Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 7, No. 2, 2024

Analysis of Health Data for the Elderly based on Medical Website Mining

Xue Tian [1, *], Xia Wang [1], Ying Li [2]

[1] School of Information Science and Technology, Taishan University, Taian, Shandong, China
[2] Zoucheng Economic and Social Development Research Center, Jining, Shandong, China

*** Corresponding author:** Xue Tian (Email: tianx1122@163.com)

Abstract: The development of Internet medicine provides a convenient platform for doctor-patient communication, and provides an important data source for paying attention to the health of the elderly. In this study, a large number of consultation records obtained from the medical website was screened and cleaned, and the medical thesaurus was generated by training of its own data to solve the problem of inaccurate professional terminology segmentation caused by using default jieba segmentation. At the same time, we use the trained medical thesaurus to conduct topic mining of the consultation data, and it is found that the most concerned problems in the field of elderly health are cerebro-cardiovascular, pulmonary and stomach diseases, so as to provide further medical advice and targeted services.

Keywords: Health of the Elderly; Medical Big Data; Jieba Segmentation; LDA Algorithm.

1. Introduction

Population aging is an important trend in the current social development, and the health of the elderly has attracted much attention [1]. Medical resources have presented significant pressure to a certain extent. With the development of the Internet, online consultation on health issues have become a mainstream trend, as doctors use their fragmented time to reply online, breaking free from the limitations of time and space, and maximize the utilization of medical resources to a greater extent. Therefore, it also provides a data source for studying the health problems of the elderly.

A medical website is a third-party platform for patients and doctors to communicate. When patients can make inquiries, doctors use their spare time to reply. People usually go directly to hospitals for treatment when emergencies occur, so the consultation records on the website reflect more of the daily needs of patients. If we can identify which health problems the elderly are more concerned about, we can provide targeted services to further reduce the pressure on medical resources.

Although most of the consultation records have filled in titles, the topics obtained are not reliable due to word limits and inaccurate expression. Topic mining in natural language Processing (NLP) is the most common technique for mining topics from a large number of records [2]. Topic mining relies on the results of text segmentation, and there are a lot of professional terms in medical texts, which will cause the confusion in segmentation if they are not discriminated. At present, there are two ways to improve the accuracy of word segmentation: improving algorithm and establishing accurate word libraries, both of which are difficult in practice. In this paper, we propose a method of using self-owned text training to train supplementary word libraries.

2. Data Resource and Method

2.1. Data Resource

Select the online consultation data from a large health platform (haodf.com/) in China, select the elderly disease classification, and crawl the last 100 pages of data provided by it [3]. Due to the platform no longer provides the elderly disease classification, a total of 8,797 records were crawled on August 2, 2021. The crawling data included title, disease description, disease, past medical history, medical record summary and disposal suggestions. Medical record summary was selected as the source of corpus for word segmentation training, and medical description as the source of corpus for topic mining. The data was screened, cleaned, and deduplicated, and records with missing data and wrong records caused by incorrect classification were deleted. Finally, 7240 and 8237 data were obtained respectively.

2.2. Method

2.2.1. Chinese Word Segmentation

Jieba segmentation is used to segment the data of medical record summary, which is the reorganization of disease description, which is more accurate and can obtain more professional word segmentation results. However, the thesaurus of Jieba segmentation contains only daily words, making it impossible to recognize the professional terms in medical records, so it is necessary to supplement the thesaurus or generate a professional medical thesaurus. At present, we can find some medical thesaurus on the Internet, but there are some limitations in practical application. On the one hand, these only provide very professional and specific terms, while online disease descriptions have some colloquial expressions in addition to professional terms. On the other hand, its thesaurus is fixed, which is difficult to meet the actual needs of this study. For example, to inquire the use of amoxicillin, various types of this drug were provided in medical thesaurus of Tsinghua University, such as amoxicillin capsules, amoxicillin tablets, etc., so we need to identify words according to our research need. Therefore, appropriate thesaurus should be supplemented or created according to the research purpose. The idea of word segmentation in this paper is as follows: 500 records are randomly selected as test data, and the remaining data are training data. Divide the training group data into groups of 500 pieces, update the custom dictionary after word segmentation in each group, apply it to

the next group to continue word segmentation and update, and so on. After each update of the custom dictionary, evaluate the test component words and obtain evaluation indicators until there is no significant improvement in the indicators [4].

