Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 2
CII 2023

Research on Patient Information Demand Preference in Network Health Community based on TF-IDF Algorithm and BTM Model

Rong Hu [1], Yiting Guo [1], Yunxia Cheng [1], Yuzhu Zhang [1], Shuang Liu [2, *]

[1] Graduate School of Tianjin University of Traditional Chinese Medicine, China

[2] School of Management, Tianjin University of Traditional Chinese Medicine, China

* Corresponding Author

Abstract. With the rapid development of information technology, the network health community has become an important way of doctor-patient communication in the new era. As the starting point of the doctor-patient communication process, patient information needs are an important basis for conducting information communication research in online health communities. In this study, we take the question data of medical patients in the "ask and answer" community from the website called "xunyiwenyao" in 2022 as samples, and the TF-IDF algorithm was applied to draw word cloud map to reveal the hot spots of patient information demand in the secondary department, apply the BTM toic model to identify the information demand topic of patients in the network health community, and extract the information demand preferences of internal medicine patients. The study found that users of patients in online health communities pay more attention to the causes and countermeasures of initial symptoms and uncomfortable symptoms, and prefer to obtain relevant information about the efficacy and side effects of medical products.

Keywords: Network Health Community; Patient Information Requirements; Topic Modeling; Text Mining.

1. Introduction

With the rapid development of the new generation of information technology, the integration of information technology and medical service industry is accelerating[1], which has given birth to a number of network health communities and online medical platforms, including 39 Health network, Xunjian Network, Haofu Online, etc.[2]. In the process of doctor-patient communication, patients can obtain medical information not only through traditional offline medical channels, but also through online health communities[3]. The rich information released by online health communities and online medical platforms provides important value for studying patients' demand preferences for medical and health information[4]. In this study, TF-IDF algorithm and BTM model are used to analyze the hot spots and topic preferences of patients' information demands by taking the data of medical patients published by the internal medicine users of the website called "xunyiwenyao" as an example, so as to prepare for further research on the differences between the supply and demand of doctor-patient communication information.

2. Research Design

2.1. Data Sources

This study selected the online health community called "xunyiwenyao" as the data source, programmed a Python crawler to climb 61,877 problem information published by internal medicine patients in 2022 on this platform, removed 170 blanks, and obtained 61,707 valid information.

2.2. Research Tools

2.2.1. TF-IDF Algorithm

TF-IDF [5] is a typical text similarity measurement method. The higher the frequency of words in the topic, the higher the relevance to the topic. The documents in the corpus are characterized by words, and the corresponding weights of feature words are determined by word frequency TF (Term Frequency) and inverse text frequency IDF (Inverse Document Frequency). TF-IDF algorithm usually adopts the following formula.

$$\mathrm{TF-IDF}(w_i) = \mathrm{tf}(w_i) \times \mathrm{idf}(w_i) = \mathrm{tf}_{d_n}(w_i) \times \log\left(N/\mathrm{df}(w_i)\right)$$

In the formula, $\mathrm{tf}_{d_n}(w_i)$ represents the frequency of the word w_i in the document d_n, N represents the number of all documents in the entire corpus, and $\mathrm{df}(w_i)$ represents the number of documents in which the word w_i appears in the entire corpus.

2.2.2. BTM Model

The BTM model (Biterm Topic Model) is a variant of LDA, the classic probabilistic topic model. It is a topic model proposed by YAN [6] for processing short text, which can mine the potential topic features of short text. BTM model expands the number of words trained by combining the words in the corpus. For example, one of the original short text corpus contents is "Which hospital is better for treating thyroid cancer", according to the BTM idea, there will be three-word pairs of "treatment, thyroid cancer", "thyroid cancer, hospital", "treatment, hospital", and the relevant data used to train the BTM topic model is composed of these word pairs.

3. Preference Analysis of Patients' Information Needs on Online Health Platform

3.1. Text Preprocessing

First, the text data obtained by crawling was cleaned, and the blank content and patient evaluations containing only punctuation, numbers, symbols, emojis, and English were deleted. Secondly, the cut method in Jieba system and the default precise word segmentation mode are used for text segmentation [7]. The results of word segmentation were observed, user dictionaries were added in combination with professional words in the medical and health field, and contents with meaningless preferences for research patient information were deleted.

