Summer 7-5-2020

Research on Medical Overtreatment Based on LDA and Structural Equation Model

Weilong Liu
School of Management Science and Engineering, Shandong University of Finance and Economics, Ji'nan, 250014, China, lwl_sdufe@163.com

Daojia Xi
School of Management Science and Engineering, Shandong University of Finance and Economics, Ji'nan, 250014, China, XDJ1281732910@163.com

Research on Medical Overtreatment Based on LDA and Structural Equation Model

Weilong Liu[1*] , *Daojia Xi*[2]
School of Management Science and Engineering, Shandong University of Finance and Economics, Ji'nan, 250014, China

Abstract: Medical overtreatment has caused a lot of waste of medical resources. In the face of the increasingly serious medical overtreatment phenomenon, it is of great significance to clarify the factors of medical overtreatment to help solve this problem in China. In this study, we use the medical overtreatment text as a corpus and use latent Dirichlet allocation (LDA) topic model for topic extraction. Based on the extracted topics, a path model is established and the structural equation model(SEM) is used to test the path model. Finally, the influence factors of medical overtreatment are obtained. The results show that this study has extracted three main reasons that affect medical overtreatment, namely doctors, hospitals and patients. The factors influencing doctors' medical overtreatment are the institutions, benefits, and induced demand. The factors that affect patients' overtreatment are health and medical insurance. The factors that influence hospitals' medical overtreatment are monopoly, economics, and management. These factors significantly affect the occurrence of medical overtreatment. Therefore, public health organizations should proceed from these three aspects and formulate effective measures to solve the problem of medical overtreatment.

Keywords: medical overtreatment, influencing factors, LDA, SEM

1. INTRODUCTION

Medical overtreatment is a diagnosis and treatment that exceeds the actual needs of patients, causing unnecessary waste of medical resources and lost of patients [1]. With the development of medicine and the advancement of science and technology, although more and more medical problems have been solved, medical overtreatment problem have become increasingly prominent and have gradually become world-class medical problems. Although different countries have different systems, they all have serious over medical treatment. Lyu Heather[2] surveyed the extent to which 2,106 physicians at the American Medical Association used medical overtreatment care in their practice. The results showed that 20.6% of medical care, 22.0% of prescription drugs, 24.9% of examinations, and 11.1% of surgery were unnecessary at the time of treatment. The WHO recommends the use of antibiotics in hospitals as 30%. In China, the median use of antibiotics in patients is as high as 79%, which is more than double the recommended use rate worldwide [3]. In the face of the increasingly medical overtreatment phenomenon, exploring the causes of medical overtreatment and finding solutions is of great significance for maintaining the smooth operation of China's health service.

Many scholars have studied and published opinions on the causes of medical overtreatment. Shi[4] pointed out from the perspective of large hospitals that the extensive operation and management of hospitals will lead to the increase of operating costs which are ultimately transfered to patients, and inevitably lead to the phenomenon of medical overtreatment. Zhu[5] said that the information asymmetry between doctors and patients will lead to doctors to implement medical o*vertreatment behaviors. Chioleroa[6] pointed out in the research on prevention of medical overtreatment problems that patients have a strong subjective demand for diagnosis and medication to promote medical overtreatment problems.

* Corresponding author. Email: lwl_sdufe@163.com (Weilong Liu) , XDJ1281732910@163.com (Daojia Xi)

However, the way of the above research is subjective inference of influencing factors, lack of empirical analysis of the results, and few literatures use empirical analysis to study medical overtreatment issues. In this context, this paper uses machine learning and empirical analysis to explore the issue of medical overtreatment. We use the LDA topic model and perplexity to determine the optimal number of topics, reveal the influencing factors of medical overtreatment and establish a path model, and finally use the structural equation model to verify the path model.

2. STRUCTURE EQUATION MODEL AND LATENT DIRICHLET ALLOCATION

2.1 Structure equation model

Karl Joreskog first constructed a structural equation model (SEM) to study multivariate relationships [7].SEM is a commonly used empirical analysis model to find the structural relationship between variables. In SEM, latent variables can be used for target analysis, but they cannot be directly observed. Observed variables can be directly measured to estimate latent variables and support target analysis. SEM can reflect the relationship between latent variables and observed variables, making variables that could not be directly observed can be measured and embedded in the system of equations that reflect causality [8].