2.2.2. Evaluation Indicators

The evaluation indicators include precision, recall, and F1 score, which are often used in machine learning classification problems. The formulas are as follows.

$$precision = \frac{TP}{TP+FP} \quad (1)$$

$$recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1 - score = \frac{2 \times recall \times precsion}{recall + precsion} \quad (3)$$

In binary classification problems, TP (true positive) indicates that the true category of the sample is positive, and the predicted result is also positive. FP (false positive) indicates that the true category of the sample is negative, and the predicted result is positive. FN (false negative) indicates that the true category of the sample is positive, and the predicted result is negative. TN (true negative) indicates that the true category of the sample is negative, and the predicted result is negative [5]. However, the segmentation problem is not a classification problem, it only includes standard answers and actual segmentation results, so a transformation is needed. We use A and B to represent the standard segmentation result and the actual segmentation result, respectively. A is equivalent to $TP \cup FN$, and B is equivalent to $TP \cup FP$. The converted formulas are as follows.

$$precision = \frac{A \cap B}{B} \quad (4)$$

$$recall = \frac{A \cap B}{A} \quad (5)$$

2.2.3. Topic Mining.

Common topic mining methods include TF/IDF algorithm, LSA/LSI algorithm, LDA algorithm, etc. In this study, LDA algorithm was selected [6]. The topic mining of the medical record description provided by the patients is carried out to grasp the focus of the current health problems common in the elderly, in order to provide suggestions on the problems found. When applying the LDA model, it is necessary to determine the number of topics, and there are usually two indicators when selecting the number: confusion and consistency. The degree of confusion indicates how uncertain a text belongs to a certain topic, so the lower the confusion, the better the classification effect. However, generally, the more topics there are, the lower the degree of confusion, which is not reasonable enough. Consistency indicates the degree to which the text supports the topic, determined by the coherence value. The higher the value, the better the support. In this study, we choose confusion and consistency to jointly determine the number of topics.

3. Results

3.1. Supplementary Medical Library

Table 1. Disease Names

No.	Disease Names	No.	Disease Names
1	COPD	**7**	neoplasia
2	Stripe stove	**8**	Gastric erosion
3	Myocardial Bridge	**9**	ground-glass nodule
4	Helicobacter Pylori	**10**	Inflammatory lesions
5	Adrenal adenoma	**11**	non-atrophic gastritis
6	Intestinal metaplasia	**12**	bradycardia-tachycardia syndrome

Table 2. Medical Terms

No.	Medical Terms	No.	Medical Terms
1	TPO-Ab	**7**	trigeminy
2	colonoscopy	**8**	TI-RADS
3	bigeminy	**9**	T-wave
4	Fasting blood glucose	**10**	Anti-thyroid peroxidase antibody
5	Ceruloplasmin	**11**	Sinus heart rate
6	COVID-19 vaccine	**12**	Myocardial enzyme

Table 3. Drug Names

No.	Medical Terms	No.	Medical Terms
1	Febuxostat	**7**	Montmorillonite powder
2	Mirtazapine	**8**	Febuxostat
3	ilaprazole	**9**	Statins
4	Sigliptin	**10**	Oshitinib
5	Diovan	**11**	Rosuvastatin
6	Diane-35	**12**	Asme

In the process of word segmentation, some terms that do not appear need to be judged manually. There is some difficulty in word segmentation due to its complexity in expression structure, so different strategies are chosen according to different situations. In the process of discrimination, words can be identified through the network, while phrases require specific analysis based on the specific situation. For example, "stomach discomfort" means that there are problems in stomach that need attention, and the appearance of the word "stomach" has reflected the attention of netizens to stomach problems. In later research, using the term "stomach" to explore support is more likely. Benign lesion" refers to a condition where, although a lesion has occurred, the benign condition to some extent reflects better

outcomes. However, if the phrase is divided into two words, the tendency of expression is. changed, so it is more reasonable to split it as a phrase. After discrimination, the added disease names, medical terms, and drug names are shown in the table 3.

3.2. Evaluation Indicators

Jieba was used to segment the test group corpus with default thesaurus to obtain the indicators, among which the accuracy rate was 0.86, recall rate was 0.81 and F1 was 0.83. After adding the medical vocabulary of Tsinghua University, the accuracy rate was 0.86, the recall rate was 0.83, and F1 was 0.85. The improvement degree of the indicators was 0%, 2%, and 2%, respectively. The improvement effect was not significant.

According to the method mentioned above, 6 rounds were set up, and the results for each round are shown in Figure 1. After the 6th round, the indicator results showed an accuracy rate of 0.89, a recall rate of 0.85, and an F1 rate of 0.87, with an improvement of 3%, 4%, and 4%.