3.2. Analysis of Patient Information Needs in Network Health Community based on TF-IDF Algorithm

In order to analyse the information needs of patients in each department, this study applied TF-IDF algorithm to conduct word frequency analysis and draw word cloud map. it can be intuitively seen that there are some similarities in the information needs of patients in different departments. The words with high frequency in the questions of patients in different departments of internal medicine mainly include "examination", "hospital" and "reason". Combined with the original content, it can be seen that patients pay more attention to the examination results and the meaning of the indicators of the examination results, the differences in the medical level of different hospitals, the occurrence of symptoms or the causes of illness.

Among the word cloud maps, it can be seen that patients in the endocrine department pay more attention to "nodules", "hyperthyroidism", "hormones", "Blood sugar". Patients in the department of neurology paid more attention to "symptoms", "cerebral infarction", "headache" and "cerebral hemorrhage ". Patients in the respiratory medicine department were more concerned about "COVID-19", "Yang rejuvenation", "lung", "throat" and "bronchitis". Patients in the department of gastroenterology pay more attention to "stomach", "enteritis", "nausea", "gastritis", "acute" and "food". Patients in the renal department were more concerned about "kidney disease", "nephritis", "uremia", "urinary protein" and "kidney stone". Patients in the department of cardiovascular medicine

pay more attention to "hypertension", "coronary heart disease", "angina pectoris" and "myocardial infarction". Patients in the oncology department pay more attention to "cancer", "surgery", "cyst" and "gastric cancer". Patients in the department of hematology were more concerned about "anemia", "purpura", "platelets" and "iron deficiency". Other patients in the internal medicine department pay more attention to "efficacy", "side effects", "sensation", "operation" and "pain".

3.3. Subject Analysis of Information Needs of Patients in Network Health Community based on BTM Model

3.3.1. Determine the Optimal Number of Topics

The number of extraction topics needs to be set in the BTM topic model. Perplexity [8] is a standard index for evaluating the generalization ability of models. Generally, the smaller the perplexity value is, the better the extension of the model and the better the model effect. The lowest degree of confusion, or the point corresponding to k is the best topic number. Therefore, according to the degree of confusion, the optimal number that can explain the user's demand preferences is determined. The confusion degree is calculated as follows.

$$\text{Perplexity} = \exp\left\{-\frac{\sum \ln p(b)}{B}\right\}$$

Where B represents the pair of words. $p(b)$ represents the probability that the BTM model generates word pairs of b.

$$p(b) = \sum_z p(z)p(w_i|z)p(w_j|z) = \sum_z \theta_z \varphi_i|z \varphi_j|z$$

The line chart of confusion degree is drawn according to the results of confusion degree calculation under different topic numbers. As can be seen from Figure 1, confusion degree roughly decreases with the increase of the number of topics, and the decline rate gradually slows down with the increase of the number of topics. When the number of topics k=12, the decline trend of the confusion degree line is obviously slowed down, and it is considered that the data results have good explanation and low redundancy. Therefore, in order to avoid over-fitting and combine linear features, k=12 is used as the subject quantity in this study. The words that embody more than 80% of the characteristics of the topic are counted as the topic characteristic words, so that the topic has a relatively good interpretability.

