Analysis of Epidemiological Characteristics of COVID-19 Based on Web Data Mining

Zhaobo Zeng[a], Kai Tang, Ke Xing, Yingyin Wu, Haisong Xiong

Computer Science and Technology, Beijing Jiaotong University, Weihai, Shandong, 264401

[a]duoerxs@163.com

Keywords: COVID-19, data Mining, Epidemiology, Social Media

Abstract: Based on Selenium data mining technology, the epidemiological characteristics of real help-seeking cases in Sina Weibo were obtained through the analysis of 690 effective cases in Sina Weibo "pneumonia patients seeking help" from February 4 to 22, 2020. The study found that 97.6% of the patients seeking help came from Wuhan, focusing on central urban areas such as Wuchang, Dengkou and Hanyang, which was proportional to local medical resources and population density. The cases of Weibo seeking help were mainly distributed from February 4 to 7. With the relief of the shortage of medical resources, the number of cases seeking help through Weibo decreased significantly. The diagnosis date of the patients seeking help was mainly distributed from January 16 to February 6, which was basically consistent with the case distribution released by the Chinese Centers for Disease Control and Prevention. The median age distribution of patients seeking help is 60 years old, which is significantly higher than the data released by the Chinese Centers for Disease Control and Prevention, but it is basically consistent with the data of Wuhan Central Hospital. The results show that for major sudden infectious diseases, social media such as Weibo not only play a role in the spread of public opinion, but also play an important role in epidemiological analysis. Based on the real-time and extensiveness of social media, combined with data mining and big data analysis, it is helpful for decision-makers to quickly grasp the real situation.

1. Introduction

COVID-19 (novel coronavirus pneumonia, NCP), or "COVID-19" for short, refers to the pneumonia caused by 2019-nCoV infection. Since December 2019, a number of cases of unexplained pneumonia have been found in some hospitals in Wuhan, China, all of which were diagnosed as viral pneumonia. The main clinical manifestations were fever, fatigue, dry cough and so on. It has been confirmed that it is an acute respiratory infectious disease caused by 2019-nCoV infection. After the outbreak in Wuhan, the epidemic quickly spread to the whole country in a short time. On January 7, 2020, the Chinese Centers for Disease Control and Prevention isolated COVID-19 from a patient's throat swab sample, which was later named COVID-19 by WHO [1].

Since the first batch of novel coronavirus infection was found in Wuhan in December 2019, the infected population of COVID-19 has spread rapidly with the help of the Spring Festival transportation, and quickly spread to the whole country. As of February 20, 2020, more than 70, 000 cases of infection have been found in the country. At present, relevant scholars and scientific research institutions have carried out research and analysis on COVID-19 virus [2]. China has basically mastered the etiology, epidemic characteristics and pathogenic mechanism of COVID-19 virus, which not only provides a scientific basis for the formulation of epidemic prevention and control strategies and measures, but also provides an important reference for the global community to understand novel coronavirus. In the fields of genetics and virology, researchers have carried out research on the sequence of virus genome, which has accelerated the progress in tracing the source of the virus, preventing diseases, developing vaccines and so on. In the aspect of epidemic situation prediction and evaluation, by establishing SEIR model and adding different parameters for simulation analysis, it is not only proved that the model analysis is basically consistent with the real

Published by CSP © 2020 the Authors

performance of the development of the epidemic situation, but also further confirmed the effectiveness of the prevention and control measures for 2019-nCoV pneumonia, which has a good guiding value for the prevention and control of the epidemic situation. It is gratifying that China has made great progress in the research and development of COVID-19 drugs. A variety of drugs play a positive and effective role in the treatment of COVID-19 patients, accelerating the pace of our country to overcome the epidemic [3].

It is worth noting that most of the above research results focus on a certain aspect of COVID-19's transmission model, epidemiological characteristics, etiology and pathology, treatment and nursing, and a large part of the data comes from the data published on the official website of the national or local health commission, and the source of the data is single. The most important thing is that in the early stage of the epidemic, the government is unable to quickly obtain real and effective data, which is not conducive to the prevention and control of the epidemic, nor is it conducive to the implementation of effective measures. As one of the seven strategic emerging industries identified by the State Council in the 12th five-year Plan, the importance of the new generation of information technology in traditional epidemiological research has not been shown.

