Design and Implementation of Web User Behavior Analysis Method Based on Big Data

Hongwei Huang

Shandong University, Weihai, 264299, China

Abstract. In China, the Internet has developed to a relatively mature scale, and Internet applications have gradually transitioned from being singular to diversified. The Internet is changing people's ways of learning, working, and living, and even influencing the progress of the entire society. Against the backdrop of rapid Internet development, we have gradually entered the "big data era." Faced with such a vast amount of data, single-machine statistics have become inadequate. This paper first introduces the background and significance of the research, providing a preliminary description of network traffic from the perspective of network services. It then elaborates on the concepts and classifications of network user behavior, along with key data, mainly introducing analysis methods. The primary method used in this paper is cluster analysis, including distance and similarity coefficients in cluster analysis, with a focus on the main steps and algorithm process of k-means clustering. Finally, the paper conducts user behavior analysis based on the clustering results.

Keywords: China Internet, Big Data Era, Network Traffic, Network User Behavior, Cluster Analysis, Distance and Similarity Coefficients, K-Means Clustering, User Behavior Analysis, user behavior analysis, clustering analysis.

1. Introduction

1.1. Background and Significance of the Study

In recent years, with the rapid advancement of the Internet and cloud computing technologies, internet applications have evolved from simplicity to diversification. Nowadays, people can access a vast array of useful knowledge from the Internet's expansive knowledge base. The swift growth of the Internet has led to the continuous generation of data across various industries.

According to the "China Internet Development Report" released by the China Internet Network Information Center, as of June 30, 2018, the number of Internet users in China had reached 802 million, with 788 million being mobile phone users, accounting for 98.3% of the total. The Internet penetration rate is 57.7%. This clearly demonstrates that the Internet is intricately linked to people's daily lives.

1.2. Research Status Domestically and Internationally

Today, the storage capacity of computers has steadily increased, and algorithms have become increasingly sophisticated. Recent years have seen exponential growth in data. It is estimated that by 2020, the volume of data will reach 40 ZB. Big data holds immense value.

Data mining involves extracting valuable information from large datasets, revealing connections and trends among the data. Effective data mining requires the collection of comprehensive data, as richer data sets yield more accurate insights. By selecting appropriate methods, one can swiftly and efficiently acquire an understanding of users' internet habits and preferences.

1.3. Main Research Content of the Thesis

This thesis focuses on the analysis of network user behavior in the context of big data. The main research content includes the concept and classification of network user behavior, key data points, calculation of similarity coefficients and distances in cluster analysis, k-means clustering, and a detailed understanding of the main steps and algorithmic processes of k-means clustering. Additionally, corresponding algorithms are designed for simulation purposes.

1.4. Organization of the Thesis

The thesis is structured as follows:

Chapter 1 introduces the research background, significance, and current status, concluding with an overview of the thesis structure.

Chapter 2 conducts empirical research, divided into three steps: first, the data preprocessing stage, which involves classifying 65,000 data points; second, the text classification stage, where the raw data is split into 650 groups using a distributed approach, followed by the combination of classified labels and time nodes to generate classification results, which are then subjected to k-means clustering analysis; and third, the analysis of user behavior based on the clustering results.

Chapter 3 focuses on analyzing the results, categorizing time nodes into 39 clusters through cluster analysis. It extracts the frequency of occurrences within specific time nodes as a measure of user activity.

Chapter 4 provides a summary and outlook, summarizing the research conducted and discussing future prospects.

2. Research Data and Methodology

2.1. ntroduction to Python

Python is an interpreted language, meaning there is no compilation stage during development, similar to PHP and Perl.

- • Python is an interactive language, allowing code to be executed directly at the Python prompt >>>.
- • Python is an object-oriented language, which means it employs object-oriented styles or encapsulation of code in programming techniques.
- • Python is considered a beginner-friendly language.
- • Features of Python:
- Easy to learn: Python has fewer keywords and a relatively simple structure, making it easier to learn.
- Easy to read: Python's code definitions are clear and straightforward.
- Easy to maintain: Python's source code is relatively easy to maintain.
- Extensive standard library: Python is well-supported on UNIX, Windows, and Macintosh platforms.

