An Identification Model of Water Army Based on Data Analysis

Jianhua Dai and Xinyang Wang

Department of management science and Engineering, Communication University of China, School of Economics and Management, Beijing, China.
Email: daijianhua66@163.com

Abstract. The water army creates a lot of inaccurate and misleading information on Weibo which has greatly undermined the users experience. It is necessary to use machine learning model to identify and exclude water army accounts. In this paper, we analyzed the data of water army and their behavior. We created some indices for the identification model, and used the web crawler to get the data of water army. In the end, we used support vector machine and k-means to build the model by the R software. Compared with the earlier models, the model achieves better identification results.

1. Introduction
In China, sina Weibo has become a byword for micro-blog. In 2017, the annual earnings report of sina Weibo showed that as of December 2017, the number of monthly active users of Weibo (MAU) has reached 392 million. The Weibo water army identification is to separate the water army account and the normal account from a large number of user accounts, so the Weibo water army identification problem can be simplified as a two classification problem. This paper will conduct an in-depth study of the Weibo water army, and find out the important characteristics of the water army through data analysis, and establish an identification model.

2. Literature Review
At present, the research on network water army can be divided into three directions at home and abroad: identifying network water army based on user behavior characteristics, identifying network water army based on user relationship characteristics and identifying them based on environmental characteristics.

The identification of network Water army based on user behavior characteristics is a representative method for the research of Water army recognition in the social network field. This method was first used by Benevenuto and others in the study of Youtube network Water army [1]. Mo and others think: compared with normal users, the Water army in the social network does not have normal social relations, and its formation has a special network structure [2]. Xu used the organizational relationship formed by the network water army to discover the network of water army formed by network water army, calculated the similarity between the water army and its neighbors on this basis, and modified the classification and determination of the network water army by using the K nearest neighbor algorithm [3]. Las-Casas et al. proposed a way to identify the source from the network Water army [4]. In order to minimize the burden of work and maximize the efficiency of the network, the network will concentrate on a large amount of garbage opinion for a period of time, so the network load will increase suddenly in this period, and the traffic will also be concentrated in some links. This method has high recognition accuracy.
Unlike classification technology, clustering technology is an unsupervised learning mode. Before clustering, we do not know the classification of data sets. Clustering converges similar information by similarity of information. The ultimate goal of clustering is to make the distance between the same objects as small as possible, and the distance between different classes is as large as possible. Common clustering techniques include K-means algorithm, Clara algorithm and so on.

3. Acquisition of Data from Weibo Water Army

3.1. Analysis of the Characteristics of Weibo Water Army
Before grabbing the data of Weibo Water army, it is necessary to make a simple analysis of them. According to previous studies, Weibo Water army was divided into zombie fans and forwarded and commented Water army.

Zombie fans is mainly to play the role of increasing the number of fans for Weibo's account or target account. At first, they are mainly batch registered robot accounts. The earliest features of the zombie fans are more obvious. They usually have no head images, fewer fans, more attention, no meaning, no user tags and other detailed user information, and almost no original Weibo. With Weibo's cleansing of zombie accounts, such zombies have lost their living space. This kind of zombie fans is not the object of our study.

Although the appearance of Weibo's water army account is similar to a real user, we can reasonably speculate on its features before we do rigorous data analysis modeling:
1. The only purpose of the Weibo Water army is to complete the employer's requirements.
2. No normal user will take the initiative to pay attention to a Weibo Water army.
3. The interaction of Weibo water army account should be very poor.
4. The only purpose of the Weibo water army is to complete the employer's request; its original Weibo is only to create a false image of its real user.
5. The number of followers in the water army accounts is often very high and has a large gap with the number of fans.

3.2. Determination of Quantitative Indicators
Before grabbing Weibo account data, we must first determine which data is useful, that is, determine reasonable quantitative indicators.

Combined with the above analysis and hypothesis, the following quantitative indicators are selected for grasping.

The indicators shown in Table 1 are related to the first part of the analysis and hypothesis. Specifically:

Table 1. Related indicators of Weibo account

| Index                                                                 |
|-----------------------------------------------------------------------|
| The Number Of Friends/The Number Of Followers                        |
| Account level                                                         |
| Hot Weibo number                                                      |
| The number of recent 5 original Weibo Reviews                        |
| The last 5 original micro-blog points                                |
| The most recent 5 forwarding micro-blog's proportion of micro-blog in advertising category |
| Is there an attestation                                               |
| Whether it is a member or not                                        |
| Whether there is educational information                             |
| Does it has work information                                          |
| Does it add V                                                         |
| The number of zombie fans in the first 50 vermicelli                 |
Select whether "member" and "add V" index is because the water army accounts to control costs, usually do not add V and member recharge.

