A Regression Prediction Algorithm for the Popularity of English Forum Posts Based on Support Vector Machine

Tian Li

Department of Foreign Language, Shaoyang University, Shaoyang, China

Keywords: Support Vector Machine; Post Keywords; Post Popularity; Influencing Factors

Abstract: Predicting the popularity of English forum posts has become a key research direction of public opinion monitoring. This paper proposes a new regression prediction algorithm for the popularity of English forum posts. In the process of data processing, firstly, the influencing factors of posts are quantified to study the prediction of the development of public opinion popularity. In this paper, the first thing to do is to study the distribution characteristics of the popularity of English forum posts; then quantify the factors that affect the popularity of posts, and use the regression based vector machine to generate the development trend of public opinion; finally, the prediction method of this paper is tested, and the test results show that the prediction algorithm proposed in this paper has good accuracy Accuracy.

1. Introduction

With the popularity of the Internet, the number of network users shows an explosive growth, network platform has become a platform for ordinary people to express their opinions. And in the network platform, the place that netizens like to express their personal opinions and opinions most is the network forum. The Internet Forum has a large number of Internet users, which is the gathering place of public opinion information. At present, there are more than 70 million registered users in Tianya forum, tens of thousands of Posts appear in the forum every day. Monitoring these posts timely and effectively, grasping the public opinion direction of Internet users, can help the decision-making departments of the government to adjust relevant policies and behaviors in time, and promote the harmony of the whole society Harmonious and stable development ". Based on this need, it is very important to analyze and grasp the key information quickly when monitoring the public opinion of English forum post information, so this field has become a research hotspot.

In the process of data processing, firstly, the influencing factors of posts are quantified to study the prediction of the development of public opinion popularity. In this paper, the first thing we do is to study the distribution characteristics of the popularity of English forum posts; then we quantify the factors that affect the popularity of posts, and use the regression based vector machine to generate the development trend of public opinion; finally, we test the prediction methods of this paper, and get the results of comparison.

2. A New Regression Prediction Algorithm for the Popularity of English Forum Posts

2.1. User behavior and hot post distribution analysis

In the network community, users can publish articles or posts related to or interested in themselves in various forums. Any user can sign up and participate in the discussion of a topic in the cloud. In the Tianya forum, the research object of this article, users can easily and quickly see the latest status of their friends on their home page, and can also add any user to their friends or pay special attention to them. Friends and people with special attention always appear at the top of the post list for easy search, which forms the phenomenon of "the richer the richer". When users browse in the list of posts, they usually browse from top to bottom according to the order of posts, so the popularity of topics is directly related to the interest of netizens.

2.2. Pretreatment of data

Store the posts in the forum into the collection: Posts = {Posti | i = 1, 2, ..., m}. i represents the serial number of the post, and m is the number of posts. The authors of all the posts in the forum are stored in the collection: Authors = {Authorj | j = 1, 2,⋯, n}. n is the number of authors, and PostofAuthorj is the collection of posts posted by the corresponding author, which meets PostofAuthorj⊆Posts. All the posts in the collection are segmented using the word segmentation machine of the Chinese Academy of Sciences. After the words are removed, the results provided a basis for subsequent research.

2.3. Extraction of post feature

The response of users can be considered as the spread of forum messages. Some posts with thousands of replies and hundreds of thousands of page views have a profound influence on the real society. After research, it is found that in the internet forums, there are many factors influencing the popularity of posts, and any possible influencing factors cannot be ignored. Otherwise, the prediction results may have large errors. It is necessary to combine multiple factors and make a predictive analysis of the popularity of the post. This section is mainly a quantitative analysis of the characteristics that influence the popularity of posts.

User mature and relationship

1) User activity degree

The role of active forum users in post enthusiasm is very obvious, especially the opinion leaders of the forum, who are the users with great influence in the forum. Although they account for a small number of the whole user group, the posted posts or comments often makes the posts become hot posts because ofits popularity. In the network community, the popularity of posts made by opinion leaders is different from those posted by other users.