2.2. Cloud Computing Platform BosonNLP

The BosonNLP engine provides services via REST APIs and supports multiple programming languages. Python utilizes BosonNLP through its SDK. The BosonNLP Python SDK is a developer toolkit officially supported by BOSON, offering simplified access to REST interfaces.

2.3. Data Overview

To explore user behavior during the early morning hours, data from 23:00 to 23:59 was extracted from the Sogou Lab, totaling 65,000 data points. This data was then classified using cloud computing technology (As shown in Table 1).

Table 1. Data Classification

Index	0	1	2	3	4
Category	Sports	Education	Finance	Society	Entertainment

5	6	7	8	9	10
Military	Domestic	Technology	Internet	Real Estate	International

11	12	13
Women	Automotive	Gaming

2.4. Methodology

First, cloud computing technology is utilized to classify the extracted data. The classified labels and time nodes are then combined to generate classification results, and finally, the clustering results are analyzed to understand user behavior.

To explore the behavior of netizens during the early morning hours, we first use Python's pandas and numpy libraries to extract search data from the Sogou Lab for the time period between 23:00 and 23:59, totaling 65,000 data points. This constitutes the data preprocessing stage.

Next, the BosonNLP cloud computing technology is employed to classify the extracted data into 13 categories, such as sports, education, finance, society, etc. This stage is referred to as the text classification stage. The main technique used here involves distributed computing, as the cloud platform requires the raw data to be divided into 650 batches. Each batch is processed in a queue on the cloud platform, with the output results stored in a pandas dataframe and subsequently saved to an Excel file.

Afterward, the classified labels and time nodes are combined to generate the final classification results. The pyhanlp library is then utilized to perform k-means clustering analysis on these classified results. Finally, the clustering results are analyzed to understand user behavior. This constitutes the data analysis stage, where techniques such as deduplication and merging of multiple arrays are used to process and analyze the predicted and actual data.

2.5. Data Analysis

The code development for this project is structured around the data processing workflow, which is divided into three main stages: data processing, text classification, and cluster analysis.

2.5.1. Data Processing

Initially, search data for the time period between 23:00 and 23:59 is extracted from the Sogou Lab, totaling 65,000 data points. After preprocessing, the data segments are as shown in table 2:

Table 2. Data Processing

Time	User ID	Search contents
23:00:00	7.31E+15	Not wearing clothes, beautiful
23:00:00	8.21E+15	Price of Hetian jade
23:00:00	7.12E+15	Huari Casting Steel treasure chest
23:00:00	5.37E+15	Myopia
23:00:00	8.48E+15	Master Jingkong
23:00:00	5.98E+15	Watching other people's videos
23:00:00	7.36E+15	Kobe vs Jordan video
23:00:00	6.38E+15	Ni Ping's marriage history
23:00:00	6.38E+15	Ni Ping's marriage history
23:00:00	3.49E+15	Shaanxi earthquake
23:00:00	6.53E+15	Su Yu and Xu Shiyou's grievances
23:00:00	3.91E+16	Causes of the Wenchuan earthquake
23:00:00	7.41E+15	Pictures of Wu Xiubo
23:00:00	6.91E+15	What delicious foods are there in Chongqing
23:00:00	3.95E+15	Host income
23:00:00	2.58E+16	Assets and liabilities
23:00:01	7.99E+15	Baidu
23:00:01	5.62E+15	Flat warts + treatment
23:00:01	1.94E+16	Battle of Beijing, Shanghai, and Hangzhou
23:00:01	5.57E+15	Do hotel booking websites make money
23:00:01	5.60E+15	Watch Forensic Heroes II online
23:00:01	6.07E+14	Analysis
23:00:01	2.96E+15	Banning Sharon Stone
23:00:01	2.75E+15	About the ancient Sanwei Bookhouse
23:00:01	7.97E+15	Looting disaster relief supplies
23:00:01	4.66E+15	Looting disaster relief supplies
23:00:01	7.67E+14	Comrade Hua Guofeng passed away

The above data is saved as a txt file in preparation for the next stage, which is the text classification stage.

