ISSN: 2664-9640

Research on data analysis of online health community platform based on Text Mining

Tao Wang [1], Zhichao Sun [2], Yi He [1], Huiwen Guo [3,*]

[1]International School of Hanyang University, Seoul, 04763, South Korea

[2]Business School of Changwon National University, Changwon,51140, South Korea

[3]Shandong Yantai lezhan Technology Co., Ltd, YanTai, 264009, China

These authors are co-authors

*Corresponding Author : oucghw@163.com

Abstract

Internet-based social media provide many channels for doctors, patients and nurses to obtain and exchange information. Online health community platform is professional. Theoretically, this study applies the theory and method of machine learning to online medical text data mining, and expands the extraction theory and method of medical health information. From the practical application aspect, this study deeply digs the data value of online health community platform. On the one hand, it can provide information retrieval ways for public self-diagnosis and self-treatment; On the other hand, it can analyze and summarize the evolution trend of health hotspots in China. Firstly, this paper reveals the characteristics of sub-languages of Chinese online medical texts, which lays a foundation for the construction of online medical information extraction methods based on machine learning.

Keywords

Online medical,text mining,data analysis.

1. Theoretical Basis

1.1. The Background Of The Topic

In 2020, the epidemic in novel coronavirus (COVID-19 for short) broke out on a large scale and quickly swept the world.

On the one hand, the outbreak of COVID-19 epidemic has brought heavy losses to the global economy. In Asia, for example, the relevant data of the Asian Development Bank show that the government revenue has dropped sharply due to the slowdown of industrial and commercial development, and in order to cope with the persistent impact of the epidemic on health, economy and society, the government expenditure of various countries has increased significantly. As a result of both revenue and expenditure factors, by the end of 2020, the fiscal deficits of many Asian countries have reached double digits.

On the other hand, the epidemic has forcibly changed people's living habits and consumption demands, thus giving birth to new markets and new formats and accelerating industrial iteration. For example, online office collaboration, live broadcast retail, community group purchase, "cloud life" and other production and life styles that rely on the Internet are in the ascendant, constantly catering to the change of consumers' preferences. In addition, the vigorous development of online medical community is also a major change brought about by the outbreak of the epidemic. The epidemic has greatly reduced the mobility of people, and one of its impacts is that patients turn to online consultation. The major hospitals in China have

ISSN: 2664-9640

launched online consultation clinics one after another. In addition, the online consultation platform of Internet+medical treatment launched by Internet companies has greatly increased the number of consultations, which has become an important substitute for traditional medical treatment.

As early as 2016, "Healthy China 2030" Planning Outline issued by the Central Committee of China and the State Council put forward "developing health services based on the Internet and encouraging the development of health services such as physical examination and consultation", which affirmed the development of medical care in "internet plus". In 2018, the General Office of the State Council issued the Opinions on Promoting the Development of "Internet+Medical Health", which further guided and standardized the "Internet+Medical Health" industry. In 2020, in response to the COVID-19 epidemic, the National Medical Insurance Bureau and the National Health and Health Commission issued the Guiding Opinions on Promoting the Medical Insurance Service in internet plus during the Prevention and Control of COVID-19 Epidemic, which clearly stated that the eligible medical service expenses in internet plus should be included in the scope of medical insurance payment. The introduction of this guidance is an important driving force to encourage people to participate in "internet plus" medical treatment.

Obviously, the practice of "internet plus" medical treatment can weaken the obstacles of time and space in people's visits, thus improving the efficiency of medical services to a certain extent. More importantly, with the help of the Internet platform, interrogation, an interactive process, can generate and save a large amount of text data. In the information age, it is of great practical significance to dig and analyze the real data. The Guiding Opinions on Promoting and Standardizing the Development of Big Data Application in Health Care issued by the General Office of the State Council clearly pointed out that "integrating public information resources in social networks" and "strengthening key technologies such as cleaning, analysis and mining of massive data storage in health care, and protection of security and privacy". Personally, the mining and integration of online health community platform information is helpful for patients to acquire targeted medical health knowledge. As far as society as a whole is concerned, using machine learning model to analyze data is of great value in disease early warning.