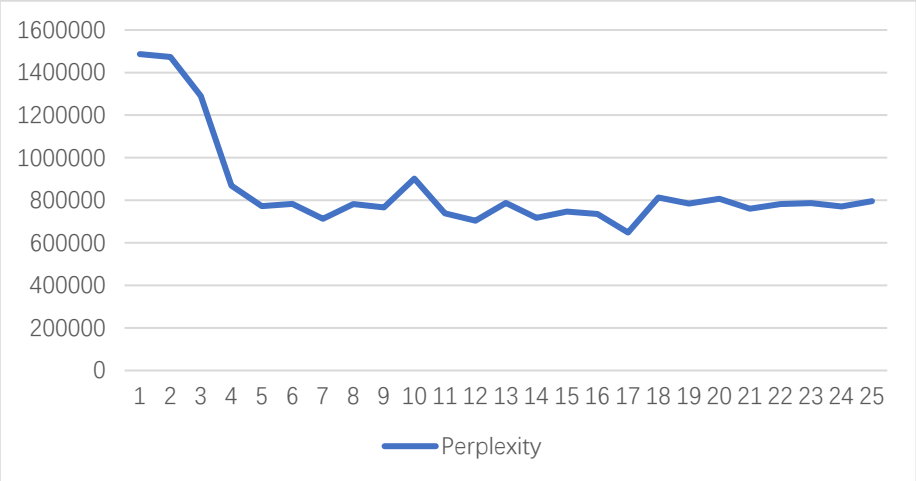

Fig 1. Confusion degree of BTM model under different number of topics

3.3.2. Topic Mining Results

The BTM model was used to conduct topic feature mining on the word segmentation data results obtained in the previous section. Parameters $\alpha=50/k$ and $\beta=0.01$ were selected according to the

empirical value [9], and the number of iterations was set to 100. The topic extraction results and feature words were shown in Table 1.

Table 1. Topic extraction results and feature words.

ID	Topic name	Thematic feature words
1	First visit	excuse me, what, examination, hospital, doctor, know, situation
2	Discomfort symptom	recently, feeling, What, body, symptom, always, occurrence, discomfort, finding, puffy, somewhat, often
3	Anaemia	anemia, what, lacks, iron, sex, examination, bald, doctor, can, hospital, symptom, recently, know, often, serious, body, what, food, now
4	Respiratory discomfort symptom	feeling, cough, throat, chest, throat, breathing, chest tightness, cold
5	Colds and the efficacy of medicines	cold, cough, fever, runny nose, particles, temperature, treatment, cold medicine, efficacy, capsule, taking
6	Nephropathy	purpura, examination, allergic, hospital, treatment, nephritis, kidney, disease, what, symptoms, sepsis, suffer, discover, what, serious
7	lithocyst	check, hospital, recently, body, treatment, kidney, stone, uremia comfortable, disease, suffer, what, doctor, need, now, nephritis, cyst, ialysis, hydronephrosis
8	Inspection result	examination, physical examination, creatinine, urine protein, platelets, body, normal, result, how much, meaning, display, low, routine urine, white blood cells, positive
9	Cardiovascular disease	hypertension, heart, coronary heart disease, acute, blood pressure, thrombosis, angina, heart failure, cardiomyopathy
10	Food and preparation methods	hypertension, can, diabetes, patient, how and what and cause, note, sugar, food, endocrine, disorder, to do, food, symptoms, can't, need, did
11	Medical level and cost	treatment, hospital, which, early, what, Shanghai, injury, best, Excuse me, know, how much, surgery, skin cancer, transfer, cost
12	Indicator meaning	meaning, examination, echo, physical examination, antibody, hormone, globulin, function, calcification, low, blood flow, report sheet, thyroxine

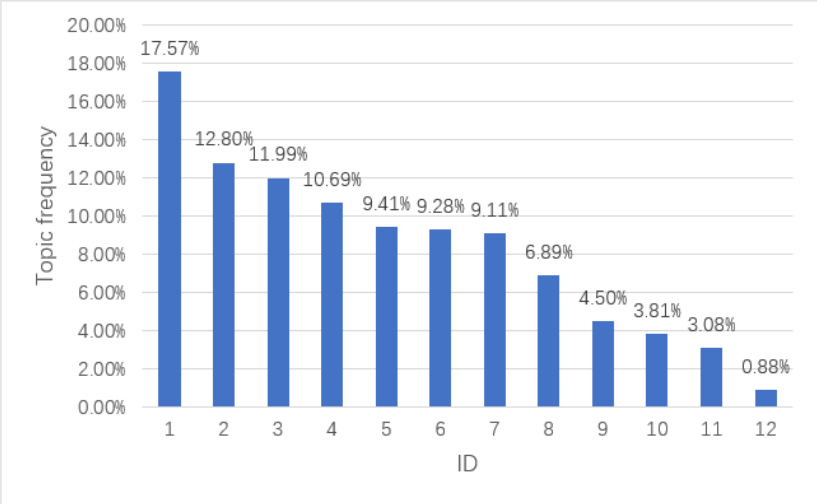

Fig 2. Frequency distribution of the topics

The frequency distribution of different topics is shown in Figure 2. The horizontal coordinate of the bar chart indicates the topic number, and the vertical coordinate indicates the topic distribution frequency.