The number of "zombies" in the top 50 fans was chosen because the fans of the water army account were usually zombies. For statistical purposes, the zombie fans are used to determine the ratio of the number of fans to the number of fans. If the number of fans / fans is >=5, the account is zombie fans.

3.3. The Way To Gather a lot of Weibo Water Army
In order to facilitate the accurate acquisition of large numbers of Weibo Water army data, we need to gather them together. This article adopts the method of buying water army. To this end, this article specially registered the Weibo experiment account, bought 200 zombie fans, and issued a Weibo, which asked the Weibo water army to forward and comment, and in this way, a considerable sample of the water army was gathered to facilitate access to the data of these accounts. At the same time, in order to compare with the Weibo Water army account, we chose the Weibo account designated as the real user and grabbed the same index data.

3.4. Weibo Account Data Capture
In general, as long as information displayed on browsers can be crawled through web crawlers. The web crawler is a program that can automatically extract the information of web pages. It mainly provides Web data for the search engine. It is an important technology to obtain network data.

The basic workflow of the web crawler is as follows:
1. First, a part of the seed URL is selected.
2. Put these URL into the URL queue.
3. From the URL queue to be fetching, fetching in the URL, parsing the DNS, and getting the IP of the host, and downloading the web page corresponding to the URL and storing it in the downloaded Web Library. In addition, put these URL into the URL queue that has been grabbed.
4. Analyze the URL in the URL queue that has been grabbed, analyze other URL in it, and put URL in the queue to be grabbed so as to enter the next loop.

![Figure 1. Width first traversal strategy](image_url)
In the reptilian system, it is important to grab the URL queue. It is also an important problem to grab the order of URL in the URL queue, because it involves grabbing that page first, and which page to be grabbed after. And the way to determine the order of these URL is called grasping strategy. Grasping strategies include depth first strategy and breadth first strategy.

1. Depth first traversal strategy. The depth first traversal strategy is that the web crawler starts from the starting page, and a link is tracked by a link, after the line is processed and then transferred to the next starting page to continue tracking the link.

2. Width first traversal strategy. The basic idea of breadth first traversal strategy is to insert the links found in the new download page directly into the end of the URL queue to be grabbed. That is, the web crawler will grab all the pages linked in the starting page, and then select one of the link pages, and continue to grab all the pages linked in this web page. Take the Figure 1 as an example; the traversal path is A-B-C-D-E-F-G-H-I.

Usually Weibo crawlers use breadth first traversal strategy. Sina Weibo provides API interface for developers, and it can get Weibo data conveniently. In this paper, we use R’s XML package and rvest package to grab data.

3.5. Data Cleaning
Because there are a lot of missing values and wrong formats in the original data, it can not be used for analysis and modeling directly, so the original data should be cleaned first. In addition, some indicators can not be directly obtained, and we need to use available data and calculate them.

4. Acquisition of Data from Weibo Water Army
In this paper, the R software is used as the operation platform, and the Weibo Water army identification model is established through K-means clustering algorithm and SVM classification algorithm.

4.1. Brief Introduction of Tools and Algorithms
The K-means algorithm is to cluster the samples into K clusters (cluster). The specific algorithm is described as follows:

1. Randomly select k clustering centroid points (cluster centroids) as: \( \mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n \).

2. Repeat the following process until convergence

   For each sample i, calculate the class that it should belong to.

\[
e^{(i)} := \arg \min_j \| x^{(i)} - \mu_j \|^2
\]  

(1)

For each class of j, the centroid of the class is recalculated.
\[ \mu_j := \frac{\sum_{i=1}^{m} \mathbb{1}_{\{c^{(i)} = j\}} x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}_{\{c^{(i)} = j\}}} \]  

(2)

Support vector machine SVM (Support Vector Machine) is a supervised machine learning model, usually used for pattern recognition, classification, and regression analysis.

4.2. Using K-Means Algorithm to Verify Account Difference

The detailed analysis steps are as follows:

The first, read the data into R and transform the data into the form of data frame.

The second, draw a pie chart of the account type, as shown in Figure 2. And then draw the density function diagram of “The Number of Friends/the Number of Followers”, as shown in Figure 3.

The third, the KNN package is used for K-mean clustering analysis.