On the other hand, in recent years, the development of social media in China has attracted attention, with the emergence of social software such as Wechat, Weibo, Douyin and so on. The platform of social media software with the help of the Internet covers all forms of network services with human social interaction as the core, helping the Internet to expand from research departments, schools, governments, business applications and other platforms to everyone. At the same time, it has also led to the explosive growth of social data. There is a lot of information and knowledge between social media data, and can be widely used in a variety of application scenarios, including business management, production control, engineering design, market analysis and scientific exploration. However, how to extract and use this information effectively has become a great challenge. In order to solve this problem, the network data mining technology of directionally crawling related web page resources arises at the historic moment. Network data mining can automatically grab the information of the World wide Web through programs or scripts according to certain rules, and realize the data resource analysis of related web pages. At present, data mining technology is widely used in electric power, economy, communications, people's livelihood and other fields [4].

In this paper, through Selenium data mining technology, from social media (Sina Weibo) to obtain effective information about COVID-19 's help-seeking cases. Then, using the data set, the epidemiological characteristics of COVID-19 were analyzed from four aspects: the geographical space of the patients seeking help, the number of people seeking help, the date of diagnosis and the age of the patients. Finally, combined with the extensive and real-time characteristics of social media, the effectiveness of various policies of the management department during the occurrence of major infectious diseases were discussed and evaluated.

2. Materials and methods

This paper is mainly based on the open source Web application Selenium testing tool, combined with the collection program written by Python, in the case of simulating the operation of the browser, to realize the automatic collection of COVID-19 case data on the social media tool.

During the COVID-19 epidemic, Sina Weibo quickly became an important platform for the public to understand the dynamics and trend of the epidemic. On average, more than 200 million netizens use Sina Weibo every day to follow the latest information on the epidemic, get epidemic prevention and control services, and participate in public fund-raising. The number of epidemic topics on Weibo is growing. As of February 22, 2020, a total of 880000 individual certified users had posted 16.88 million Weibo posts, including medical care, science popularization and other fields.

In response to the COVID-19 epidemic, Sina Weibo officially released a "super call for help for pneumonia patients" on February 4, 2020, and relevant government departments have also set up special channels to verify and dock with people seeking help. As of Feb. 22, the super topic had collected 1,222 posts, followed 575000 followers and read more than 2.9 billion people. Based on the social media data mining method, this paper selects the help information posted on the help area of Sina Weibo "help for pneumonia patients" as the object to study the epidemiological characteristics of help cases on social media during the epidemic of COVID-19. Among them, the help-seeking super call contains the help-seeking personnel information field as shown in Table 1.

Table.1. Helper Information Field

Serial number	Field
1	Name
2	Age
3	The city where it is located
4	Community
5	Time of illness
6	Disease description
7	Contact information

As of Feb. 22, a total of 690 effective help-seeking cases were obtained from Weibo "help for pneumonia patients" by data mining method. The information data of some people seeking help is shown in Table 2. In view of the fact that the focus of this study is the analysis of the epidemiological characteristics of COVID-19's case. Therefore, the information of help-seeking personnel mainly focuses on “age, city, community, community, time of illness, and date of help”. It should be noted that the time of illness in the microblog super-talk specifically refers to the time of diagnosis of the patient, and most of the patients who have asked for help have provided detailed confirmation of the diagnosis. In addition, in order to protect personal privacy, the final data set of this study hides the information such as the name of the person seeking help, specific residential address, contact information and detailed disease description.

3. Result

3.1 Regional distribution of patients seeking help

Table.2. The personnel information field contained in help-seeking super call

Serial number	Region name	Number of people seeking help	Area / km2	Resident population / 10,000	Population density
1	Outside the province	4	/	/	/
2	In the province	12	/	/	/
3	Jianghan	55	28.29	68	2.4036
4	jiangan	118	70.75	100	1.4234
5	qiaokou	101	41.46	62	1.1727
6	wuchang	120	107.76	126.37	0.5313
7	hongshan	106	220.5	117.16	0.6710
8	qingshan	44	80.47	54	0.5851
9	hanyang	86	111.54	65.27	0.0502
10	huangpo	8	2256.7	113.23	0.1170
11	xidonghu	15	499.71	58.48	0.0426
12	caidian	4	1093.57	46.66	0.0699
13	xinzhou	7	1500.66	105	0.0452
14	jiangxia	10	2018.3	91.37	0.0124

The detailed regional statistics of 690 effective help-seeking patients based on social media data mining are shown in Table 2. It can be seen from Table 2 that although the "Super call for help of pneumonia patients" is open to all members of the public, the final statistics show that only 4 cases come from outside Hubei Province, only 12 cases come from Hubei Province, but do not belong to Wuhan City. The vast majority (97.6%) of the patients who asked for help came from Wuhan. On the other hand, the number of social media patients seeking help varies significantly in different regions. Wuchang, Jianghan, Hanyang, Dengkou, Jiangan and Hongshan are concentrated areas for seeking help, while Hannan and Jiangxia are less.