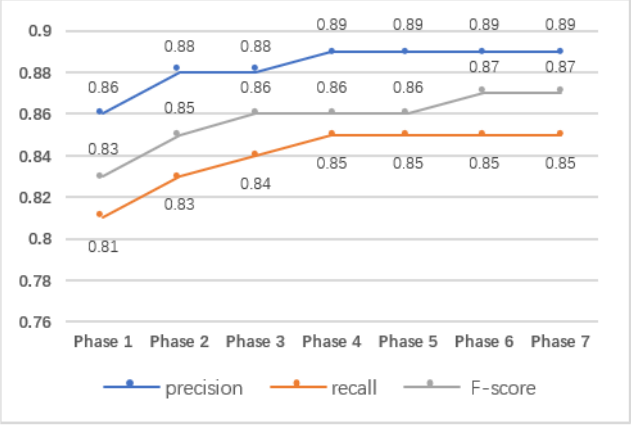

Figure 1. Specific segmentation indicator results

3.3. Topic Mining Results

After updating the vocabulary, topic mining was conducted on 8237 disease descriptions. The number of topics depends on the coherence value. The range of topic numbers was determined to be between 1 and 10 as needed. The coherence values for different topic numbers are shown in Figure 2. The figure shows that when the number of topics is 2, the coherence value is the highest. However, since there are only 2 topics, the confusion level of the topics is very high. Therefore, considering the confusion level and coherence value both, the number of topics is selected as 5.

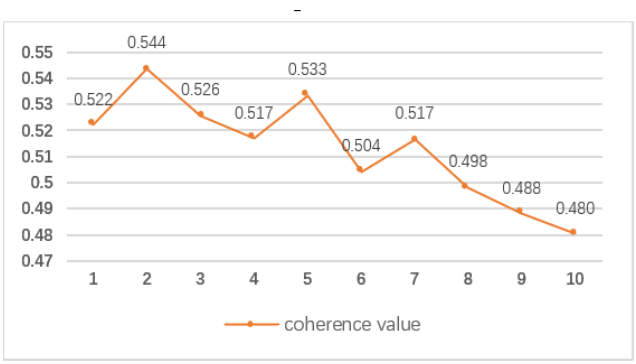

Figure 2. Coherence values under different number of topics

According to 5 topics, we mined the description of the disease and obtained the distribution of topic words as shown in Figure 3. According to the keywords under each topic, topic titles can be got, which are cardiovascular and cerebrovascular diseases, stomach problems, dizziness and headache, treatment process, and lung problems. Topic 4 does not belong to the health concern, so this topic is deleted. Dizziness and headache can be said to be a symptom and also a manifestation of cardiovascular and cerebrovascular diseases, so the topic is merged into cardiovascular and cerebrovascular diseases.

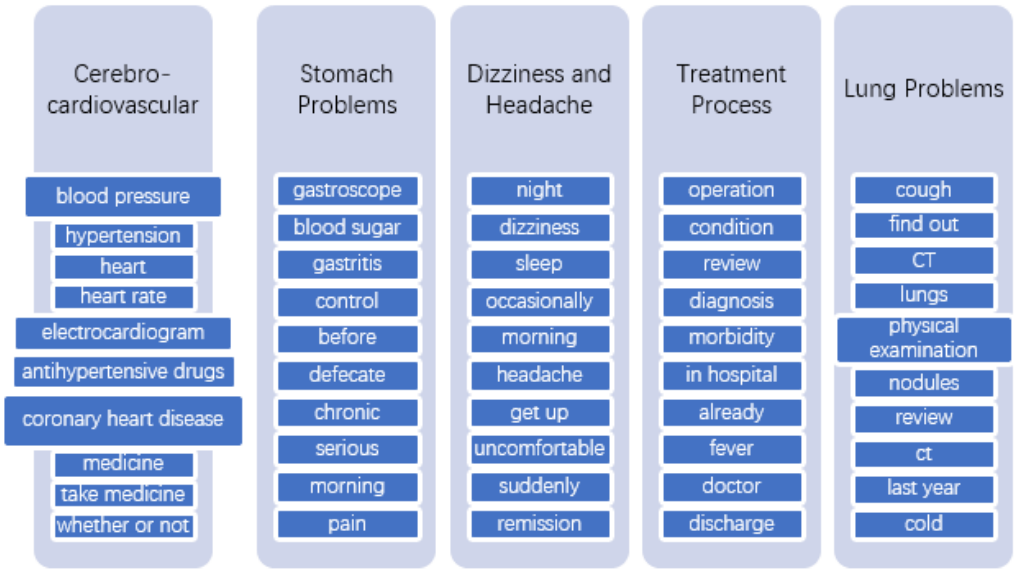

Figure 3. the Results of Topic Mining

4. Discussion

During the process of word segmentation, it was found that training the vocabulary it's our own corpus is effective. The reason is that the health problems concerned by the elderly group are relatively concentrated, and commonly used terms such as treatment methods, medication, and key terms repeatedly appear. Therefore, this segmentation method of "transforming original soup into original food" has a good application effect in this research field.

