Based on the Tensor Flow Framework and the Improved Logistic Regression Algorithm, the Detection Model of Weibo Water Army was Constructed

QI SHEN, YUNXUE GAO and LIKUN NIU

ABSTRACT

With the development of Sina Weibo and the rapid growth of user volume, Sina Weibo has become the largest comprehensive social networking platform. But its hasty progress has further caused the rise of Weibo water army and gradually makes it appear in a rapid increase in the trend. The existence of the water army has a significant impact on the quality of Weibo information, but also causes the shared environment is not healthy and other serious problems. Built on the above reasons, this paper extracts the characteristics of the water are of the characteristics of water army’s user attributes, behavioral attributes and temporal attributes by studying the characteristics of the water arm's users and normal users. At the same time, based on the Tensor Flow depth learning framework and the improved logistic regression algorithm, the training model is effectively identified the water army. The accuracy and validity of the model can be demonstrated by the experiment.

INTRODUCTION

According to CNNIC in January 2013 Published in China Internet Development Statistics Report[1]. Weibo households at the end of 2012 to reach 309 million, the annual growth rate of 23.5%. Internet users in Weibo users reached 54.7%, the report that Weibo has become the mainstream application of Internet users in China, in the network of public opinion dissemination center status, now Weibo users than to be much higher than this figure, Weibo is not just a simple social platform, however; it is a media platform, people thought it to publish, receive and disseminate information. And because the existence of the water arm makes Weibo information appear good and bad phenomenon, a large number of water army through the propaganda campaign to guide users to an event attitude, resulting in Weibo information authenticity decline. Therefore, how to automatically identify Weibo water army to avoid the Internet users by the adverse factors in order to improve the quality of Weibo information has a significant role.

The water army studied in this paper refers to the group of people who are specialized in business or private purposes, and they have achieved the purpose of promoting, selling or attacking certain people or commodities by employing a large number of people on the internet[2]. With the rise of domestic Weibo, the current domestic has appeared for the Weibo identification of the water army detection
technology. Gao Fan has studied the communication environment of Weibo and passed a series of micro blogs the water army incident analyses the influence of the water army on the network environment and puts forward some countermeasures\textsuperscript{[3]}.

Yuan Xuping based on the comprehensive index and entropy method, the automatic identification of Weibo water army is established, the entropy method is used to determine the weight of the water army, and the automatic identification model is established with the multi-index comprehensive index \textsuperscript{[4]}. Han Zhongming constructs a probability graph model to identify the user's probability of the water army through the fusion of the user's behavioral characteristics and the probability graph model \textsuperscript{[5]}.

Due to the rise of the deep learning, it has not applied identification of the water army. So this paper aims to the TensorFlow based on the depth of learning framework and improved logic regression algorithm fusion, through Weibo water army attributes analysis, training out accurate Weibo water army identification model.

FEATURE SELECTION AND MODEL DESIGN OF WATER ARMY

Select the Characteristics of the Water Army

THE USER ACCOUNT ATTRIBUTE OF THE WATER ARMY

1) Water Army users concerned about many users, but fans are few. "Zombie powder" through the attention, mutual fans other ways to achieve profit purposes; "waistcoat" is responsible for the manufacture of public opinion, affecting the truth; Mainly water army through forwarding, comments etc. to expand the negative side, so Weibo marketing\textsuperscript{[3]}. As a result of the above several forms of activities determine the water army users and normal users have different attributes, in order to improve an influence, they will focus on a large number of users. These users pay little attention to each other; at the same time Weibo water army is not a real person, so in their Weibo on their circle is very small, that is, the number of fans of the water army users is very small. Figure 1 is the normal user relationship diagram. Figure 2 is the relationship between the water army users, figure citation literature \textsuperscript{[6]}. The center of the hollow circle is the user, the surrounding solid circle is the user's concern friends, the connection line is the relationship between the two users. It is easy to see that the social circle of normal users is complex and the social circle of the water army is relatively simple.

The formula 1 represents the characteristics of the water army. FFR stands for the fan follow ratio. Friend refers to fans; follow indicates the user's concern. Depending on the formula and the above analysis, the larger the FFR value is, the closer the FFR is to the normal user, and the smaller the FFR is, the closer to the water army user.

\[
FFR = \frac{\text{number of friends}}{\text{number of follows}}
\]
2) Non-authenticated users. Water army users are not real people, so it is difficult to get Sina Weibo certification, while normal users will be able to authenticate their user attributes. Identify = \{0, 1\}, 1 for unauthenticated, 0 for authenticated.

3) Low activity. Active degree of normal user is high. Active degree of water army user is low. Use an active representation.