SEM has been widely used in various fields, such as high-order factor analysis, path analysis and causal analysis, etc. It can measure the relationship between variables, such as causal and co-occurrence. The relationship between the variables can be visualized by using a path model. In the schematic diagram of the structural equation model, an ellipse is used to represent the latent variable, a rectangle is used to represent the observed variable, a single arrow indicates a causal relationship, and a two-way arrow indicates a co-occurrence relationship. As shown in Figure 1, the figure contains three observation variables X1, X2, X3 and one latent variable Y, among which X1 and Y, X2 and Y are causal relationships, X3 and Y are co-occurrence relationships, ε is the error term. μ is the path coefficient, indicating the degree of relationship between variables [9].

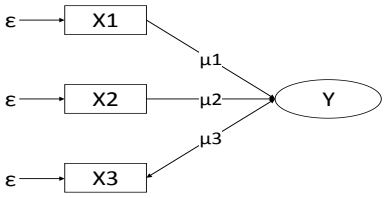

Figure 1. SEM diagram

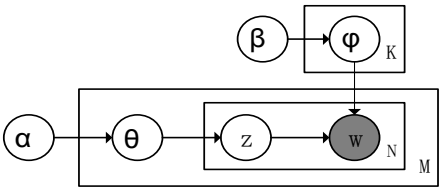

Figure 2. LDA model

2.2 Latent Dirichlet allocation

The latent Dirichlet allocation (LDA) topic model was proposed by BLEI In 2003. It is essentially a three-layer Bayesian model, which can extract the research topics contained in the text [10]. The LDA model consists of three parts: documents, topics, and words. It can retain the essential statistical information in the corpus and process the documents quickly and efficiently.

The LDA topic model is an unsupervised algorithm that can effectively analyze unstructured document sets and extract multiple topics from it. When the model is generated, it is assumed that "each word selects a certain topic with a certain probability, and selects a certain word from this topic with a certain probability". The LDA model is shown in Figure 2. M represents the total number of articles in the corpus, K represents the number of topics set, N represents the number of all words, and W is the number of words observed. θ is a matrix of M * K, which represents the topic distribution of the document. Φ is a matrix of K * V(V represents the vocabulary of all words that appear in all training corpora), which represents the word distribution of the topic. α is the hyperparameter of the Dirichlet distribution of θ, and β is the hyperparameter of the Dirichlet distribution of Φ [11]. Therefore, the probability of the ith word in the document can be calculated by equation (1).

$$P(wi)=\sum_{j=1}^{K} P(wi|zi=j)P(zi=j) \quad (1)$$

In the formula, P (z_i = j) indicates the probability that the word selected from the article is topic j, and P ($w_i|z_i$ = j) indicates the probability that the word taken is i when the topic is j[7].

The process of LDA topic modeling is expressed as follows:

（1）Select $\vec{\theta}$i～Dir($\vec{\alpha}$)，i∈{1,2...M}

（2）Select $\vec{\varphi}$k～Dir($\vec{\beta}$)，k∈{1,2...K}

（3）For each word position wi,j，j∈{1,2...N}，i∈{1,2...M}

① Choose a theme from zi,j～Mul(θi)

② Choose a word from wi,j～Mul(Φzi,j)

2.3 Perplexity

In natural language processing, perplexity evaluation is one of the important methods in measuring the pros and cons of language probabilistic models. The lower the model's perplexity, the stronger the generalization ability of the model, and the better the model's effect [11]. The perplexity formula is expressed as equation(2):

$$\text{perplexity}(D)=\exp\left(-\frac{\sum logp(wd)}{\sum_{d=1}^{M} Nd}\right) \quad (2)$$

In the equation(2), D is the test set in the corpus, M is the total number of documents, N_d is the number of words in each document d, w_d is the word in document d, and $p(w_d)$ is the probability of the word w_d in the document.

3. ANALISIS PROCESS USING LDA WITH SEM

This paper combines LDA and SEM methods to explore and analyze the important factors affecting medical overtreatment. The main process flow is corpus extraction→topics extraction→construction of path model→analysis by SEM. The process is shown in Figure 3.