In order to make a further quantitative analysis of the geographical distribution information of patients seeking help, Table 3 also records the data of geographical area and resident population of each region of Wuhan, and expresses the population density by defining the ratio of resident population to geographical area. By comparing the distribution of population density in different regions of Wuhan, it can be found that except for the abnormal data in Jianghan District because of the large population density, there is a significant positive correlation between the number of patients seeking help in other regions and the population density in this region.

3.2 Time series analysis of patients seeking help

Based on the data of 690 effective help-seeking patients obtained by social media data mining, the daily distribution of help-seeking patients over time is shown in figure 2.

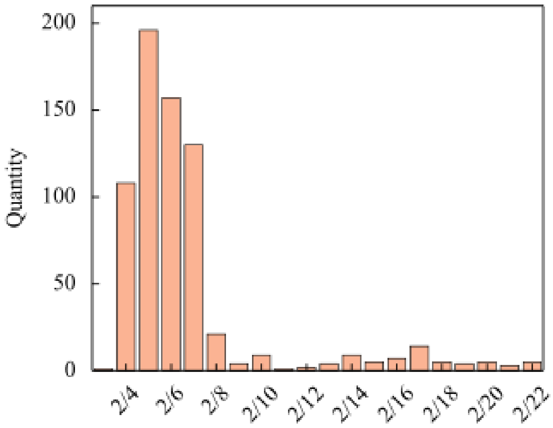

Figure 1. The number of patients seeking help is distributed over time

It can be seen from figure 1 that the daily help-seeking patients are mainly distributed from February 4 to 7, in 2020. During this period, the average daily number of help-seeking patients is more than 100, and the peak of the number of help-seeking patients is on February 5. The number is close to 200. On the other hand, since February 8, the number of requests for help from patients with pneumonia has dropped sharply, with an average of no more than 20 times a day.

3.3 Distribution of diagnosis date of patients seeking help

Figure 2 shows the distribution of diagnosis time of patients seeking help every day. It can be seen from the figure that the diagnosis time of patients seeking help is mostly from the middle of January to the day of seeking help from February 4 to 7. As can be seen from the blue dotted line in the picture, with the passage of time, the number of patients with early diagnosis is getting smaller and smaller, which is basically in line with the national policy requirements of "all receivables, Brooks no delay." it also shows that the vast majority of patients have been effectively assisted and arranged.

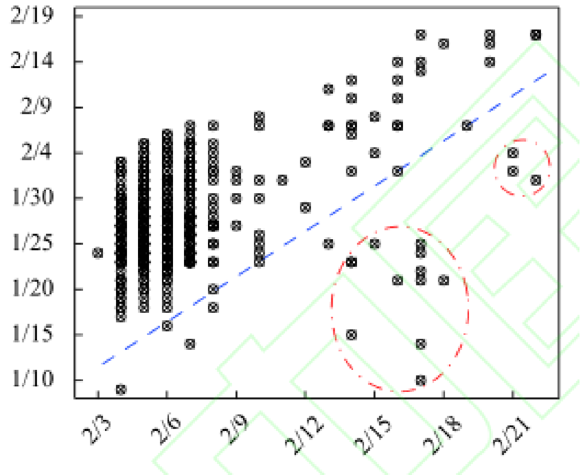

Figure 2. Daily date distribution of patients seeking help

However, since February 13, from the red oval area in figure 6, there have been a number of requests for help from patients with early diagnosis. Through the further analysis of the information of 14 patients in the red area, it is found that these patients can be divided into two groups: one group is the plasma of patients who have been hospitalized normally to seek help, and the other group is that there are other basic diseases to seek treatment. During the fight against COVID-19, in addition to the diagnosis and treatment of patients, another important work was the research on COVID-19's new drugs and new treatments. From the data of patients seeking help from social media pneumonia, it can be found that "plasma therapy" has a certain impact on the medical front line. On the other hand, during the epidemic of major infectious diseases, people with other underlying diseases are often susceptible to infection because of low immunity. In particular, patients receiving chemotherapy have become susceptible to novel coronavirus. Patients can only be monitored at home and treated with drugs. In addition, during the epidemic, a number of cancer hospitals in Wuhan were requisitioned to fight the epidemic, resulting in the delay of chemotherapy for some patients, so these patients sought help through social media.