In this paper, when measuring the users' activity index, the users' relationship network diagram is used to compare the number of interactions between the user node as well as other user nodes. When calculating the importance of these relation nodes, currently, the more usedalgorithms include PageRank, HITS, and statistics-based algorithms. The formula of PageRank algorithm is as follows:

$$LR(A)=1-d+d\left(\frac{LR(U1)}{C(U1)}+\frac{LR(U2)}{C(U2)}+\cdots,+\frac{LR(Un)}{C(Um)}\right) \quad (1)$$

In the above formula (1), LR (A) represents the degree of activity at the user node A; d represents the damping coefficient, which will normally take the value 0.85; C (Ui) represents the relationship with the number of nodes when there is user Ui nodes interaction. Then, the algorithm performs multiple iterations to derive a stable user activity value.

In the HITS algorithm, two new calculation parameters are used: (1) the authoritative value of the post contents, which represents the weight of a node pointed by other nodes; (2) the link authoritative value (Hub), which representsthe centrality of a node pointing to another high authority node. The detailed calculation formula is as follows:

$$A(i)=\sum_{j=1}^{k} H(j) \quad (2)$$

$$H(j)=\sum_{i=1}^{n} A(j) \quad (3)$$

In the above formula, H (j) represents the center value in node j, and its value is the sum of the authority values of all thenodes that pointto the user node. A (i) represents the contents authority central value in node I, which is the sum of the Hubs that pointto node i among all the nodes.

PageRank algorithm and Hits algorithm can obtain more accurate results in the experimental environment, which can accurately find the opinion leaders in the forum. However, in the real

network community, the results of thetwo kinds of algorithms are not as good as the statistical algorithms. The complexity level is smaller than the two algorithmsmentioned above. Thus, this paper uses statistical algorithms to calculate user activityand defines two nodes whichinfluence the features: (1) node degree value, Di = ΣEi (2) average daily posts, Post_per_day = p / days. Between thetwo characteristics, Di represents the number of connections of the node in the relationship network; p is the total number of users who take part in the discussion of posts, and “day”means the time that the user has logged in since the last login.

2) Friends and relationships of users

In the TianYa Forum, because of the existence of the friend mechanism, some users with more followers tend to be opinion leaders in a topic area. In the network community, opinion leaders have a greater influence on the popularity of posts. In general, there will be more people involved in the discussion of opinion leaders, so it is necessary to analyze this phenomenon.

2.4. Support vector machine regression prediction

Support vector machine regression prediction is a classification prediction method on the basis of the VC dimension of statistical learning theory. The reason why the author of this paperadopts this method is that it can obtain a globally unique solution, which no longer depends on the dimensions of the input space. In text processing, it can achieve better results.[7]. Thus, this paper uses this method to predict the popularity of posts.

Before predicting the popularity of posts, the test data was divided into two categories, namely training data and test data. The regression function f was used to model the training data, and then the heat value of the test post data set was calculated. For the convenience of calculation, this paper normalized the features of the text, and the result interval was [0,1]. When the value obtained was greater than 0.5, the corresponding postswere classified as a popular post.

3. Heat Prediction Results and Analysis

3.1. Evaluation Index

When detecting the prediction results, three indicators were used, namely: recall rate, precision rate, and accuracy rate. Set the total number of posts in the test data as N. Among them, there are M active posts. There are L active posts in the prediction results. Compared with M, there are C cases. At this time, the specific judgment methods of these three indicators are as follows:

Recall rate detection:

$$Recall = C / M \quad (4)$$

Precision detection:

$$Precision == C / L \quad (5)$$

Accuracy detection:

$$Accuracy = (2C + N-M-L) / N \quad (6)$$

3.2. Analysis of prediction results

The data set used in this paper mostly comes from the TianYaBBS. The data period is from 2002 to 2011. After processing the data, 107732 available experimental data were obtained. In the course of doing the experiment, the first 5000 data of the experimental data are selected as the training set, and the rest were used as the test data set. Thus, the amount of data is very large. In this paper, the threshold for the number of hot posts was set between 1, 000 and 5, 000. In this paper, the above support vector machine regression prediction model was used for prediction, and the results are shown in Fig. 1.