2.5.2. Text Classification

Primarily, the processing targets txt files, using Python's open function to read the txt content and delivering it to the BosonNLP cloud computing platform for text classification. The BosonNLP cloud computing platform has a limit on the number of data entries for a single input, which is restricted to 100 entries. Therefore, the source data needs to be split into streams of 100 entries each for input into the BosonNLP cloud computing platform for this data classification. For detailed code, see the appendix. The text classification results are as follows (Table 3):

Table 3. Text Classification

ID	Category	Label
0	4	Entertainment
1	2	Finance
2	4	Entertainment
3	4	Entertainment
4	13	Gaming
5	13	Gaming
6	0	Sports
7	11	Women
8	11	Women
9	6	Domestic
10	4	Entertainment
11	6	Domestic
12	0	Sports
13	6	Domestic
14	4	Entertainment
15	2	Finance
16	8	Internet
17	11	Women
18	6	Domestic
19	8	Internet
20	1	Education
21	11	Women
22	4	Entertainment

2.5.3. Cluster analysis

(1) The clustering analysis stage primarily uses Python's pandas to create the clustering dataset. The detailed code can be found in the appendix. The results are shown in Figure 1.

23:00:00	Entertainment	Finance	Entertainment	Entertainment	Gaming	Gaming	Sports	Women	Women	Domestic	Entertainment
23:00:01	Internet	Women	Domestic	Internet	Education	Women	Entertainment	Technology	Domestic	Domestic	Domestic
23:00:02	International	Finance	Finance	Entertainment	Entertainment	Education	Entertainment	Technology	Gaming	Gaming	Entertainment
23:00:03	International	Real Estate	Military	Entertainment	Education	Domestic	Domestic	Finance	Entertainment	Finance	Entertainment
23:00:04	Entertainment	Automotive		Automotive	Education	Entertainment	Domestic	Domestic	Domestic	Domestic	Entertainment
23:00:05	Education	Entertainment	Entertainment	Education	Women	Gaming	Entertainment	Education	Domestic	Military	Finance
23:00:06	Entertainment	Entertainment	Sports	Women	Entertainment	Entertainment	Society	Automotive	Women	Gaming	Finance
23:00:07	Entertainment	Entertainment	Entertainment	Domestic	Technology	Education	Entertainment	Entertainment	Society	Entertainment	Gaming
23:00:08	Entertainment	Technology	Finance	Automotive	Domestic	Domestic	Entertainment	Women	International	Entertainment	Society
23:00:09	Women	Technology	Education	Women	Domestic	Finance	Domestic	Domestic	Gaming	Automotive	Domestic
23:00:10	Entertainment	Military	International	Technology	Entertainment	Entertainment	Entertainment	Women	Education	Society	Domestic
23:00:11	Sports	Entertainment	Entertainment	Entertainment	Domestic	Technology	Women	Entertainment	Society	Entertainment	
23:00:12	Entertainment	Entertainment	Domestic	Domestic	Entertainment	Entertainment	Entertainment	Education	Entertainment	Military	Women
23:00:13	Finance	Military	Technology	Domestic	Education	Gaming	Internet	Internet	Domestic	Domestic	Education
23:00:14	Women	Internet	Education	Technology	Domestic	Domestic	Domestic	Sports	Military	Entertainment	Sports
23:00:15	Sports	Entertainment	Finance	Entertainment	Entertainment	Women	Society	Technology	Education	Technology	Entertainment
23:00:16	Entertainment	Finance	Education	Domestic	Domestic	Entertainment	Entertainment	Entertainment	Entertainment	Entertainment	Women

Figure 1. Clustering Analysis

(2) In the clustering analysis phase, the PyHanLP platform is used to perform clustering analysis on the clustering data, converting the clustering data into a txt file for input into the PyHanLP platform. The code is detailed in the appendix. The results are shown in Figure 2.

