Personality Prediction of Social Network Users

Chaowei Li*, Jiale Wan*, Bo Wang†
*Tianjin Polytechnic University, P. R. China
†Hubei University of Economics, P. R. China
*teamfinlab@qq.com, wjl.finlab@qq.com; †bo@vip.suwfc.com

***Abstract*—Through weibo users, we extract social data and questionnaire, and focus on how to use the user text information to predict their personality characteristics. We use the correlation analysis and principal component analysis to select the user information, and then use the multiple regression model, the gray prediction model and the multitasking model to predict and analyze the results. It is found that MAE values of the gray prediction are better than the multiple regression model Multi-task model, the overall effect of the prediction between 0.8 and 0.9, the overall accuracy of good prediction. This shows that gray prediction in the user's personality prediction shows a good generalization and non-linear ability.**

***Keywords*-Social Network; Text Information; Personality; Gray Prediction**

I. Introduction

The popularity of social networks makes people's social changes, making the exchange of people, communication and cooperation between the changes. On the one hand, people can use the network platform to contact friends, comment, discuss public topics and so on. On the other hand, the role of social networks in daily life is increasing, and even to a large extent affected the reconstruction of the network of real social relations. At the same time, because the behavior and status of social networks are easily recorded, acquired and analyzed, social computing has become an important research content in the field of information technology and computer. Because people's behavior and personality are closely linked, making personality prediction has a broad academic value and business prospects. The results show that there is a significant correlation between commercial personality and commodity selection. Personality information can be widely used in advertising and merchandise personalized recommendation [1], [2]. Merchants can push music based on the music preferences of different personality characteristics of users [3]. Thus, the use of social networking to tap the user's social information has become the focus of attention of enterprises and scholars. Personality is a high degree of generalization of different individual characteristics of mankind, even in the same environment, different people will show different behavior, which comes from the different personality of each person. Personality psychology is one of the branches of psychology, mainly through people's external behavior to distinguish between people's intrinsic characteristics, and to study the relationship between them [4]. Psychology usually use personality traits to define people's personality, explain the user's behavior and preferences [5]. Commonly used personality models are MBTI (Myers Briggs Type Indicator) and the big five personality model (openness, agreeableness, conscientiousness, extraversion, neuroticism). The user's social network behavior is closely related to his personality traits. As early as 2000, Hamburger et al. studied the relationship between user behavior and personality. Their research shows that there is a correlation between the user's personality and the social services it uses. Among them, the outgoing personality more inclined to entertaining services; neurotic personality is less use of information services, and more popular social services [6]. [7] used the mouse and keyboard usage habits to predict for personality traits . Since the data of social network users can reflect their true personality traits to a large extent, it is natural to use the user's social data to predict personality. Bernd Marcus believes that the individual's personality can be analyzed using the user's personal website, there is a correlation between the two [8]. [9] used Facebook's social data analysis to find that the extraversion and conscientiousness character had a significant positive relationship with the user's use of social network comfort . [10] through the development of Facebook social networking applications, invite users to participate in the big five personality test, the use of customer information as a data set to study, analysis found that the user's network density does not reflect the user's social status and importance. Scholars use different research methods to predict the relationship between user behavior and their personality traits. In order to predict the personality of 335 users online, Bai and others based on the number of friends and users recently released the state update, using a variety of machine learning algorithms such as Naive Bayesian (NB), support vector machine (SVM) and decision tree. The results show that the C4.5 decision tree algorithm can get the best predictive effect, and verify that the machine learning algorithm can effectively carry out social network user personality prediction [11]. In order to solve the problem of insufficient training data, Verhoeven B et al. used an integrated method to predict the user's personality, and integrated the data of different organizational forms with the integrated method and proved the validity by experiment [12]. Through the existing academic results, we can find that there is a certain correlation between the social network and the user's personality, but the use of social networks to predict the personality of the user's research results are relatively small. Based on these, this paper studies the Sina weibo, extracts the user characteristics and personality characteristics of weibo users, establishes the forecasting model, analyzes and predicts

2473-3636/17 $31.00 © 2017 IEEE
DOI 10.1109/DCABES.2017.25

the weibo user's personality, and verifies the feasibility of the model.