THE USER BEHAVIOR ATTRIBUTE OF THE WATER ARMY

1) Weibo URL usage is high. Water Army users often by adding a link in the blog post, to induce other users click on the link to browse the page, so as to achieve their unique purpose. Formula 2 represents URL use radio. \( \text{Nu} \) represents the number of URLs in the user blog. \( \text{N} \) represents the number of Weibo posted by users; the size of the urlR is proportional to the probability of the water army.

\[
\text{urlR} = \frac{\text{Nu}}{\text{N}}
\]

2) Original Weibo low proportion. Water army users rarely published personal opinion Weibo, most Weibo are forward or comment for this Weibo content publicity campaign, even if there are some personal opinion of the original Weibo, but also for other water army users to provide the conditions for forwarding comments. Such as formula 3, CreateR on behalf of the original Weibo ratio, \( \text{Nc} \) is the number of original Weibo posts, and \( \text{N} \) represents the total number of Weibo. Obviously the greater the value of CreateR, the higher the likelihood of the normal user, the smaller the value of the more inclined to the water army users.

\[
\text{CreateR} = \frac{\text{Nc}}{\text{N}}
\]

3) Weibo get a low rate of interaction. The number of comments, point praise and forwarding was less than those of normal users. The average Weibo comments, point praise, the number of forwarding description of the interaction.

THE USER TIME ATTRIBUTE OF THE WATER ARMY

1) Weibo release time interval is more fixed. As water army users are usually controlled by the machine to release a series of content, so the timing of the release
often has a certain mechanical law. Such as formula 4 and 5, \( P(x_i) \) represents the user's time series is divided into \( m \) sub-sequences, the number of blogs in each sub-sequence the same. \( K \) is the normalization factor, that is, each sub-sequence of the blog array is the same. The formula is 6.

\[
T_r = -k \sum_{0 < x_i \leq a} p(x_i) \log p(x_i) \tag{4}
\]

\[
K = \frac{-1}{\sum_{0 < x_i \leq a} \frac{1}{m} \log \left( \frac{1}{m} \right)} \tag{5}
\]

Speed of attention. Water army users usually pay attention to a large number of users within a short time after the registration is completed, so a user with a high rate of concern is likely to be a water army user. Since it is difficult to get the user's attention time, it will use the account creation time to approximate the time of attention, and the number of users concerned in the 10 minutes of this attention time is the rate of attention. Equation 6 is the attention rate.

\[
Fr = \text{number of follows} \tag{6}
\]

In summary, this paper selected as the characteristics of the water army identification as showed in Table I.

### Table I. User Characteristics Description.

<table>
<thead>
<tr>
<th>User characteristics</th>
<th>Symbolic representation</th>
<th>Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>User account attributes</td>
<td>Id</td>
<td>User id</td>
</tr>
<tr>
<td>User account attributes</td>
<td>Identify</td>
<td>Whether to certify, 0, 1</td>
</tr>
<tr>
<td>User account attributes</td>
<td>FFR</td>
<td>Fans concerned about the ratio</td>
</tr>
<tr>
<td>User account attributes</td>
<td>Active</td>
<td>Activity</td>
</tr>
<tr>
<td>User behavior attributes</td>
<td>urlR</td>
<td>URL usage</td>
</tr>
<tr>
<td>User behavior attributes</td>
<td>CreateR</td>
<td>Original ratio</td>
</tr>
<tr>
<td>User behavior attributes</td>
<td>Acomment</td>
<td>Average number of comments</td>
</tr>
<tr>
<td>User behavior attributes</td>
<td>Apraise</td>
<td>Average number of praises</td>
</tr>
<tr>
<td>User behavior attributes</td>
<td>AForward</td>
<td>Average number of forwards</td>
</tr>
<tr>
<td>User time attribute</td>
<td>TR</td>
<td>Each Weibo time interval</td>
</tr>
<tr>
<td>User time attribute</td>
<td>FR</td>
<td>Attention radio</td>
</tr>
</tbody>
</table>

### Design of Water Army Identification Model

The identification of Weibo water army is from a large number of user accounts. The water army accounts and normal accounts are separated, so the water army identification problem can be converted into a dichotomy problem \([7]\).
Set Sina Weibo all users belong to the set \( M=\{x_1, x_2, \ldots, x_n\} \), in which \( x_n \) represents the \( N^{th} \) user from the user. Represents the user type with \( K=\{A, B\} \). Where A represents the user of the water army, and B represents the normal user. So the target function of dichotomies is:

\[
H(x_n) = \begin{cases} 
1, & x_n \in A \\
0, & x_n \in B
\end{cases}
\]  

(7)

The simplified target function can be mapped by \( M->\{1 \ 0\} \).