The test data set is constructed to set the account type column of the test data set to null value.

```r
newdata <- data1
newdata$accountType <- NULL
```

Run K-means clustering analysis on the data set newdata, and save the clustering results in KC. In the kmean function, the number of clusters needed to be generated is set to 2:

```r
kc <- kmeans(newdata, 2)
```

Output results: K-means clustering with 2 clusters of sizes 191, 35. The K-means algorithm generates 2 clusters with 191 and 35 sizes respectively. Clustering vector: represents the clustering of each row record (1 represents the first cluster 2 represents second clusters).

```
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
......
[211] 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

Figure 3. The density function diagram of “The Number of Friends/The Number of Followers”

Available components: Represents the 20 components contained in the object returned by running the K-means function. "Cluster" is an integer vector, which is used to represent the clustering of records, and "center" is a matrix that represents the center of each variable in each cluster.

```
[1] "Cluster" "centers" "totss" "withinss"
```

Create a table. Does the statistical clustering result agree with the actual situation?
There are two kinds of clustering algorithms, which are in good agreement with the actual situation. It shows that there is a significant difference between the Weibo water army account and the normal users. Finally, scatter plot, Figure 4, is plotted according to the clustering results, and the data are listed in the dataset "ratio of fans to fans" and "popular Weibo numbers".

```r
plot(newdata$NumberOfFriends/Newdata$NumberOfFollowers,newdata$hotWeibo,col=kc$cluster)
```

**Figure 4.** Scatter plot of clustering result

### 4.3. Using SVM Algorithm to Construct Classification Model

To call the SVM function, you need to introduce the e1071 package in the R software.

A brief introduction to the SVM function in R: the key parameters of function SVM () are kernel, cost and gamma. Kernel refers to the type of support vector machines, which may be linear SVM, multinomial SVM, radial SVM or Sigmoid SVM. Cost is a cost function that violates the constraint. Gamma is a parameter used by all the other SVM except linear SVM.

It can be seen that the classifier distinguishes between the Weibo Water army and the normal users.

**Table 3.** The table of classification result

| pred     | water army | normal users |
|----------|------------|--------------|
| water army | 59         | 0            |
| normal users | 0          | 13           |

### 5. Summary

This paper discusses the hazards of Weibo Water army and shows the importance of harnessing Weibo’s Water army. By studying the results of previous studies, the main ideas of establishing Weibo Water army identification model are summarized. Based on the R software environment, the machine learning algorithm is applied to the identification model by using related packages. Specifically, in this paper, the K-means clustering algorithm is used to verify the significant difference between the water army account and the normal user account. Then, the linear SVM algorithm is used to train the
classification model with the related indexes of the previous analysis, and the test data are used to verify the classification results, and a good classification effect is obtained.

6. Acknowledgments
The helpful comments of the editor and two anonymous referees are gratefully acknowledged. The work has been supported by National Social Science Fund (17BC059) and the research tasks of Communication University of China (3132014XNG1458, 3132018XNG1863, CUC18QB12, CUC16A17, CUC17A16). Moreover, this work has been partly performed in the project “A Study on Interconnection Charges during China’s Integration of Telecommunications, Broadcasting and Internet Networks” supported by National Natural Science Fund (71203202).

7. References
[1] Benevenuto F, Rodrigues T, Almeida V, Almeida J, Zhang C, Ross K. Identifying video spammers in online social networks. In: Castillo C, Chellapilla K, Fetterly D, eds. Proc. of the 4th Int’l Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2008). New York: ACM Press, 2008. 45-52.
[2] Mo Qian, Yang Ke. Research on network Water army recognition [J]. Software journal, 2014.25 (7): 1505-1526.
[3] Xu C, Zhang J, Chang K, Long C. Uncovering collusive spammers in Chinese review Websites. In: He Q, Iyengar A, Nejdl W, eds. Proc. of the 22nd ACM Conf. on Information and Knowledge Management (CIKM 2013). New York: ACM Press, 2013. 979-988.
[4] Las-Casas PHB, Guedes D, Almeida JM, Ziviani A, Marques-Neto HT. Spa De S: Detecting spammers at the source network. Computer Networks, 2012, 57 (2): 526-539.
[5] Feng Dian. Design and implementation of data acquisition and analysis system for Weibo [D]. Beijing: Beijing University of Posts and Telecommunications, 2013.
[6] Xue Yi. Statistical modeling and R software [Z]. Beijing: Qing Hu University Press, 2006.
[7] Zhou Aiwu, Yu Yafei. K-Means clustering algorithm research [J]. Computer technology and development, 2011, 21 (2): 62-65.
[8] Zhang Xuegong. On statistical learning theory and support vector machine [J]. Journal of automation, 2000, 26 (1): 32-42.
[9] Fang Kuang Nan. Classification and clustering algorithm based on data mining and software realization [D]. Guangdong: Jinan University, 2007.