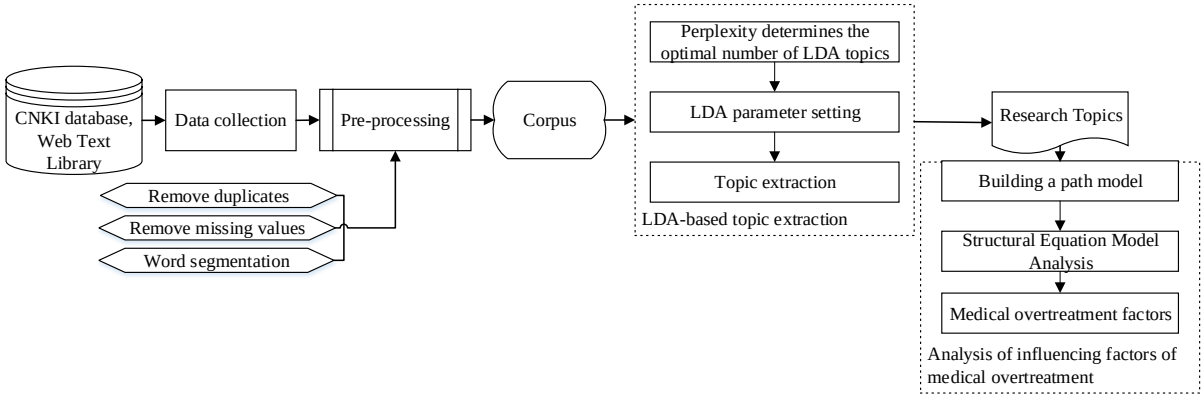

Figure 3. LDA and SEM construction process

3.1 Corpus extraction

The step of corpus extraction is mainly based on the data mining method to obtain text data, and then through filtering the text, deleting missing items, segmenting words, removing stop words, loading custom dictionaries and other pre-processing operations to form a corpus for easy reading and analysis.

3.2 Topics extraction

The topic extraction uses LDA topic modeling method. During the topic modeling, a suitable number of topics needs to be selected. Too few topic selections will lead to reduced interpretable information and accuracy; while too many topic selections may cause unrecognition topics to lead to a decrease in the reliability of the data [12].Therefore, the choice of the number of topics is generally 3-8. In the traditional LDA topic extraction, the number of topics is often determined based on experience. Generally, the optimal number cannot be directly

selected. In order to ensure a suitable number of topics, a perplexity auxiliary topic selection can be selected. The lower the value of perplexity, the better the corresponding number of topics[13]. The topic names is determined according to the keyword contribution, the relationship between topics and the relationship between documents in LDA analysis results.

3.3 Construction of path model

The words in the results generated by the LDA topic model are arranged in descending order according to the degree of contribution. After removing the words that have no practical meaning to the topic, three words with higher contributions are selected as observation variables, and the word frequency is used as a measure of the observed variables. When choosing words, we avoid choosing the same words for different topics. After the latent variables and observation variables are determined, the path model of the latent variables pointing to the target variable can be determined.

3.4 Analysis by SEM

The word distribution data and the established path model were used for SEM analysis, and the Amos software was used to detect various indicators. In order to verify the fitting degree of the model, representative model fitting indexes GFI (goodness fitness index), CFI (comparative fitting index), and RSMEA (root mean square error of approximation) were selected for measurement. The GFI index needs to be greater than 0.9 and less than 1, the closer to 1, the better the effect; RSMEA should be less than 0.1, preferably less than 0.08; the CFI value is greater than 0.9 and less than 1, the closer to 1, the better the effect [9].

4. MODEL BUILDING TESTING

4.1 Corpus extraction

In order to study the influencing factors of medical overtreatment, this paper uses data mining to crawl the data. The main sources of data include web review articles, web news and CNKI text related to medical overtreatment. In the final collected data, there were 27 web review articles, 54 web news, 160 CNKI papers, and a total of 241 relevant texts. After preprocessing, this was used as a corpus for subsequent analysis.

4.2 Topics extraction and construction of path model

In the process of topic extraction, this paper uses the method of perplexity evaluation to determine the optimal number of topics. This paper calculates the perplexity of different topic number models.

As the topic number changes, the value of the perplexity will also fluctuate. When the topic number is 4, the perplexity values the lowest, as shown in Figure 4. Therefore, the best topic number for the model can be determined to be 4.