SIGMOID FUNCTION, DECISION BOUNDARY AND REGULARIZES COST FUNCTION

Based on the previous user characteristic data and the objective function of the two categories, the logical regression model is established on the sigmoid function, and the sigmoid function is used to plan the value of any range into the interval [0, 1]. The formula is shown in 8.

\[
h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}
\]

(8)

Where \( x \) is a multidimensional vector representing the user characteristics mentioned above, and \( \theta \) is also a multidimensional vector, which is the eigenvalue corresponding to \( x \). When \( h_\theta(x) = 1 \) is for the water army user, 0 is for normal users. But because the sigmoid function output value is between 0 and 1 between the numbers, and we have to solve the two classification problems, to make the final output value can only be 0 or 1. So we set a decision boundary, the output is greater than the boundary of 1 that is the water army, less than the boundary of 0 is the normal user.

However, due to the fact that the decision boundaries are set to over-fit the known data. The algorithm cannot make a decent prediction of unknown user characteristics. So we are regularization of the Cost function to reduce the Cost function value to achieve the optimal fit. The regularized Cost method likes Reference [8]. The Cost function of regularization is as formula 9.

\[
J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[ y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right] + \frac{\lambda}{2m} \sum_{j=1}^{g} \theta_j^2
\]

(9)

The difference between the regularized cost function and the original cost function is that the regular term of the ultimate weight is increased at the end of the formula.

GRADIENT DOWN

According to \( J (\theta) \) after the regularization of the previous section, we know that the accuracy of the model prediction has a strong relationship with \( J (\theta) \).
The J (θ) value must be lower. In order to obtain the minimum J (θ) value, we use the gradient algorithm to iterate the value of θ, except that the original gradient algorithm also increases the regularization so that its weight can be maintained in a small range to ensure the stability of the algorithm. The regularization method is not only a method to avoid over-fitting in the logic regression method, but also provides reference and improvement in other complex machine learning and depth learning to solve the problem of fitting.[14] Regularization gradient descent algorithm is like formula 10.

\[
\theta_j = \theta_j - \alpha\left( \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i^j - y_i^j)x_j^i) + \frac{\lambda}{m} \theta_j \right),\ j \in \{1,2,\ldots,n\} \tag{10}
\]

Where α is the learning rate, and the choice of α requires that the number of iterations be as few as possible. Each parameter is updated using the procedure of parameter matrix in the process of calculation. Finally, the optimal value of θ is achieved by multiple iterations. The J (θ) value is smaller.

**EXPERIMENT**

![Figure 3. The experimental process.](image)

Experiment is conducted according to figure 3.

**The Data Collection**

In the Sina Weibo network using crawler tools to obtain user account attribute information, user behavior information, user time information and a series of data. Artificially marked data of the user attribute a total of 1000, the number of the water army and non-water army users for the 4:6, at the same time bought 200 water army data for experimentation.
Data Preprocessing

Because the collected data are mostly in the form of text and the model is applied in the numerical model, so the collected data preprocess to obtain the user eigenvalues corresponding to the specific value.

Training Model

We selected 600 samples as training set, including 400 water army users and 200 non-water army users. Apply the depth learning TensorFlow framework and the regularized logistic regression algorithm of the previous section about to fuse the water army model.

Model to Evaluate

Using 600 tests sets data to evaluate the trained water army identification model. Accuracy and recall rate as an evaluation index, with A on behalf of the water army users to determine the correct, B on behalf of the normal user was mistakenly referred to as the water army users, C said the water army users are not detected on behalf of the normal user was correctly judged as normal user. According to the experimental results calculated $P = 0.9232$, $R = 0.9134$, from the evaluation of the results can be seen from the training of the water army detection model can be more accurate to determine whether a user is a water army user or ordinary users. Through this water army model can be Weibo social network to cope with too much water army users' problem, to solve the water army caused Weibo information good and bad phenomenon.

Epilogue

This paper mainly introduces the training model based on TensorFlow depth learning framework and regularized logical regression algorithm to detect Sina Weibo water army users. Through the analysis of the attribute of the water army users and the normal users, the feature set of the water army is obtained, and the model is used to train the more accurate the water army identification model. Although water army detection model is obtained by the method introduced in this paper can solve the problem of Weibo caused by water army, but there are a lot of living men that the water army all kinds of big V marketing. Their user attributes are more complex on Sina Weibo, so in order to completely purify the micro blog environment, we still face many problems.

REFERENCES
1. CNNIC. The 30th statistical report on Internet development in China, 2013.