II. THEORY OF PERSONALITY PREDICTION AND DATA EXTRACTION

In a certain period of time, people's behavior is consistent, the consistency of this behavior is precisely the theoretical premise of personality traits, network behavior and personality relevance is the theoretical basis of personality prediction. If you do not have this cross-domain consistency, then the theory of personality prediction will lack the basic theoretical support. The consistency of behavior makes personality prediction an effective tool for predicting individual behavior [13]. Individuals of different personality, their social network attitude and the use of the way also has some differences [14], [15]. But the same personality of the individual for different types of network use situation there is a big difference, and sometimes even completely contrary [15], which makes the personality prediction has a certain theoretical basis, but because in reality, people use different Of the network has the difference, leading to the relationship between network attributes and personality prediction has become confusing. Firstly,We through the users of the Sina weibo fill out the big five personality model questionnaire to obtain the user's personality score.Then, under the user's authorization, we obtain the user's data. Based on the data, first of all,we extract the user's inherent characteristics; Secondly, we go on the emotional analysis to user released weibo, and extract the user's emotional characteristics; Finally, the user relationship is analyzed, the characteristics of the network relationship are extracted, the user characteristics are put into the data set, the user data is marked according to the user's personality score, and finally the user's eigenvector is obtained. The text message which users release in the use of social networking, usually contains personal emotions, and personal personality is closely related. Based on the characteristics of weibo text, this paper puts forward a kind of weibo emotion analysis method based on improved emotion dictionary, carries on the emotion analysis to weibo text published by weibo user, and extracts the emotional characteristics of user weibo.

III. PERSONALITY PREDICTION MODEL

A. Feature selection

As the user in the social network contains more information, the information dimension is higher, so we have to reduce the dimension of the index processing, and processing is divided into two steps. First, the use of correlation coefficient for the selection of indicators, we delete the correlation between the larger indicators, because the greater the correlation, indicating the greater the information between the indicators of redundancy. Secondly,We use the significant coefficient of principal component analysis to select the index, less than 0.9 to be deleted [16]. As the correlation analysis is relatively simple, here we focus on the index screening process of the principal component analysis, the specific process is as follows:

Step 1: Find the correlation coefficient matrix of the normalized index value $R_{m\times m}$.

Step 2: Find the eigenvalue λ_j of the matrix $R(j = 1, 2, 3, ..., m)$, λ_j represents the total variance of the original data index explained by the jth principal component .The contribution of the principal component F_j to the original data is w_j ,given by: $W_j = \frac{\lambda_j}{\sum_{j=1}^{k} \lambda_j}$.

Step 3: The eigenvalues λ_j are arranged in descending order, and the principal components corresponding to the k eigenvalues are selected according to the requirements of the cumulative contribution rate , and the factor load matrix of the ith index on the jth principal component is obtained. Among them: $a_{ij} = \frac{b_{ij}}{\sqrt{\lambda_i}}$.

Step 4: According to the absolute value $|b_{ij}|$ of the factor load on the principal component F_j, the greater the $|b_{ij}|$, the greater the effect of the indication i on the evaluation result, the more should be retained; The smaller the $|b_{ij}|$, the less the impact of the indication on the evaluation results,so the indication should be eliminated. In this paper, the selected significance threshold is 0.9.

B. Type style and Fonts

The personality prediction model is evaluated, mainly considering the fit between the real value and the predicted value. Where the true value is the personality score obtained from the personality questionnaire that the user has filled out, and the predicted value is the result of the prediction by the user characteristic. The predicted results of the continuous prediction model are specific values that can be evaluated by the correlation index and the prediction error index. We use the mean absolute error and root mean square error to represent the prediction error index, the letters are expressed as MAE.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| r - r^{'} \right| \tag{1}$$

Where r is the true score of the user's personality trait and $r^{'}$ is the predicted score.

C. Prediction method

After obtaining the user's behavior data, we can extract the user's characteristic data, such as the use of the frequency of words, the number of exclamation points and other independent variables, recorded as $X_1, X_2, X_3, ..., X_n$, for a specific personality characteristics we set to T, there are $T = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$. Among them,$k \leq n$ this is because the network data acquisition of its large amount of information, high dimension, so inevitably there are a lot of indicators of information less, so that the role of small indicators can be ignored. We use the correlation analysis to a large extent to avoid the problem of multiple collinearity, at the same time, we use the principal component analysis of information significance to filter the indicators, and remove the indicators of low significance indicators, which played a reduced dimension effect. In this paper, gray prediction has four main processes: data conversion, prediction, error

checking. First, in order to ensure the feasibility of the modeling method, it is necessary for us to test the known social network data. We assume that the reference data is $x^{(0)} = x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), ..., x^{(0)}(n)$, Calculate the $\lambda(k)$ value of the sequence,

$$\lambda(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}, \tag{2}$$

where $k = 2, 3, ..., n$.