23:03	23:00	23:14:11	23:41:34	23:05:13	23:16:11	23:20:38	23:37:20	23:21:07	23:42:03	23:19:06	23:40:24	23:19:03	23:46:26
23:06	23:04	23:15:40	23:22:19	23:57:27	23:14:10	23:19:25	23:44:10	23:43:35	23:37:31	23:18:13	23:46:48	23:39:15	23:44:44
		23:17:21	23:19:47	23:04:44	23:42:01	23:42:45	23:34:21	23:34:42	23:09:43	23:16:58	23:15:20	23:03:17	23:15:41
		23:01:19	23:14:56	23:52:14	23:45:37	23:18:14	23:42:42	23:04:29	23:21:20	23:17:24	23:26:14	23:43:39	23:25:49
		23:08:34	23:18:55	23:17:42	23:17:02	23:58:56	23:23:46	23:09:47	23:25:03	23:22:57	23:27:05	23:01:13	23:08:33
		23:04:25	23:42:20	23:34:12	23:39:37	23:49:18	23:09:22	23:01:39	23:34:19	23:02:46	23:08:15	23:08:09	23:47:36
		23:08:38	23:22:10	23:57:41	23:39:13	23:57:42	23:04:17	23:05:03	23:00:51	23:04:28	23:03:35	23:31:45	23:48:48
		23:00:02	23:00:49	23:12:33	23:39:56	23:59:42	23:41:15	23:27:34	23:54:49	23:04:47	23:09:07	23:30:31	23:25:48
		23:04:13	23:23:06	23:58:43	23:05:39	23:28:01	23:20:24	23:31:00	23:25:12	23:01:40	23:04:34	23:06:31	23:08:36
		23:55:36	23:23:07	23:44:51	23:00:23	23:31:47	23:25:41	23:55:54	23:26:23	23:33:08	23:32:37	23:29:34	23:00:21
		23:02:33	23:08:59	23:09:27	23:20:58	23:28:24	23:29:21	23:24:41	23:02:51	23:05:23	23:58:28	23:06:50	23:08:37
		23:53:36	23:27:28		23:23:48	23:53:58	23:47:22	23:13:18	23:35:02	23:29:53	23:54:05	23:33:04	23:06:06
		23:49:25	23:00:50		23:56:05	23:48:10	23:44:30	23:34:37	23:13:16	23:58:47	23:01:20	23:14:03	23:07:18
		23:48:35	23:28:48		23:55:58	23:59:38	23:38:16	23:22:07	23:37:04	23:22:41	23:04:33	23:18:27	23:42:28
		23:02:50	23:29:59		23:52:02	23:04:31	23:50:02	23:38:35	23:11:17	23:13:38	23:51:35	23:41:47	23:30:34
		23:10:00	23:07:44		23:49:42	23:27:31	23:34:26	23:45:23	23:13:14	23:31:29	23:55:11	23:09:30	23:31:01
		23:18:45	23:21:30		23:54:26	23:32:54		23:18:42	23:11:58	23:37:24	23:55:13	23:33:42	23:30:32
		23:45:02	23:46:11		23:58:48	23:57:56		23:25:35	23:52:07	23:25:59	23:32:32	23:01:28	23:58:04
		23:48:14	23:30:18		23:23:53	23:49:02		23:01:03	23:45:48	23:02:35	23:15:17	23:03:07	23:32:55

Figure 2. Clustering Analysis

3. Research Results and Data Analysis

Firstly, we extracted search data from Sogou Labs between 23:00 and 23:59, totaling 65,000 entries. Using BosonNLP's cloud computing technology, the extracted data was classified into 13 categories. By categorizing the online search content from 23:00 to 23:59, we found that users were most interested in entertainment information during this time period, followed by topics related to women, finance, and games. Interest in other areas was relatively balanced.

From an emotional perspective, most users felt relatively happy between 23:00 and 23:59, showing significant interest in entertainment, domestic news, finance, and games, indicating that most were in a non-working state. Topics related to women, which are considered sensitive, also garnered significant attention during this time.

We then combined the classified labels with time nodes to generate clustering data, followed by k-means clustering analysis. Using the PyHanLP Python tool, we performed time-node clustering analysis on the clustering data, describing user behavior at different time nodes within the same category. The analysis results are shown in Figure 3. The time nodes were classified into 39 categories through clustering analysis.

