Personality Prediction Based on All Characters of User Social Media Information

Danlin Wan, Chuang Zhang, Ming Wu, and Zhixiang An

Beijing University of Posts and Telecommunications, Beijing 100876, China
{wandanlin2014,azx-c}@163.com,
{zhangchuang,wuming}@bupt.edu.cn,

Abstract. In recent years, the number of social networks users has shown explosive growth. In this context, social media provides researchers with plenty of information about user behavior and social behavior. We are beginning to understand user's behavior on social media is related to user's personality. Conventional personality assessment depends on self-report inventory, which costs a lot to collect information. This paper tries to predict user's Big-Five personality through their information on social networks. We conducted a Big-Five personality inventory test with 131 users of Chinese social network Sina Weibo, and crawled all of their Weibo texts and profile information. By studying the relevance between all types of user generated information and personality results of users, we extracted five most relative dimensionalities and used machine learning method to successfully predict the Big-Five personality of users.

Keywords: Social media, Personality prediction, Machine learning.

1 Introduction

Social media is the social interaction among people in which they create, share or exchange information and ideas in virtual communities and networks. Nowadays, people use more and more time in social media, every time there is a flood of information creation and dissemination. These information includes user's social behavior, user generated texts and language habit, to some extent reflects user's personality.

Personality uniquely characterizes an individual, and profoundly influences user's mental status and social behaviors [1]. Conventional personality assessments use self-reported inventory [2]. Self-reported method has a profound theoretical significance, but it costs long time and lots of manpower.

This paper tries to use user generated information on social network which is easy to get to predict user's personality.

First core research point in this paper is why user generated information on social network can be used to predict user's personality, in other words, the relationship between information on social network and user's personality. Previous work has shown that the information in users' Facebook profiles is reflective of their actual personalities, not an "idealized" version of themselves [3]. Gosling et al. [10]

H.-Y. Huang et al. (Eds.): SMP 2014, CCIS 489, pp. 220–230, 2014.

delivered a mapping between 11 features of users' online behaviors on Facebook and users' personality. They verified there is a correlation between them.

Second core research point in this paper is how to use user generated information to predict user's personality. Most recent research tends to use one aspect of information on social network to predict user's personality. Globeck tried to predict web users' personality traits through text features on Facebook and Twitter [5], [6]. Quercia proposed to predict web users' personality traits through three features (i.e., following, followers and listed counts) available on profiles of Twitter [7]. Shuotian Bai and Bibo Hao used multi-task regression method to predict user's personality with their behavioral characteristics on Sina Weibo [8].

In this paper, we are interested in Chinese social media. Sina Weibo is one of the most popular sites in China. Akin to a hybrid of Twitter and Facebook, it is in use by well over 30% of Internet users, with a market penetration similar to the United States' Twitter [9]. Sina Weibo has been the leading micro blogging service provider in China. By the time March 2014, Sina Weibo had 143.8 million monthly active users, 66.6 million active users. So it's chose as the major source of user generated information.

In order to achieve these two main research cores. We conducted a Big-Five personality inventory test with 131 users of Sina Weibo, through result analysis, we got their Big-Five personality dimensions scores. We take these scores as ground truth. Then we crawled all of their Weibo texts and profile information including text features, social behavior features and interaction features. Through a Pearson correlation analysis between all features and user's Big-Five personality dimensions scores, we verified there is a correlation between them. According to the correlation results, we extracted the first five relative features and used logistic regression and Naive Bayes algorithms to learn and successfully predict user's personality.

In psychology, the Big Five personality traits are five broad domains or dimensions of personality that are used to describe human personality. The theory based on the Big Five factors is called the Five Factor Model (FFM) [11]. It was one of the most well-researched and well-regarded measures of personality structure in recent years. Tupes and Christal came up with five domains of personality, Openness, Conscientiousness, extroversion, Agreeableness, and Neuroticism first through analyses of previous personality tests [12], [13]. Latter research has proved that different tests, languages, and methods of analysis do not alter the model's validity [14], [15], [16]. Such comprehensive research has led to many psychologists to accept the Big Five as the current definitive model of personality [17], [18]. Following is a summary of the factors of the Big Five and their constituent traits.

