Digital Life Analysis Based on R

Yuewen Guo
School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China
winifred0918@163.com

Abstract. Based on R language ggplot2 and wordcloud2, this paper analyzes, collects and visualizes the collected personal behavior data sets, and completes the interpretation of personal digital life.

1. Introduction
In recent years, with the rapid development of the Internet, the term "personal behavior data" has entered our lives. Personal behavior data refers to data related to user-generated data and related to personal behavior, including social media data such as Weibo and WeChat, shared bicycle track data, bus card record data, mobile phone call record data, and personal mobile terminal. Consumption data, etc. With the advent of the Internet big data era, traditional analysis methods are increasingly difficult to meet the growing demand for personal behavior data processing. In order to solve this problem, a large number of new tools and technical means have emerged. Among them, R language is an excellent tool for statistical calculation and drawing.

This paper aims to make a variety of visual analysis of personal behavior data through R software and related expansion packs, so as to establish a personal image map and propose relevant suggestions for improving life.

2. Data pretreatment
The whole data preprocessing links are mainly divided into two actions. Firstly, we directly download through the corresponding network platform, or make corresponding information collection program to collect and obtain the initial data needed by the research. Secondly, we conduct data preprocessing such as data filtering, deduplication, and null processing, and normalize the data into a format suitable for data analysis. We have four kinds of data sets. They are:

1) Personal mobile communication records
   The personal mobile communication record can be directly downloaded from the official website. This study summarizes the personal communication records for the first six months of 2017.

2) Personal bank card consumption record
   The personal bank card consumption record can be directly downloaded by logging in to the personal bank card online banking. The data used in this study is the transaction details of the debit card issued by the Bank of Communications Chongqing Branch from September 2013 to June 2017.

3) Personal social media records
   Personal social media recording sources can be fallen into two categories: QQ chat records and WeChat chat records. QQ chat records can be directly exported into txt format through personal
computer client. The QQ chat record used in this study ranged from August 2014 to December 2017. WeChat chatting records can be exported by the synchronous assistant software (version v3.3.0_64bit).

(4) Personal shopping records

Personal shopping records are collected through a node.js-based crawler. The shopping information collected in this study ranged from 2013 to 2017.

3. Data analysis

The main software used in the data analysis is the R language, supplemented by Excel for simple descriptive statistics.

3.1. Personal mobile communication records

Through R language, there are 347 records of personal communication in the second half of 2017, the total call cost is 1.5 yuan and the total call duration is 1 hour and 50 minutes. Among them, there are 231 communication records within 1 minute, 68 communication records for 1 to 2 minutes, 43 communication records for 2 to 9 minutes and 5 communication records for 9 minutes. We use the ggplot2 extension package in R to visually analyze the duration of the call:

![Figure 1. Call duration distribution scatter diagram](image)

As can be seen from the figure 1, the longest distribution of the four record lengths is mostly for personal dialing. Among them, the two highest ordering time are 33 min 29s and 14 min 19s, which are dialed out by the owner to the friend. The third highest time is 13 min and 20 s dialing out to her friend Zhao. The height of the fourth high period is 11 minutes and 06 seconds, which is allocated by the individual to Teacher Yan. The time of the fifth highest is 9 min and 49 s, which comes from her mother.

In the 347 communication records, we screened by month and the number of communication times from seven to December was 90, 64, 68, 49, 34, and 42. We use the ggplot2 extension package in R to visually analyze the number of calls:
We can see from the figure 2 that the number of individual calls has generally declined. After entering into postgraduate study in late September, due to the busy schedule, the number of calls was significantly reduced compared with the holidays. As can be seen in conjunction with Figure 1, the length of the call is greatly reduced in the first month after the start of the school.

3.2. Personal bank card consumption record
There were 947 records of bank card consumption, including 185 in total, and 762 expenditures; the total amount of money flow was 218,066.66 yuan, of which income was 108,972.88 yuan and expenditure was 109,093.78 yuan.

As can be seen from Figure 3, the main method for cardholders to receipt and pay is online payment (555), which accounts for 59%. The second is for other payments (121), which accounts for 13%. The third is to issue other funds (107), accounting for 11% which followed by other (92, 8%), inter-city consumption (48, 6%) and UnionPay online payment (24, 3%).

From the sum of consumption of credit card, among the total 947 records, the amount of credit card is distributed between 0 and 5000 yuan. Filling in the income and expenditure, we use ggplot2 to create histogram, which is as follows:
We can see that the amount of consumption is mainly 1000 yuan. Among those data, the smaller the amount, the larger the proportion of expenditure.

### 3.3. Personal social media record

For personal social media records, we use wordcloud2 for word cloud creating, so that data analysts can directly understand the purport of the chat. According to the chat text, we determine the stop words and add them to a txt file. Then, use wordcloud2 to perform word cloud drawing of chat records, which are as follows:

**Figure 5.** WeChat chat record (buying agent group) - word cloud

It can be clearly seen that the group chat is carried out around the words "like", "cute", "good-looking", "excellent" and "periphery", etc. The overall word sentiment tends to be positive. It reflects that this group has a lively and free atmosphere. Next, word cloud of chat content with mother is drawn and the result is as follows:

**Figure 6.** QQ chat record (with mother) - word cloud
From Figure 6, we can know that their topics are mainly focus on "Dad", "Night", "Video", "Back", "Mom", "Body", "Teacher", etc. It can be seen that the social media user's mother is very concerned and caring for her, and she talks about most topics related to family, learning and health.

3.4. Personal shopping records
Through R, a total of 279 Taobao shopping records were obtained, with a total cost of 34,604.05 yuan from nine categories. We can see that cosmetic & skin care, life household and book & audio & video account for nearly 70% of the total goods, nearly three quarters of the whole consumer spending. In contrast, buyer purchased the least number of digital home appliances and food health products with only 10.4% of the whole amounts. We can guess that the buyer is a fashion girl who be keen about dressing and making up. The pie chart is as follows:

![Figure 7. Taobao shopping category analysis](image)

Next, the purchase price is analyzed according to years:

![Figure 8. Goods price changing with year analysis](image)

In 2015, the buyer completed the highest amount of consumption of 2503.3 yuan and 1760 yuan in clothing. The third highest transaction was completed in 2016 with amount of 1585 yuan from cosmetic & skin care. In general, the purchase price in 2017 has been significantly reduced compared with the previous two years, which proves that the buyer may convert large transactions from Taobao to other channels.

4. Conclusion
This article uses the R software to conduct various visual analysis with personal behavior data and interpreted a personal digital life. From the communication record, the telephone exchange time of the owner is not long and the cost is small; the social relationship is relatively simple. From the bank card consumption record, Cardholders’ expenditures are significantly more than consumption. From the social media use, most of the topics discussed by user are related to study, family, and tourism. From
the shopping record, the cost of clothing is decreasing year by year. It is recommended that the data owner can maintain contact with family and friends. Hope she can develop a good consumption pattern and reduce unnecessary online shopping expenses.

References
[1] X. Wu, Y. Li, L. Li, Influence Analysis of Online Social Networks, Chinese Journal of Computers. 04 (2014) 735-752.
[2] X. Yang, D. Wu, Application of R language in Big Data Processing, Science & Technology Information. 23 (2013) 19-20.
[3] E. Silverman, Statistics for Ecologists Using R and Excel, Journal of Agricultural, Biological and Environmental Statistics. 23 (2018) 171-173.