3.4 Age distribution of patients seeking help

Figure 3 shows the age distribution of 690 active patients seeking help from social media. According to the analysis from the chart, most of the patients seeking help are in the age group of 50 to 80 years old (71.88%), and the median age distribution is 60 years old. This age distribution is basically consistent with the distribution characteristics of COVID-19 confirmed cases released by the Chinese Centers for Disease Control and Prevention. As can be seen from the picture, elderly patients are more likely to be infected by novel coronavirus.

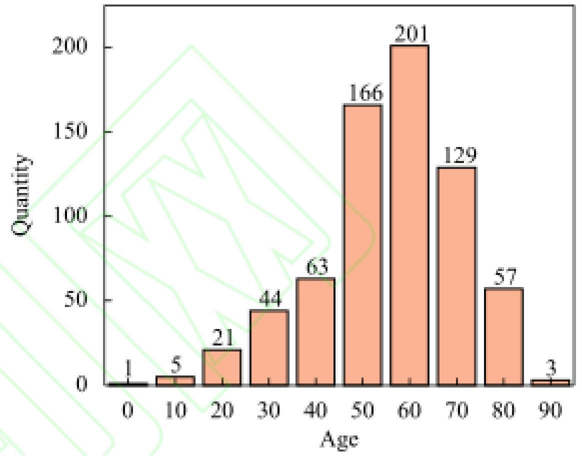

Figure 3. Age distribution of patients seeking help

4. Conclusion

Based on Selenium data mining technology, 690 valid cases in the help area of "help for pneumonia patients" on Sina Weibo from February 4 to 22 were obtained and analyzed from four aspects: the geographical space of the patients, the number of people seeking help, the date of diagnosis and the age characteristics of the patients. This paper finds that the results of real and effective case analysis obtained on social media show that the main reason for COVID-19's rapid spread is the shortage of medical resources, resulting in inefficient emergency management and causing people to panic. After meeting the demand for medical resources such as medical teams and hospitalized beds, the number of patients seeking help has been significantly reduced, and the epidemic situation has been effectively curbed. On the other hand, it is found that the information of patients seeking help during the epidemic can be visually presented and statistically analyzed through social media, which can obtain its epidemiological characteristics more effectively and timely. In the next step of work, the transmission model of infectious diseases can be introduced on this basis to effectively analyze the spatio-temporal evolution and spread characteristics of the epidemic, so as to provide important data reference for further supporting Wuhan and sniping the spread of the epidemic.

To sum up, for major sudden infectious diseases, social media not only plays a role in public opinion, but also plays an important role in epidemiological analysis. Relevant technical departments can make full use of the universality and timeliness of social media to obtain effective data cases through social media, and then combine methods such as data mining and big data analysis to help national decision-making departments quickly grasp the real situation on the front line, and help government departments to quickly carry out epidemic prevention and control work.

References

[1] Wuhan Municipal Health Commission. Wuhan municipal health and health commission's briefing on the current pneumonia epidemic situation in our city [EB/OL]. (2020-01-18). http://wjw.wuhan.gov.cn/front/web/showDetail/2020011809064. opens in new tab.

[2] ZHOU Juan, LI Dan, LONG Yun-zhu. Advances in related research on novel coronavirus (2019-nCoV) [J]. Chinese Journal of Infection Control, 2020, 19 (3): 1-5.

[3] Han Minghui, Fang Hongyi, Yang Dong see, et al. Analysis of the current situation and trend of COVID-19's incidence abroad [J/OL]. [2020-02 Murray 21]. Shanghai Preventive Medicine.

[4] Zhuang Yingjie, Chen Zhu, Li Jin, et al. Clinical and epidemiological characteristics of 26 confirmed cases of COVID-19 [J/OL]. [2020-02 MUL21]. Chinese Journal of Hospital Epidemiology.