- Openness to experience: Openness reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety a person has.
- Conscientiousness: A tendency to be organized and dependable, show self-discipline, act dutifully, aim for achievement, and prefer planned rather than spontaneous behavior.

- Extraversion: Energy, positive emotions, surgency, assertiveness, sociability and the tendency to seek stimulation in the company of others, and talkativeness.
- Agreeableness: A tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others. It is also a measure of one's trusting and helpful nature, and whether a person is generally well tempered or not.
- Neuroticism: The tendency to experience unpleasant emotions easily, such as anger, anxiety, depression, and vulnerability. Neuroticism also refers to the degree of emotional stability and impulse control and is sometimes referred to by its low pole, "emotional stability".

2 Related Studies

In the area of predicting user's personality according to their online behavior of social network, most of foreign research is based on Facebook or Twitter.

Globeck shown that users' Big Five personality traits can be predicted from the public information they share on Facebook. They collected fifty users' text information on Facebook and their Big-Five personality results, then used two regression algorithms ZeroR and Gaussian Processes to predict scores on each of the five personality traits [5], [6]. They can predict scores on each of the five personality traits to within 11% - 18% of their actual value.

Shuotian Bai and Bibo Hao used multi-task regression method to predict user's personality with their behavioral characteristics on Sina Weibo [8]. They examined the personality inventory test of 444 users, and extracted users' profile information and interaction information. Then used Muti-task regression and incremental regression to predict the Big-Five personality.

All in all, the current researches on this area are more concentrated on only one type of information on social network, such as text features or profile features. So they can't reveal the relationship between all types of information and personality dimensions. In this paper, we choose most benefit features for each personality dimensions, and successfully predict user's personality.

3 Method

The overall frame work in our research is shown in Figure 1. It shows the main procedures to collect user's personality and Social network information and how we use it to predict user's personality.

3.1 Participants

The survey was conducted during March 2014 to May 2014. A total of 589 individuals participated in our test, and 131 participants were recruited. Because this research

defines qualified participants as who has more than 200 Weibo statuses and whose answer time was longer than 100 seconds.

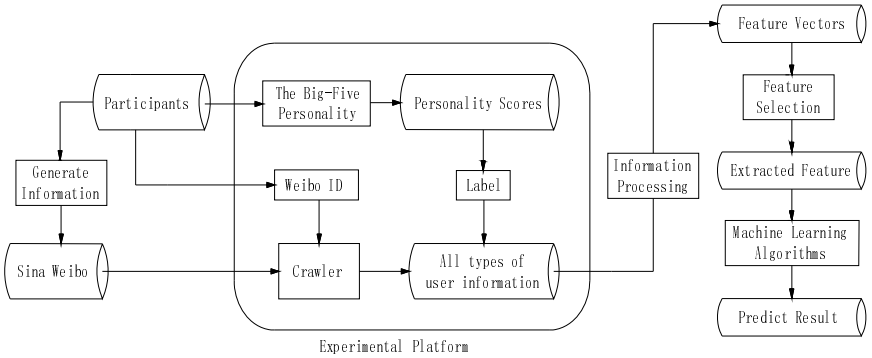

Fig. 1. Frame work of personality prediction

3.2 The Big-Five Personality Scores

Using standard Big-Five personality test scale, we conducted a personality test of Sina Weibo users. Standard Big-Five scale contains of 60 multiple-choice questions, each question on one dimension for evaluation. Options for each question are strongly agree, agree, not sure, do not agree, strongly disagree, to show measured by the bias in the measured dimension.

Personality values for each dimension range from -24 to 24. The larger positive value indicates more positive related, the larger negative value indicates more negative related, closer to 0 indicate that the characteristic dimensions is not obvious. Each dimension was divided to -24 to -8, -8 to 8, 8 to 24, these three levels representing a negative prominent, not obvious and positive prominent. In the following, we will use -1, 0, and 1 to represent these three levels.