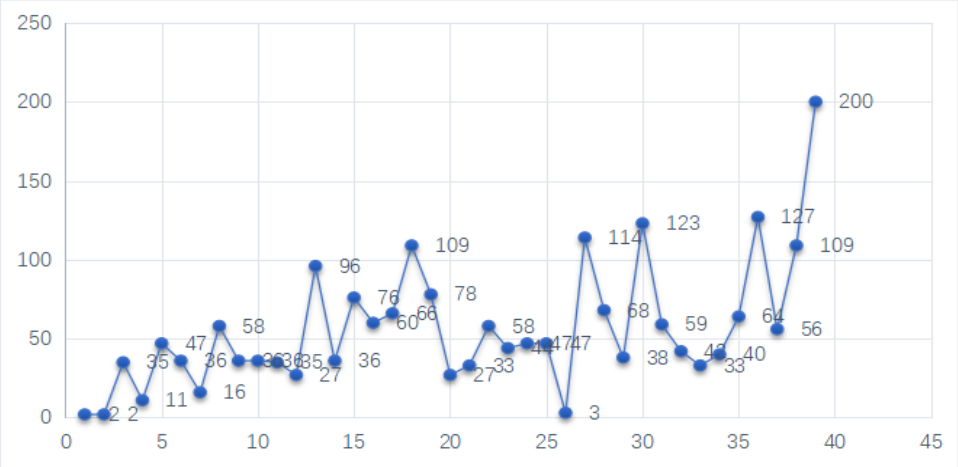

Figure 3. Time Point Classification

From Figure 3, it can be seen that the number of time nodes covered in the 39 clustering results varies. This indicates that categories with a larger number of time nodes have more frequent user activity. By counting the frequency of each time node within each category, we can identify the primary time nodes where users are most active.

In this study, several time nodes with more frequent user activity were selected. The frequency of occurrences within a particular time node was extracted to serve as an indicator of the time nodes where users in that category are most active.

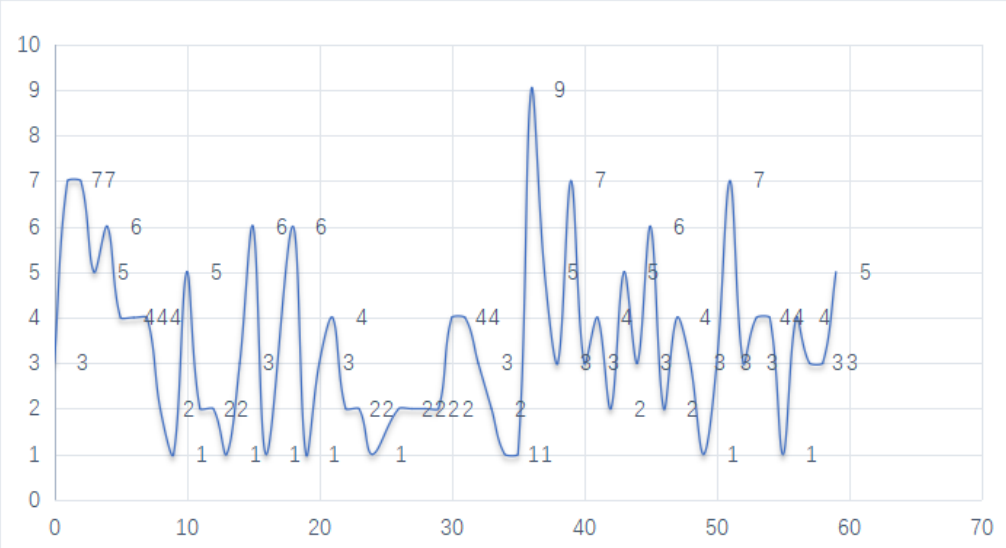

Figure 4. The visit frequency of a certain type of user within one hour

From Figure 4, it can be seen that this type of user exhibits a high level of activity, appearing at every time point with a frequency of more than 4 times on average. The peak frequency of visits, 9 times, occurs between 23:39 and 23:40. Therefore, this user group is more active between 23:00-23:20 and 23:35-23:59, with increased activity as midnight approaches.