1.2. Research Purpose

In this study, the text information of online health community platform is taken as the data source, and data mining and analysis are carried out with the help of machine learning model, aiming to achieve the following goals.

This paper reveals the sub-language features of online medical information text, and analyzes the influence of the sub-language features on online medical information extraction.

Based on online medical information extraction, we can identify health hot topics, and then classify and retrieve effective information to improve the efficiency of information retrieval.

Taking health hotspots as the research object and clustering method as the basic tool, the evolution path of health hotspots with time is identified.

1.3. Theoretical Significance And Practical Application Value

Internet-based social media provides many channels for doctors, patients and nurses to obtain and exchange information, among which the online health community platform represented by "Sanjiu Health Network" and "Lilac Garden" has unique advantages. First, the online health community platform is professional. Compared with comprehensive platforms such as Tianya Community and Sina Blog, the former is characterized by high specialization and strong pertinence. Second, because the online health community platform has interactive function, members will have user stickiness due to their own historical data iteration and emotional connection among members, which makes the information more diverse, long-term and

ISSN: 2664-9640

comprehensive. Therefore, mining and analyzing the text data of online health community platform has the following two meanings.

(1)Theoretically, this study applies the theory and method of machine learning to online medical text data mining, and expands the theory and method of medical health information extraction.

(2)In practice, this study deeply explores the data value of online health community platform. On the one hand, it can provide information retrieval for public self-diagnosis and self-treatment; On the other hand, it can analyze and summarize the evolution trend of health hotspots in China.

1.4. Possible Innovation

The possible innovations of this study are reflected in the following two aspects.

(1)To reveal the features of the sublanguage of Chinese online medical texts. As far as the linguistic features of Chinese texts are concerned, domestic scholars have conducted in-depth research in clinical texts, news texts and Weibo texts. However, the research on online medical texts is still in its infancy. Firstly, this paper reveals the features of Chinese online medical texts, which lays a foundation for the construction of online medical information extraction method based on machine learning.

(2)Text clustering method is used to identify the hot topics in the network health community. Traditional text analysis mostly adopts manual statistical annotation method based on grounded theory, which is interfered by subjective factors and inefficient. In this study, the text clustering method is applied to the health topic recognition in online community, which improves the efficiency and accuracy of recognition.

2. Literature Review At Home And Abroad

In 2005, Simpson, an American scholar, used the data of American health care database to analyze the gaps in insurance coverage, medical utilization, medical expenditure and medical quality of children from different income families in the United States. which opened up a new research field of medical and health big data analysis.

The research objects of health big data can be divided into the following three categories. The first is hospital medical big data, which is the relevant data generated in the routine clinical diagnosis and treatment and scientific research of hospitals.. For example, Chen Donghua (2020). Focusing on the related methods of medical text semantic analysis for decision support, we can dig deep into the text information in patients' electronic medical records, and apply it to clinical decision support, personal health management and other fields. Second, medical research data based on a large number of people, such as large sample survey data, disease monitoring data and cohort study data. The typical research is CANHEART team in Canada (2015). A 20-year follow-up study was conducted on 9.8 million adults in Ontario, hoping that the results of project data analysis could improve the cardiovascular level of their citizens. The third is medical big data based on interconnection, which is mainly composed of two parts.: (1) self-health management data generated by health monitoring equipment, and (2) Internet data resources generated by network-based physician consultation and drug purchase. Thanks to the rapid development of the Internet and big data analysis and processing technology, the third category of research objects is becoming increasingly important in the field of big data medical health research, and its research results are mostly used for daily life health monitoring and personalized disease prevention..

The embryonic form of a healthy network community is an online support group that was active on the Internet in the early days. Researchers found that the participating members of this online support group are composed of people with different roles, including doctors, patients,

ISSN: 2664-9640

caregivers, social workers and psychological counselors, etc. The above members can be roughly divided into two types.: First, information consultants represented by patients and their caregivers, who are also the subjects participating in the network health community; Second, information providers represented by various medical experts. Although the proportion of this group in the overall membership is not high, they can usually provide more valuable and trustworthy information. In view of the research topic of the network health community, the existing research covers the functions of the network health community, discussion topics. And emotional support. And so on.