3.3 Features Extracted from User Generated Information

In order to obtain user's generated information on Sina Weibo, we used crawler to get all of the 131 participants' profile information and Weibo text information.

All types of information can be categorized into 3 groups. The first group is user behavior which contains Weibo statues count, followers count, followings count, time since registration Sina Weibo. These information reflects user's basic use condition of Sina Weibo. The second group is interaction behavior which extracted from text information and contains expressions count, topics count and @ mentions. The third group is text features which reflect user's language habits on social network.

For the purpose of analyzing the content of Weibo content, we try to use the Linguistic Inquiry and Word Count (LIWC) dictionary [19]. LIWC produces statistics on 71 different features of text in five categories. These include Standard Counts (word count, words longer than six letters, number of prepositions, etc.), Psychological

Processes (emotional, cognitive, sensory, and social processes), Relativity (words about time, the past, the future), Personal Concerns (such as occupation, financial issues, health), and Other dimensions (counts of various types of punctuation, swear words). We can use this dictionary to catch characteristics of the text. So we choose LIWC2007 Simplified Chinese Dictionary. Firstly, using IKanalyzer tool (a Chinese word segmentation tool) to get word segmentation results. Then bring word segmentation results to match LIWC2007 Simplified Chinese Dictionary, and get all mapping count of each feature of LIWC.

The second group and third group are both extracted from text information, we need some processes to obtain standard text features. In second group, three features' average frequency was computed in every status. In third group, through dividing mapping count by user's total words count, we get 71 LIWC features' frequency in per Weibo status of each user.

Finally, through information collecting, processing and computing, we collected 131 participants' 5 dimensions personality scores, 4 features about user behavior, 3 features about interaction behavior and 71 text features.

3.4 Feature Selection

Considering the size of the data set and the numbers of extracted features, we can't use all these features to predict user's personality. So we need to select features. In this paper, we use Pearson Correlation Coefficient as our standard to select features based on dependency metrics theory. Besides, Pearson Correlation Coefficient can reveal the relationship between all types of information and user's personality.

Pearson correlation coefficient describes the degree of tightness between two fixed variables and defined as

$$p = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} \tag{1}$$

In the use of this formula, we replaced X_i with extracted features values and replaced Y_i with different personality dimensionalities scores. And $\bar{X}$ and $\bar{Y}$ representative the mean values of them.

The Pearson Correlation value ranges from -1 to +1. If it is positive, the two variables are positively correlated (the greater the value of one variable, the greater the value of another variable). If it is negative, the two variables are inversely related (the smaller the value of a variable, the greater the value of another variable).

We firstly analyzed the Pearson Correlation Coefficient between 3 groups' information and personality scores. The results are show in Table 1, Table 2 and Table 3. Because text information has 71 features, so we only choose highly relevant characteristics to show.

Table 1. Pearson Correlation values between personality scores and interaction behavior. 3 features of interaction behavior mean frequency in every status. Bold font indicates a relatively high degree of correlation.

	Expressions	Topic	@
Neur.	-0.012	0.122	0.057
Extr.	**0.172**	0.041	**0.148**
Open.	**-0.16**	-0 .01	-0.084
Agree.	**0.161**	0.032	-0.045
Cons.	-0.041	-0.085	-0.008

Table 2. Pearson Correlation values between personality scores and user behavior. Statues means user's statues count, followers and followings mean user's followers or followings count. Time means time since registration Sina Weibo. Bold font indicates a relatively high degree of correlation.