First of all, the patient users in the network health community prefer to obtain the relevant information of initial diagnosis, understand the causes of discomfort symptoms and coping methods. Secondly, the patient will pay attention to several specific disease topics, such as "anemia", "cold", "kidney disease", "stone", "cardiovascular disease" and so on. Thirdly, the patient users prefer the combination of online and offline consultation, offline medical treatment for physical examination, and online to obtain the meaning of examination results and the level of indicators. Finally, patient users will also expect to obtain and learn about the level and cost of hospital care in the online health community.

4. Conclusion and Suggestions

In this study, by analyzing the special stylological features of the patient's information fragmentation and high noise in the online platform, combined with the study of short texts, a method based on the BTM model is applied to discover the hot spots of the patient's information needs. By using BTM model to model the topic of word pairs in the corpus, the sparsity problem of short text is overcome. And The TF-IDF algorithm is improved to meet the requirement of short text feature extraction. The accuracy and effectiveness of the proposed method are verified by experiments with real question data.

The study found that the online patient side users of the network health community were more concerned about the causes of initial diagnosis and discomfort symptoms and coping methods, and preferred to obtain relevant information about the efficacy and side effects of medical products. According to the characteristics of patient information fragmentation and high value in the doctor-patient communication process in the network health community, this study uses TF-IDF algorithm and BTM model to reveal the preference hotspots of users' information needs in different secondary departments of internal medicine, and identifies the preference topics of patients' information needs.

The focus of this study is to improve the quality of hot spot discovery of patients' information needs. In the process of the experiment, there are some problems, such as the non-standard medical terms of online health platform users and the slow modeling speed of BTM. Therefore, in the subsequent research, we will consider how to improve the efficiency of modeling, improve the medical vocabulary dictionary, and build a supervised topic mining model to adapt to the application of massive data.

References

[1] CNNIC. The 49th China Internet Development Survey Statistical Report [EB/OL]. [2021-09-15]. http://cnnic. cn/ gywm/ xwzx/rdxw/20172017_7086/202202/t20220225_71725.html.

[2] Lu Quan, Zhu Anqi, Zhang Jiyue, et al. Research on user information needs mining in Chinese online health community: A case study of tumor data of Qiuyi.com. Modern Library and Information Technology 4,22-32(2019).

[3] Zhao Dongxiang. Review of research status of online health community in China . Library and Information Work 62 (9):134-142(2018).

[4] Dong Wei, Tao Jinhu. Research on User interest group identification in online health Community based on Topic preference: A case study of Medicine.com . Information Science 39(3),88-93(2021).

[5] Salton G. The SMART Retrieval System-Experiments in Automatic Document Processing. Englewood Cliffs, New Jersey: Prentice Hall Inc, (1971).

[6] YAN X, GUO J, LAN Y, et al. A biterm topic model for short texts Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. Rio de Janeiro: ACM, 1445−1456(2013).

[7] Min Yen Wu, Chih-Ya Shen, En Tzu Wang, et al A Deep Architecture for Depression Detection Using Posting, Behavior, and Living Environment Data. Journal of Intelligent Information Systems, 54(2),225-244(2020).

[8] Cao J, Xia T, Li J, et al. A density-based method for adaptive LDA model selection. Neurocomputing, 72(7),1775-1781(2009).

[9] Hao H J, Zhang K P, Wang W G, et al. A Tale of Two Countries: International Comparison of Online Doctor Reviews between China and the United States. International Journal of Medical Informatics, 99(3),37−44(2017).