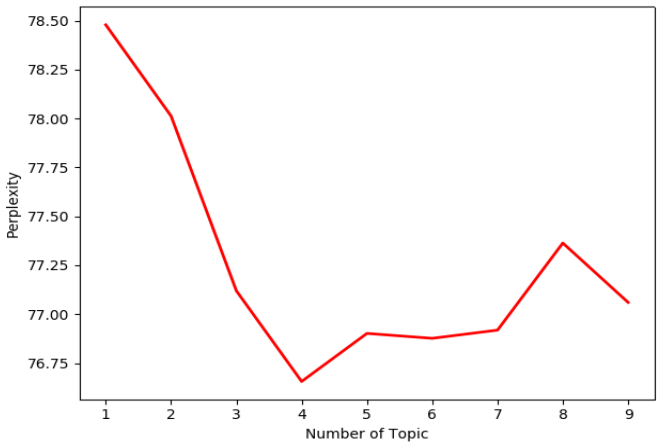

Figure 4. Number of topics-curve of perplexity

We use the LDA topic extraction method provided by the gensim package, set the number of extracted topics to 4, and iteratively filter out words that are not related to the topic and output related words. The results are shown in Figure 5.

Part of the samples are extracted from the corpus, and content mining is carried out in the way of deep reading. It is found that the categories of doctors, hospitals, and patients promote the occurrence of medical overtreatment. By observing the expression categories of the topic words in Figure 5 distributed in the sample articles, the word- category table shown in Table 1 is established.

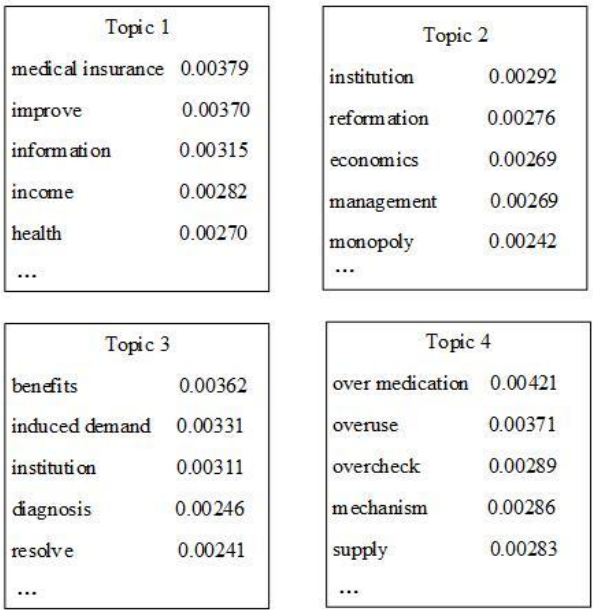

Figure 5. LDA topic distribution

Table 1. Word category table

Data	Category labeling	Source
#1:At present, the "grimace" of over treatment has been formed in China, which is characterized by minor diseases and serious diseases, **more examinations**, **more prescriptions**, **more treatment** and **long-term hospitalization**.	medical overtreatment	CNKI
#2:Doctors violate professional ethics and ethics, in order to allow them to obtain the maximum **benefits**, so as to take advantage of their positions and information advantages, to carry out diagnosis and treatment of patients beyond the actual conditions of patients. #3:This special status of doctors determines that they have absolute advantages in medicine and medical information, and this advantageous position also creates a good opportunity for them to **induce** the medical **demands** of patients.	doctor factor	CNKI
#4:But hospitals are not companies, especially public hospitals. Placing the company's **management** model in the hospital distorts the medical staff's diagnosis and treatment behavior. #5:In order to improve the **economic** benefits of the hospital, the hospital managers must allocate enough task indicators for each clinical department every year. #6:The **monopoly**, externalities and i**nformation asymmetry** of the medical service market affect the normal operation of the medical service market.	hospital factor	CNKI
#7:People's material and cultural living standards have greatly improved and increased, and their economic **income** has been increasing. Coupled with the establishment and improvement of various **medical security systems**, this has enabled the people to see a doctor for medical treatment. #8:With the continuous progress of society, people's psychological needs for **health** continue to grow. Patients are willing to try any medical treatment for the sake of their health.	patient factor	CNKI

According to the comparison of vocabulary data in Figure 5 and Table 1, topics 1 to 4 can be summarized into four topics: patients, hospitals, doctors, and medical overtreatment. Taking the three topics of patients, hospitals, and doctors as latent variables that affect the topic of medical overtreatment, a path model of doctors, hospitals, and patients influencing medical overtreatment is established. Three highly-contributing words of each topic were selected for topic measurement. Therefore, the words selected for the four topics are patients ("health", "income", "medical insurance"), hospitals ("economy", "monopoly", "management"), doctors ("benefits", "institutions", "induced demand"), medical overtreatment ("overcheck"," overuse", "overmedication"). The structural equation model test is performed with the word frequency data representing the words.