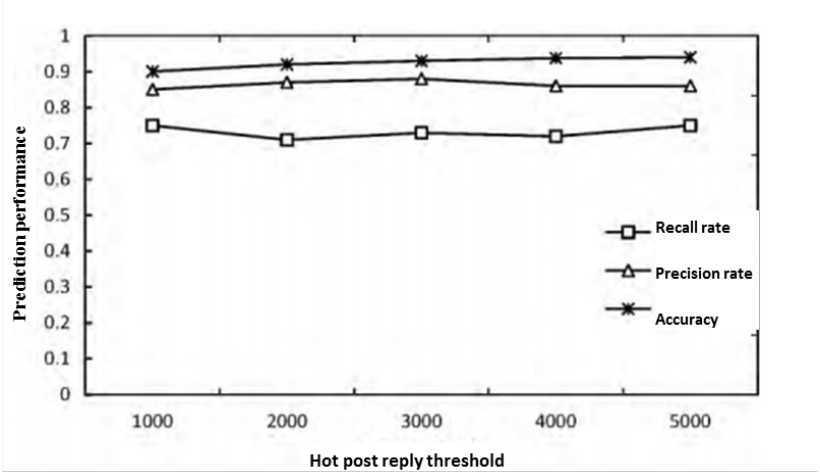

Fig.1. Predicted performance results

From the test results above, it can be seen that when predicting the popularity of posts, the precision rate and the accuracy rate can reach satisfactory values, the accuracy rate is more than 92%, and the recall rate is also about 78%. The value is lower than the accuracy rate. The main reason for this is that in the forum, only a small part of the posts can become active posts, while most of the posts have low user participationdegree.

In order to test the users' activity degree and the role of friend relationship, this paper subtracts the nature of the user and the relationship among the users, and then performs a regression test. The results are shown as follows:

It can be seen from Fig. 2 that when the characteristics of the users and the relationship between the users are not taken into account, the prediction resultsare worse than the results then being taken into account. When the number of hot posts is 1, 000, the accuracy rate is only 30%; under such circumstances, the accuracy rate is not more than 80%. Thus, the activity of forum users and the relationship among users play a very important part in predicting the popularity.

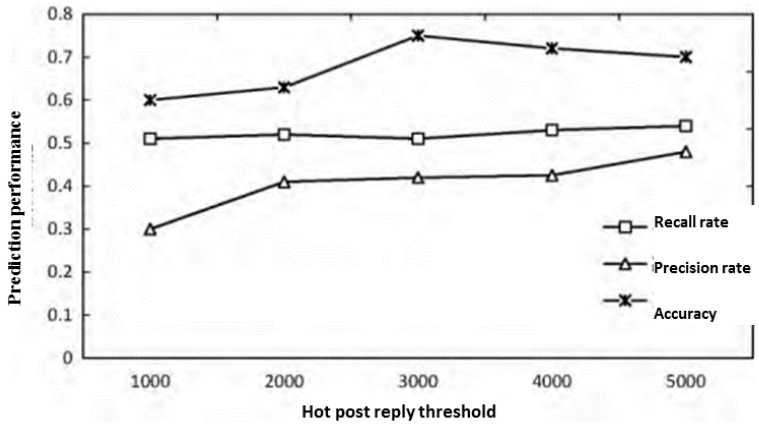

Fig.2. Prediction results of popular degree after elimination of user property relationship

4. Conclusion

Prediction of the popularity of English forum posts has become a key research direction of public opinion monitoring. This paper proposes a new regression prediction algorithm for the popularity of English forum posts. In the process of data processing, firstly, the influencing factors of posts are quantified to study the prediction of the development of public opinion popularity. In this paper, the first thing to do is to study the distribution characteristics of the popularity of English forum posts; then quantify the factors that affect the popularity of posts, and use the regression based vector machine to generate the development trend of public opinion; finally, the prediction method of this paper is tested, and the test results show that the prediction algorithm proposed in this paper has

good accuracy Accuracy.

Acknowledgement

Scientific Research Project of Hunan Education Department

References

[1] Ravi S,Balasubramanian V,Nasser S N. Influences of post weld heat treatment on fatigue life prediction of strength mis - matched HSLA steel welds [J]. International Journal of Fatigue,2005,27 (5): 547-553.

[2] Yun R,Kim Y. Post-dryout heat transfer characteristics in horizontal mini-tubes and a prediction method for flow boiling of CO 2 [J]. International Journal of Refrigeration,2009,32(5):1085-1091.

[3] Wei H,Zhang X. Study on Post-Dryout Heat Transfer by Using Wavelet Neural Network [C]/ / Second International Conference on Innovations in Bio-Inspired Computing and Applications. IEEE,2012:229-232.