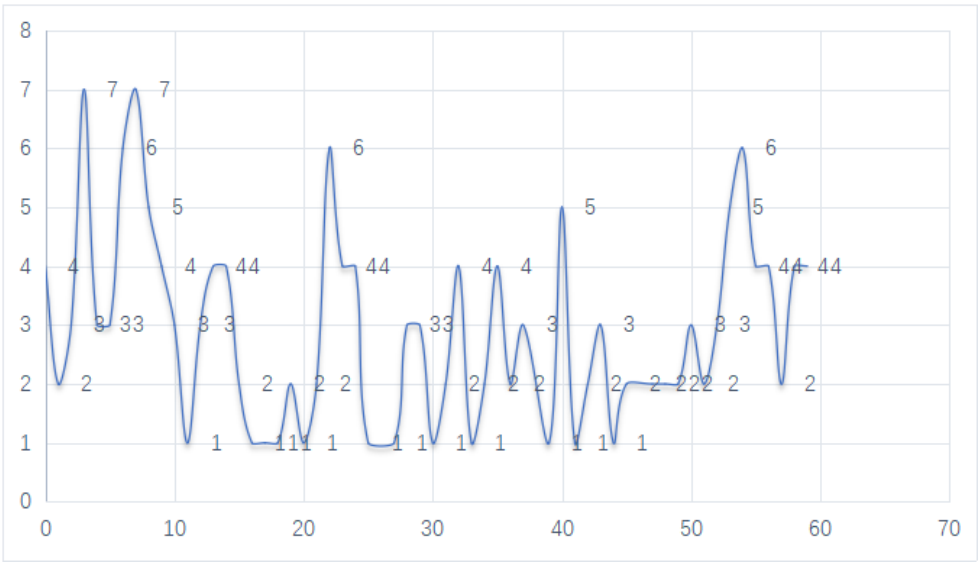

Figure 5. The visit frequency of a certain type of user within one hour

From Figure 5, it can be seen that this type of user exhibits a relatively high level of activity, appearing at every time point with a frequency of more than 3 times on average. The peak frequency of visits, 7 times, occurs between 23:06 and 23:07. Therefore, this user group is more active between 23:00-23:10 and 23:50-23:59, with increased activity as midnight approaches.

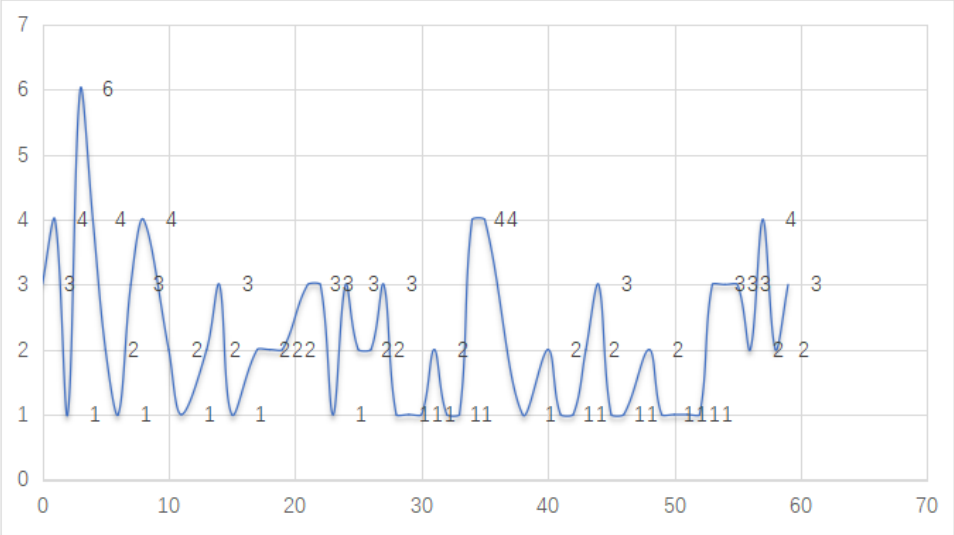

Figure 6. The visit frequency of a certain type of user within one hour

From Figure 6, it can be seen that this type of user exhibits a moderate level of activity, appearing at some time points with a frequency of more than 2 times on average. The peak frequency of visits, 6 times, occurs between 23:04 and 23:05. Therefore, this user group is more active between 23:00-23:10 and 23:53-23:59, with increased activity as midnight approaches.