4.3 Structural equation model testing and evaluation

From the statistical word frequency data, 50 pieces of data that are all 0 are eliminated, and 191 experimental data are finally obtained. By using amos software to verify the model and observe the indicators. The results show that GFI = 0.925, greater than 0.9; RSMEA = 0.071, less than 0.08; CFI = 0.885, which is close to 0.9 and slightly lower than the standard. Based on these indicators, it can be found that the model has an acceptable degree of fit [14]. The final path model results are shown in Figure 6.

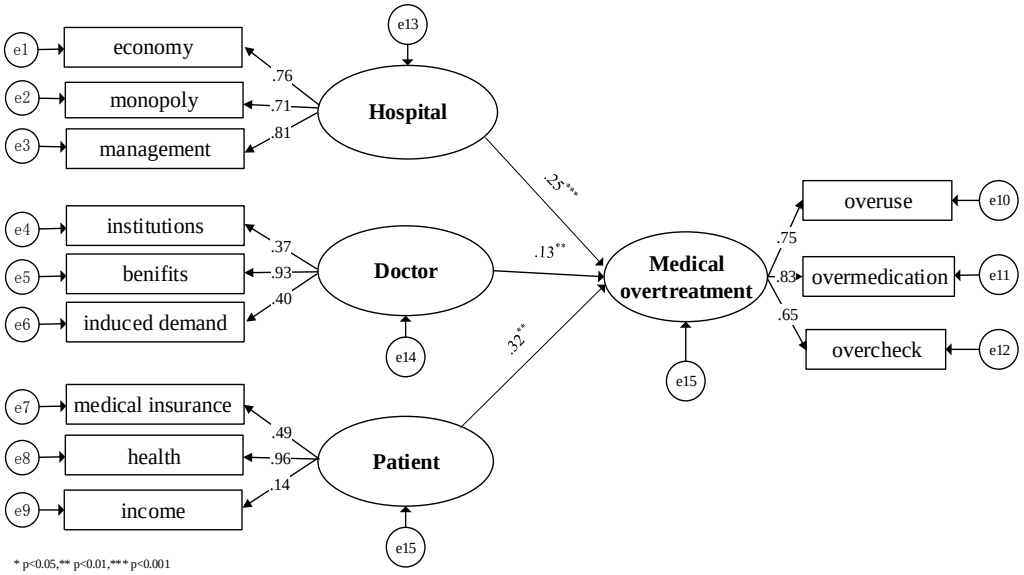

Figure 6. Path model results

5. RESULT AND DISCUSSION

According to the results of the path model in Figure 6, it can be found that the factors that affect medical overtreatment are mainly doctors, hospitals, and patients. These three factors can positively affect the generation of medical overtreatment problems.

5.1 Doctor factors

According to the results in Figure 6, "institutions" and "induced demand" have a certain effect on doctors' medical overtreatment behavior, but "benefits" are the most important influencing factor. Due to the imperfect institution of the hospital, the economic benefits of doctors are linked to the patient's medical services, and even performance competition between doctors appears. In order to maximize their own benefits, doctors will inevitably promote the occurrence of medical overtreatment. Patients are often in a passive position when they receive medical services from doctors, which increases the uncertainty of medical consumption. In order to

obtain economic benefits, doctors use the initial position of doctors relative to patients to induce patients to have excessive demand for medical services. Therefore, they issue expensive drugs to patients and carry out expensive examinations, so that patients spend more money to buy medical services, thus resulting in medical overtreatment.

5.2 Hospital factors

According to the results in Figure 6, factors such as "monopoly", "economy", and "management" can significantly affect hospital medical overtreatment. From the results, the "management" of the hospital is the most important factor. National medical and health services are mainly monopolized by public hospital, including medical information and medical resources. Patients tend to choose large hospitals when they seek medical treatment. However, due to insufficient management mechanisms, hospitals purchase high-consumption medical equipment. In order to make up for the cost of purchasing and obtain economic and social benefits, the hospital has developed an unreasonable hospital management mechanism. This mechanism encourages doctors to prescribe more drugs and check more for patients, and even relates the income of departments to the interests of doctors, which leads medical overtreatment.