In terms of research methods, early research on the network health community mostly adopts questionnaire survey. And interviews. And so on. With the rapid development of the network health community, some scholars have done quantitative content analysis through manual annotation and statistical analysis of posting texts in the community. The methods based on questionnaire survey and interview are restricted in both sample size and survey content, so the evaluation results are difficult to objectively and comprehensively reflect the development status of online healthy communities. Although the method of manual annotation takes the actual published content in the online community as the research object, it consumes huge time and labor costs, which is increasingly infeasible in the information age when the text content explodes and grows. In recent years, the extensive application of text mining technology provides an efficient and convenient tool for text information processing in the network healthy community. E.g. Bekhuis(2011). Using natural language processing tools, we extract relevant medical vocabulary from the posting texts of online healthy communities, and classify the community posts based on this. Chen（2012）Cluster analysis is used to cluster the posting texts in different healthy communities, and it is found that the hot topics discussed in different healthy communities are different.

From the above documents, we can see that there is a lot of information hidden in the relevant text data in the network health community platform, which is worthy of further discussion. Although the traditional manual processing method can effectively and accurately discriminate information, it is inevitably inefficient and lack of scientificalness when processing a large amount of data. Machine learning and text mining technology provide a sharp weapon for text processing. However, as far as the text content in the online community health platform is concerned, because it has the characteristics of daily life and medical specialization, it is of great practical significance to choose an effective machine learning model to analyze it.

3. The Research Content

3.1. Academic ideas and ideas

In this study, the relevant methods and technologies of medical text semantic analysis are taken as the basic tools, and the text data of online health community platform is taken as the research object, and the analysis methods such as system science analysis, big data analysis technology and machine learning model are comprehensively adopted to construct a medical information extraction framework for online health community platform, and further realize the research of health hotspot identification based on information extraction.

3.2. Main research contents.

This study will be carried out in detail from the following angles:

(1)Constructing the information extraction framework of online health platform. On the basis of machine learning theory, network information extraction theory and basic theory of medical platform, the characteristics of online medical sub-language text are identified, and an information extraction method of online health platform based on machine learning is established.

(2)Construct an automatic recognition framework for health hot topics. The main body of community text is effectively represented by special sets such as n-gram features and domain-related features, and the text of online healthy community is divided into different clusters by combining with text clustering technology. Finally, the health hot spots are effectively identified by extracting keywords from the clusters.

(3)Identify the dynamic evolution path of health hotspots over time. By fitting the evolution path of health hotspots with time with statistical methods, the healthy development trend of community citizens is analyzed.

3.3. Key technologies or problems to be solved

The key technologies to be solved in this study are embodied in the following two aspects:

Combining natural language processing technology with the relevant knowledge base in medical field, this paper explores related methods to better structure and standardize narrative texts in online health community platform, so as to improve the accuracy of text information extraction and analysis.

Build an effective machine learning model parameter initialization and parameter adjustment, so as to build an information extraction model framework of online health community platform.

3.4. Research methods to be adopted.

Based on the research purpose, this study comprehensively uses the relevant theories and methods of management science, information science and system science to mine and analyze the text data of online healthy community platform. The main research methods used include:

Literature research method. By referring to a large number of related literatures on online medical information extraction at home and abroad, we can understand the frontier trends in this field, and master the research status and development trend in this field.

Big data analysis method. The related data of online health community platform is characterized by large volume, multiple dimensions and complex structure. MapReduce model can solve the problem of large-scale data set association.

Machine learning method. The information extraction model framework of online health community platform based on machine learning is constructed and refined to identify medical entities and health hotspots.

References

[1] Data Sources: ASIAN DEVELOPMENT BANK: BASIC 2021 STATISTICS.

[2] Simpson L, Owens P L, Zodet M W, et al. Health Care for Children and Youth in the United States: Annual Report on Patterns of Coverage, Utilization, Quality, and Expenditures by Income[J]. Ambulatory Pediatrics, 2005, 5(1).

