNet, J. Xiamen Univ. Natur. Sci. 54 (2015) 263–269.
[19] J. Mei, Y. Zhu, Y. Gao, Synonymy Thesaurus, Shanghai Lexicographical Publishing House, Shanghai, China, 1996.
[20] Z. Lu, Y. Lin, S. Zhao, W. Zhu, Study on the feature selection and weighting based on Tongyici Cilin in text categorization, J. Infor. 05 (2015) 130–132.
[21] X. Zhou, Y. Hu, W. Liang, J. Ma, Q. Jin, Variational LSTM enhanced anomaly detection for industrial big data, IEEE Trans. Ind. Inform. (2020), PP(99):1–1.
[22] H. Gao, J. Liu, S. Yang, Identifying topics of online healthcare reviews based on improved LDA, Trans. Beijing Inst. Techol. 39 (2019) 427–434.
[23] D. Zheng, T. Zhao, S. Li, H. Yu, Research on a novel word co-occurrence model and its application, in: Knowledge Science, Engineering & Management, Second International Conference, KSEM, Melbourne, Australia, November 2007, pp. 437–446.
[24] Y. Hu, J. Jiang, H. Chang, A new method of keywords extraction for Chinese short-text classification, in: Data Analysis and Knowledge Discovery, 2013, pp. 42–48.
[25] Chinese Center for Disease Control and Prevention, General guidelines for novel coronavirus pneumonia prevention, J. Qilu Nursing 26 (2020) 21.
[26] Epidemiology Working Group for NCIP Epidemic Response, Chinese Center for Disease Control and Prevention, The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China, Chin. J. Epidemiol. 41 (2020) 145–151.
[27] Publicity Department of Health Committee of the People's Republic of China, Novel coronavirus infection diagnosis and treatment guideline for pneumonia (third edition), http://www.nhc.gov.cn/xcs/zhengcwj/202001/f492c9153ea9437bb587ce2ffcbee1fa/files/39e7578d85964dbe81117736dd789d8f.pdf, 2020.
[28] China Internet Network Information Center, The 44th statistical report on the development of China's Internet, http://www.cac.gov.cn/pdf/20190829/44.pdf, 2019.
[29] W. Wang, H. Yu, X. Cao, J. Liu, l. Pei, Accession and use of health information with intelligent mobile phone in community elderly people, Chin. J. Med. Libr. Inf. Sci. 28 (2019) 71–76 + 80.
[30] National Health Committee of the People's Republic of China, Announcement of National Health Committee of the People's Republic of China, http://www.nhc.gov.cn/jkj/s7916/202001/44a3b8245e8049d2837a4f27529cd386.shtml, 2020.