5.3 Patient factors

According to the results in Figure 6, among the factors that affect patients' medical overtreatment, the impact of "income" is not significant. Both the "medical insurance" and "health" factors can significantly affect patients' medical overtreatment, of which health factor is the most important indicator. With the development of Chinese society, people's living standards have also continuously improved, and people have gradually shifted their perspectives to their own health fields. However, the sources of people's health information are still not extensive, mainly focused on search engines or medical apps, resulting in information asymmetry of between patients and doctors. Patients simply believe that the use of more expensive drug treatments and advanced equipment examinations can help their health, resulting in a strong demand. Benefiting from the increase in patient income, doctors are often required to prescribe medicines and perform detailed examinations during treatment. In addition, the strength of China's medical insurance has continued to increase, and the cost-sharing mechanism has further reduced patients' medical expenses. The suppressed medical needs of the masses have been released, which has further promoted patients' medical overtreatment behaviors such as excessive medication and examinations.

6. CONCLUSION

Although many scholars have studied the influencing factors of medical overtreatment, few have carried out quantitative analysis. This paper uses LDA model and SEM to explore the factors that affect medical overtreatment. Using the obtained medical overtreatment text as a corpus to perform topic mining and analyze the element structure.

According to the results of this study, China is facing a severe situation of medical overtreatment. Medical overtreatment not only increases the financial burden on patients, but also reduces the effective allocation of medical resources. At the same time, this is also one of the focuses of China's current medical reform. Therefore, relevant government departments must take measures from the perspectives of doctors, hospitals, and patients to curb the development of medical overtreatment.

The limitation of this paper is that there is a strong subjectivity in the identification and extraction of topic words and topic generalization. Different people may have different explanations for the model established in this paper.

ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China under Grant 71373144.

REFERENCES

[1] Liu Junqiang, Liu Kai, Zeng Yi. (2015). The mechanism of sustained increase in medical expenses: an analysis based on historical data and field data. Chinese Social Sciences, No.236 (08), 105-126 + 207-208. (in Chinese)

[2] Heather L, Tim X, Daniel B. (2017). Overtreatment in the united states. PLOS ONE, 12(9), e0181970.

[3] Liu Ye, Yang Yue. (2016). Analysis and Suggestions on the Status of Antibiotic Abuse in China. Modern Chinese Doctors,54 (29): 160-164. (in Chinese)

[4] Shi Zhaorong. (2011). New Thoughts on the Causes of Over-medicine in General Hospitals. Journal of Medical Postgraduates, 24 (8), 853-855. (in Chinese)

[5] Zhu Jiwu. (2015). Analysis of the Impact of Information Asymmetry on the Growth of China's Medical Expenses. Price Theory and Practice (12), 77-79. (in Chinese)

[6] Chioleroa, A, Paccauda F. (2015). How to prevent overdiagnosis. Swiss medical weekly: official journal of the Swiss Society of Infectious Diseases, the Swiss Society of Internal Medicine, the Swiss Society of Pneumology, 145, w14060.

[7] Bagozzi R P.,Joreskog K G, Sorbom D. (1980). Advances in factor analysis and structural equation models. Journal of Marketing Research, 17(1), 133.

[8] Barrett P. (2000). Latent variable models: an introduction to factor, path, and structural analysis. Personality and Individual Differences, 29(5), 999-1000.

[9] Pritchard JK, Stephens M, Donnelly P. (2000). Inference of Population Structure Using Multilocus Genotype Data. Genetics. 155. 945-59.

[10] Blei DM, Ng AY, Jordan M I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

[11] Saga R, Kunimoto, Rikuto. (2016).LDA-based path model construction process for structural equation modeling. Artificial Life and Robotics, 21(2):155-159.

[12] Guan P, Wang Y F. (2016).Research on the Method of Determining the Optimal Number of Topics in the LDA Topic Model in Scientific and Technical Information Analysis. Modern Library and Information Technology, 32 (9): 42-50.

[13] Liao Liefa, Le Fugang. (2017). Research on patent technology evolution based on lda model and classification number. Modern Information (05), 15-20. (in Chinese)

[14] Kim H., Ku B, Kim J Y. (2016). Confirmatory and Exploratory Factor Analysis for Validating the Phlegm Pattern Questionnaire for Healthy Subjects. Evidence-based complementary and alternative medicine : eCAM.