If all $\lambda(k)$ values fall within the admissible cover $\Theta = (e^{-\frac{2}{n+1}}, e^{\frac{2}{n+2}})$, the sequence $x^{(0)}$ can be greyed out as the data for the model GM. Otherwise, it is necessary to do the necessary processing for the sequence x_0, so that it can fall into the cover can be covered. That is, take the appropriate constant c, do the translation transformation:

$$y^{(0)}(k) = x^{(0)}(k) + c, \tag{3}$$

where $k = 1, 2, ..., n$, let the value of the of the sequence:

$$y^{(0)} = (y^{(0)}(1), y^{(0)}(2), ..., y^{(0)}(n)) \tag{4}$$

When you create a GM model, you can get predictions:

$$\bar{x}^{(1)}(k+1) = (x^{(}0)(1) - \frac{\bar{b}}{\bar{a}})e^{-\bar{x}k} + \frac{\bar{b}}{\bar{a}}, \tag{5}$$

$$\bar{x}^{(0)}(k+1) = (x^{(}1)(k+1) - \bar{1}(k), \tag{6}$$

where $k = 1, 2, ..., n-1$.

Finally, test the predicted values. The third method we use is the multitasking regression method. The core idea of multitasking regression is to carry out multiple learning tasks in the same training set, to establish the smallest transfer matrix, and to use weibo users to predict and analyze the data. For weibo user data, we use multiple linear regression, gray prediction, multi-task regression three ways to predict. In order to compare with the research results of different scholars, we briefly list the research results of some scholars at present, as shown in Table IV. The main results of the predict and other scholars are as follows, as shown in Table I, II, III.

TABLE I
PLE REGRESSION PREDICTION RESULTS

Personality traits	MAE
Openness to experience	0.2315
Conscientiousne	0.2099
Extraversion	0.3015
Agreeableness	0.2047
Neuroticism	0.192

TABLE II
GRAY PREDICTION RESULTS

Personality traits	MAE
Openness to experience	0.1263
Conscientiousness	0.162
Extraversion	0.1308
Agreeableness	0.1008
Neuroticism	0.1905

TABLE III
MULTITASKING REGRESSION MODEL PREDICTION

Personality traits	MAE
Openness to experience	0.1375
Conscientiousness	0.1584
Extraversion	0.1799
Agreeableness	0.1608
Neuroticism	0.1590

TABLE IV
RELATED RESEARCH RESULTS

Author	Platform	Algorithm	Prediction result
Golbeck [17]	Twitter	Gaussian process and ZeroR	MAE is about 11%
Bachrach [18]	Facebook	Multiple linear regression, decision tree based rule set, SVM and decision tree	RMSE = 0.27
Wald [19]	Twitter	Linear regression, REPTree and decision tables	For open up to 10% of the user,we can predict 74.5% of them.
Ortigosa [20]	Facebook	Naive Bayesian, C4.5	Three classification accuracy of about 70%, five classification accuracy of about 63%.

Through the results of the three prediction methods, we can see that the three methods are different in the accuracy of the prediction, and there is a big difference between the prediction error and the root mean square error. The traditional multiple regression model, although simple and easy to operate, its predictive results are not ideal, especially for non-linear prediction, it is often difficult to be satisfactory. Although multi-task regression prediction model regard the five personality traits as a whole for the forecast analysis, there is a large information missing. Due to its good generalization ability, gray prediction has excellent fitting ability in nonlinear regression, and still has good results when the state of social data is poor, and therefore worthy of promotion and use. It is seen Fig. 1 that the multitasking regression prediction model also shows good predictive ability and good predictive performance in user personality prediction.

By comparing with the predictive results of some scholars, our model is slightly less than 11%. However, the MAE of the three methods is less than 0.2, and the prediction effect is between 0.8-0.9 and the prediction accuracy is good. Through the three methods used in this paper and other scholars' machine learning methods, we found that, in addition to method selection, the prediction accuracy also includes data acquisition, data cleaning, feature selection and problem complexity.

IV. Conclusion

On the basis of the existing research, we focus on the social network and user behavior characteristics, and establish three prediction models for comparative analysis. We found that the behavioral characteristics of the user and their personality are related, through the study of social network data can be a large extent predict the user's personality characteristics, which can

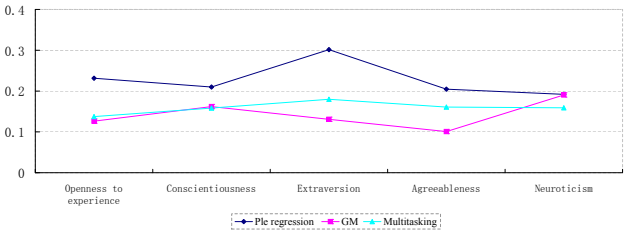

Fig. 1. Error curve

further promote the development of personalized applications. Aiming at the characteristics of weibo data volume and high data dimension, we first use correlation analysis and principal component analysis to select the feature. On this basis, we established a multiple regression model, gray prediction model and multitask regression model. Through the prediction index of MAE, gray is less error in nonlinear prediction, and it is more suitable for the prediction and research of user personality traits in social network.