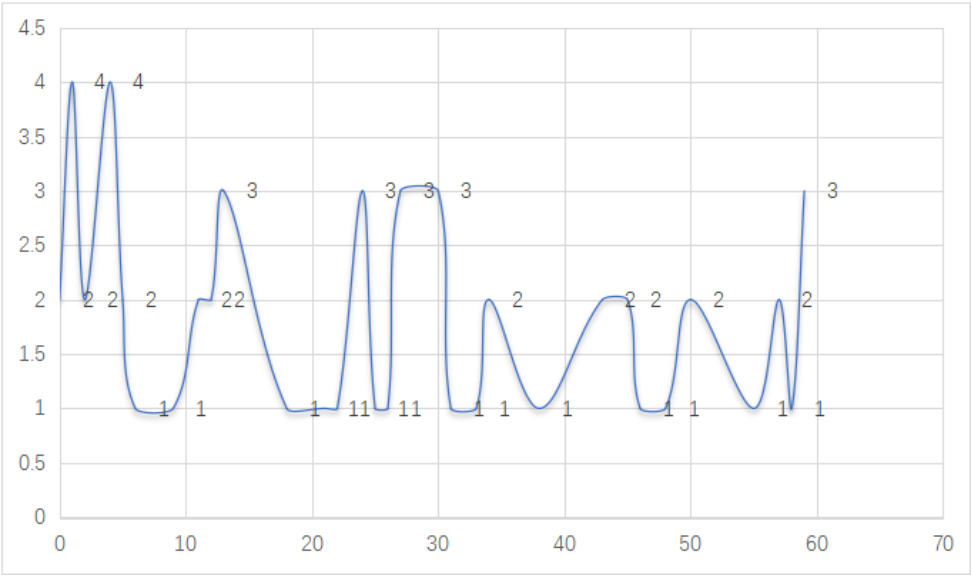

Figure 7. The visit frequency of a certain type of user within one hour

From Figure 7, it can be seen that this type of user exhibits a low level of activity, appearing at some time points with a frequency of more than 2 times on average. The peak frequency of visits, 4 times, occurs between 23:04 and 23:05. Therefore, this user group is more active between 23:00-23:07 and 23:56-23:59, with increased activity as midnight approaches.

In summary, across the four typical user clusters, we find that during the period from 23:00 to 23:59, all four user groups are highly active between 23:00 and 23:10. Additionally, between 23:55 and 23:59, all four user groups show an increasing trend in activity levels from low to high. This indicates that the time around midnight is a particularly active period for users.

4. Summary and Outlook

4.1. Summary of This Paper's Work

In today's rapidly developing internet era, users are increasingly frequent in their use of the internet, leading to a massive increase in the amount of data generated. Faced with such a vast amount of data, the challenge for all operators is how to accurately, quickly, and effectively extract the hidden value beneath big data and analyze this information. This paper begins by discussing the background, significance, and current state of research on internet user behavior, delving into the concepts and classifications of internet user behavior, introducing some common data mining methods, and focusing on the concepts of cluster analysis, distance measurement methods, correlation coefficient calculation, and types of cluster analysis.

First, 65,000 data entries were classified using BosonNLP cloud computing technology. The output results were saved into a pandas dataframe and then stored in an Excel file.

Afterward, the classified labels and time points were combined to generate classification results, and finally, user behavior analysis was performed on the clustering results.

4.2. Further Research Ideas and Prospects

This paper mainly adopts clustering analysis algorithms. The collected data is first classified, and then the k-means algorithm is used for clustering analysis, revealing some user behavior characteristics between 23:00 and 23:59. However, there are still some shortcomings in this paper, which can be expanded upon in the following aspects:

(1) This paper collected a total of 65,000 data entries from 23:00 to 23:59 from Sogou Labs, which is a relatively small amount of data. Future research can involve long-term data collection to analyze data characteristics over a day or a month.

(2) This paper only analyzes the characteristics of a few types of users, focusing on the visit frequency of users during a specific time period. Future research can involve comprehensive analysis of all user types.

(3) There are still many advanced research algorithms available. Future work will focus on strengthening the study of these advanced algorithms.

References

[1] 42nd Statistical Report on the Development of the Internet in China, July 2018 by the China Internet Network Information Center (CNNIC).http://www.cnnic.cn/.

[2] Ren, S. Y. (2014). Analysis of Internet User Behavior Based on Big Data [Master's thesis, Beijing University of Posts and Telecommunications]. Pages 18-20. December 2014.

[3] Hu Yanqing, Zhou Jinyan, Xu Xiaona. Application of Data Mining in Mobile User Behavior Analysis System JJJ. Modern Telecommunication Technology. 2013.01, 01 (01): 86 - 89.

[4] Zuo Jun. Network User Behavior Analysis Based on Big Data JJJ. Software Engineer, 2014. 05, 39 (8): 556 - 558.

[5] Chen Wenwei. Overview of Data Mining and Knowledge Discovery JJJ. Computer World News, 1997.05, 24 (8): 122 - 124.