The health issues that elderly people are concerned about mainly focus on three aspects: cerebro-cardiovascular diseases, stomach, and lungs. Cerebro-cardiovascular diseases are one of the most common diseases among middle-aged and elderly people over 50 years old, and they are also the biggest threat to their health. Most patients need to rely on medication for a long time. Cerebro-cardiovascular diseases are a collective term for cardiovascular diseases and cerebrovascular diseases, including coronary heart disease, hypertension and hyperlipidemia mentioned in the theme words, and ischemic and hemorrhagic diseases that lead to more serious consequences. Stomach diseases have always been a major problem that troubles Chinese people, which is closely related to China's dietary structure and habits, and is becoming increasingly younger. Lung disease is the third leading cause of death, which is closely related to smoking habits and air pollution. The prevalence of COVID-19 further aggravates the problem of lung disease, causing people's concern. The promotion of the Internet provides a broader channel for people to consult and solve health problems. Although most elderly people, especially those over 60 years old, are difficult to skillfully use the Internet to solve problems, and most of their posts are published by their children, it still helps the development of Internet medicine. The results of topic mining are in line with the current concerns in the field of elderly health [7][8][9][10], and verify the rationality of the segmentation results after supplementing the medical lexicon.

5. Conclusion

This paper provides a way to supplement the word library in the field of word segmentation, which improves the accuracy and efficiency of word segmentation. Based on the segmentation results, further exploration of the corpus has identified the main health concerns of the elderly, enabling targeted recommendations and services to be provided for these issues. However, due to the subjectivity in the process of manual word segmentation, there may also be some errors in word segmentation. At the same time, the supplement of stop word library also affects the results of topic mining to a certain extent. These questions will be considered and resolved in the subsequent research.

Acknowledgments

This work was financially supported by the Tai'an City Science and Technology Development Plan Project (2021GX 028).

References

[1] Xia Cuicui, Lin Bao, " International experience in dealing with population aging and implications for China's population policy, " Social Science Journal, vol. 5, 2023, pp.148-157.

[2] Gao Huiying, Liu Jiawei, and Yang Shuxin," Online medical comment topic mining based on improved LDA, " Transactions of Beijing Institute of Technology, vol. 04, 2019, pp. 427-434.

[3] Chen Shuqing, Guo Xitong," Exploring the influence of doctor-patient social ties and knowledge ties on patient selection, " Internet Research, 2021.

[4] Shao Dangguo, Huang Chusheng, Ma Lei," Research on Chinese Word Segmentation in Medical Domain Based on Bi-Lstm, " Communications Technology, vol. 55,2022, pp. 151-159.

[5] Shen Si, Chen Meng, Feng Shuyang, " ChpoBERT: A pre-training model for Chinese policy texts," Journal of the China Society for Scientific and Technical Information, vol. 42,2023, pp.1487-1497.

[6] Yang Jianliang, Liu Yuenan, Qi Tianjiao, "Topic Mining and Evolution Analysis of Public Demands During Major Public Health Events," Library Tribune, vol. 4, 2021,pp.121-131.

[7] Li Mengyu, Lian Juan, Liao Zirui, "Analysis of abnormal detection rate of health examination of elderly people in basic public health service," Chinese General Practice, vol. 26, 2023, pp. 2756-2762.

[8] Han Yingying, Mi Yueping, Cai Bo, "Analysis of death trend of cardiovascular and cerebrovascular diseases among elderly aged 65 and above in Nantong City from 2005 to 2016," Modern Preventive Medicine, vol. 46, 2019, pp. 2062-2065.

[9] Liu Yang, Chen Liping, Gao Ying," The effect of smoking on lung function in middle-aged and elderly people," Chinese Journal of Gerontology, vol. 33, 2013,pp. 4247-4248.

[10] Xu Li, Ge Jing, Yu Peng, " Study on Changes in Chronic Diseases and Comorbidity Patterns among Elderly People in China: Based on China Health and Elderly Care Tracking Survey Data," Chinese General Practice, vol. 27, 2024, pp. 1296-1302.