ACKNOWLEDGMENT

The work described in this paper was supported by Tianjin Polytechnic University Research Development Fund under Grant No. 17140/2017. The authors would like to express their appreciation for Hailin Qin for his thoughtful discussions.

REFERENCES

[1] G. Odekerken-Schrder, K. D. Wulf, and P. Schumacher, "Strengthening outcomes of retailerconsumer relationships : The dual impact of relationship marketing tactics and consumer personality," *Journal of Business Research*, vol. 56, no. 3, pp. 177–190, 2003.

[2] S. Whelan and G. Davies, "Profiling consumers of own brands and national brands using human personality," *Journal of Retailing and Consumer Services*, vol. 13, no. 6, pp. 393–402, 2006.

[3] P. J. Rentfrow and S. D. Gosling, "The do re mi's of everyday life: The structure and personality correlates of music preferences." *Journal of Personality and Social Psychology*, vol. 84, no. 6, p. 1236, 2003.

[4] G. Matthews, I. J. Deary, and M. C. Whiteman, "Personality traits (2nd. ed.)." 2003.

[5] G. W. Allport, "The general and the unique in psychological science." *Journal of Personality*, vol. 30, no. 3, p. 405422, 1962.

[6] Y. A. Hamburger and E. Ben-Artzi, "The relationship between extraversion and neuroticism and the different uses of the internet," *Computers in Human Behavior*, vol. 16, no. 4, pp. 441–449, 2000.

[7] I. A. Khan, W.-P. Brinkman, N. Fine, and R. M. Hierons, "Measuring personality from keyboard and mouse use." ACM, 2008, p. 38.

[8] B. Marcus, F. Machilek, and A. Schtz, "Personality in cyberspace: Personal web sites as media for personality expressions and impressions." *Journal of Personality and Social Psychology*, vol. 90, no. 6, pp. 1014–31, 2006.

[9] C. Ross, E. S. Orr, M. Sisic, J. M. Arseneault, M. G. Simmering, and R. R. Orr, "Personality and motivations associated with facebook use," *Computers in Human Behavior*, vol. 25, no. 2, pp. 578–586, 2009.

[10] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," 2011, pp. 253–262.

[11] C. L. Bai S, Zhu T, "Big-five personality prediction based on user behaviorsat social network sites," *Plos Neglected Tropical Diseases*, vol. 8, no. 2, p. 2682, 2012.

[12] B. Verhoeven, W. Daelemans, and T. D. Smedt, "Ensemble methods for personality recognition," 2013, pp. 35–38.

[13] M. D. Back, J. M. Stopfer, S. Vazire, S. Gaddis, S. C. Schmukle, B. Egloff, and S. D. Gosling, "Facebook profiles reflect actual personality, not self-idealization." *Psychol Sci*, vol. 21, no. 3, p. 372, 2010.

[14] T. Ryan and S. Xenos, "Who uses facebook? an investigation into the relationship between the big five, shyness, narcissism, loneliness, and facebook usage," *Computers in Human Behavior*, vol. 27, no. 5, pp. 1658–16 642 009, 2011.

[15] T. Correa, A. W. Hinsley, and H. G. D. Ziga, "Who interacts on the web?: The intersection of users personality and social media use," *Computers in Human Behavior*, vol. 26, no. 2, pp. 247–253, 2015.

[16] Q. Wang, H. N. Dai, and H. Wang, "A Smart MCDM Framework to Evaluate the Impact of Air Pollution on City Sustainability: A Case Study from China," *Sustainability*, vol. 9, no. 6, 2017.

[17] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *IEEE Third International Conference on Privacy, Security, Risk and Trust*, 2013, pp. 149–156.

[18] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell, "Personality and patterns of facebook usage," in *Acm Web Science Conference*, 2012, pp. 24–32.

[19] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner, "Using twitter content to predict psychopathy," in *International Conference on Machine Learning and Applications*, 2013, pp. 394–401.

[20] A. Ortigosa, R. M. Carro, and J. I. Quiroga, "Predicting user personality by mining social interactions in facebook," *Journal of Computer and System Sciences*, vol. 80, no. 1, pp. 57–71, 2014.