[3] Yu Guopei, Bao Xiaoyuan, Huang Xinting, et al. Types, nature and related issues of medical and health big data [C] / / 2014 China hospital information network conference. Information Management Professional Committee of China Hospital Association; China Digital Medicine, 2014.

[4] Chen Donghua. Research on semantic analysis method of medical text for decision support [D]. Beijing Jiaotong University, 2020

[5] Tu J V, Chu A, Donovan L R, et al. The Cardiovascular Health in Ambulatory Care Research Team (CANHEART): Using Big Data to Measure and Improve Cardiovascular Health and Healthcare Services[J]. Circulation Cardiovascular Quality & Outcomes, 2015, 8(2).

[6] Liu Ning, Chen min. research on application topics and related data sources of medical and health big data [J]. China Digital Medicine, 2016, 11 (8): 6-9

[7] Kim T W, Park K H, Yi S H, et al. A Big Data Framework for u-Healthcare Systems Utilizing Vital Signs[C]// International Symposium on Computer. IEEE, 2014.

[8] Finn J. An exploration of helping processes in an online self-help group focusing on issues of disability[J]. Health and Social Work, 1999, 24(3): 220–231.

[9] Richter J G, Becker A, Schalis H, et al. An ask-the-expert service on a rheumatology web site: who were the users and what did they look for? [J]. Arthritis Care and Research, 2011, 63(4):604-611.

[10] Durant K T. Identifying Temporal Changes and Topics that Promote Growth within Online Communities: A Prospective Study of Six Online Cancer Forums [J]. International Journal of Mathematical Modelling and Algorithms, 2011, 2(2):1–22.

[11] Culver J D, Gerr F, Frumkin H. Medical information on the Internet: a study of an electronic bulletin board [J]. Journal of General Internal Medicine, 1997, 12(8):466–470.

[12] Gooden R J, Winefield H R. Breast and prostrate cancer online discussion boards. A thematic analysis of gender differences and similarities [J]. Journal of health psychology, 2007, 12(1):103–114.

[13] Swan M. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking [J]. International Journal of Environmental Research and Public Health, 2009, 6(2):492–525.

[14] Cho J, Noh H I, Ha M H, Kang S N, Choi J Y, Chang Y J. What kind of cancer information do Internet users need? [J]. Support Care Cancer, 2011, 19(9):1465-1469.

[15] Castleton K, Fong T, Wang-Gillam A, Waqar M A, Jeffe D B, Kehlenbrink L, Gao F, Govindan R. A survey of Internet utilization among patients with cancer [J]. Support Care Cancer, 2011, 19(8):1183-1190.

[16] Rodgers S, Chen Q. Internet community group participation: Psychosocial benefits for women with breast cancer [J]. Journal of Computer Mediated Communication, 2005, 10(4).

[17] Klemm P, Wheeler E. Cancer caregivers online: Hope, emotional roller coaster, and physical emotional psychological responses. Computer Information Nursing, 2005,23(1): 38–45.

[18] Colineau N, Paris C. Talking about your health to strangers: understanding the use of online social networks by patients [J]. New Review of Hypermedia and Multimedia, 2010, 16(1-2):141-160.

[19] Armstrong N,Powell J. Patient perspectives on health advice posted on Internet discussion boards: a qualitative study [J]. Health Expect, 2009, 12(3):313–320.

[20] Schultz P N, Stava C, Beck M L, Vassilopoulou-sellin R. Internet message board use by patients with cancer and their families [J]. Clinical Journal of Oncology Nursing, 2003, 7(6):663–667.

[21] Attard A, Coulson N S. A thematic analysis of patient communication in Parkinson's disease online support group discussion forums [J]. Computers in Human Behavior, 2012, 28(2): 500-506.

[22] Bekhuis T, Kreinacke M, Spallek H, Song M, O'Donnell J A. Using natural language processing to enable in-depth analysis of clinical messages posted to an Internet mailing list: a feasibility study [J]. Journal of Medical Internet Research, 2011, 13(4): 98.

[23] Chen A T, Exploring online support spaces: Using cluster analysis to examine breast cancer, diabetes and fibromyalgia support groups [J]. Patient Education and Counseling, 2012, 87(2): 250–257.