	Statues	Followers	Followings	Time
Neur.	**-0.124**	-0.085	0.107	-0.042
Extr.	0.012	-0.011	0.02	**0.158**
Open.	-0.112	-0.091	-0.071	**-0.145**
Agree.	0.068	**0.15**	0.11	**0.173**
Cons.	0.066	0.02	**-0.127**	0.091

Table 3. Pearson Correlation values between personality scores and text features. Bold font indicates a relatively high degree of correlation.

	anger	sexual	certain	body	ProgM	home
Neur.	**0.23**	**0.235**	0.203	**-0.25**	**-0.204**	-0.201
	nonfl	MultiFun	time	swear	anger	sexual
Extr.	**0.178**	**0.152**	0.137	**-0.165**	**-0.144**	-0.138
	space	work	motion	see	humans	funct
Open.	0.191	0.165	0.132	**-0.348**	**-0.277**	**-0.24**
	feel	discrep	body	space	motion	relativ
Agree.	**0.242**	**0.231**	**0.174**	**-0.183**	-0.172	-0.17
	ProgM	feel	motion	funct	i	tentat
Cons.	0.155	0.132	0.111	**-0.222**	**-0.193**	**-0.192**

Through these three tables, we can see that different features are related to different personality dimensionalities. And not all features are related to personality. In interaction behavior group, expressions show relatively high correlation with 3 personality dimensionalities. In user behavior group, time shows relatively high correlation with 3 personality dimensionalities. And text group has most features that have relatively high

correlation with personality dimensionalities. And most of the relatively high correlation values are range from 0.17 to 0.35. These values mean there are relations between user generated information on Social media and personality and the correlation intensity is weak correlation. The results correspond to the research on Twitter [5].

Many of the correlations make intuitive sense. For example, neuroticism is positively correlated with words about anger (e.g. "fury", "rage"), suggesting neurotic people tend to express more anger emotion on social media. At the same time, the words about anger are negatively correlated with extraversion, indicating extraverted people tend to express less about anger emotion. Extraverted people also tend to use more expressions and @ mentions in their Weibo statues.

In order to get better predict result, we selected five features for each personality dimensionality to predict user's personality. The criterion is choosing high Pearson Correlation values features, at the same time, every personality dimensionality should have positive related features and negative related features. The eventual result shows in Table 4. All three types of features are used, it means different features match different personality dimensionalities. So we can't use only one type features to predict all five personality dimensionalities.

Table 4. Top five high Pearson Correlation values between personality scores and all types' features of every personality dimensionalities. Bold font indicates features except text features.

	sexual	anger	certain	body	ProgM
Neur.	0.24	0.23	0.203	-0.25	-0.204
	swear	anger	nonfl	**Expressions**	**Time**
Extr.	-0.2	-0.144	0.178	0.172	0.158
	space	work	see	funct	humans
Open.	0.19	0.165	-0.346	-0.24	-0.277
	feel	discrep	space	**Time**	motion
Agree.	0.24	0.231	-0.183	0.173	-0.172
	ProgM	feel	funct	i	tentat
Cons.	0.16	0.132	-0.222	-0.193	-0.192

3.5 Personality Prediction

In order to predict the scores of five personality dimensionalities, we divided our dataset into training set and test set and performed machine leaning test on Weka [20], which is a software of machine learning and data mining. Using extracted five features for different personality dimensionalities to predict personality. The predicting outcomes for one user are like the Fig. 2. Each user has five scores for five personality dimensionalities.

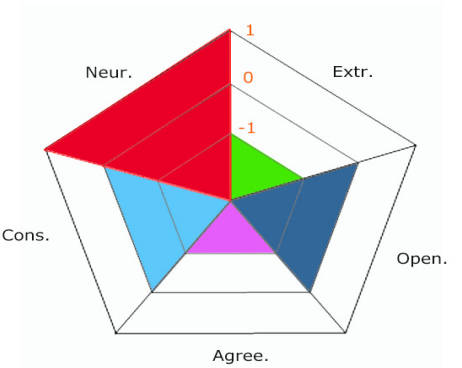

Fig. 2. Predicting outcomes for one user

We used two algorithms: Logistic Regression and Naïve Bayes to training data, each algorithm with a 10-fold cross-validation with 10 iterations. Then used them to compute personality results of test set, and used personality test result which is considered as ground truth in this paper to verify the predicted results. The precision and recall of each algorithm is shown in Fig. 2 and Fig. 3.

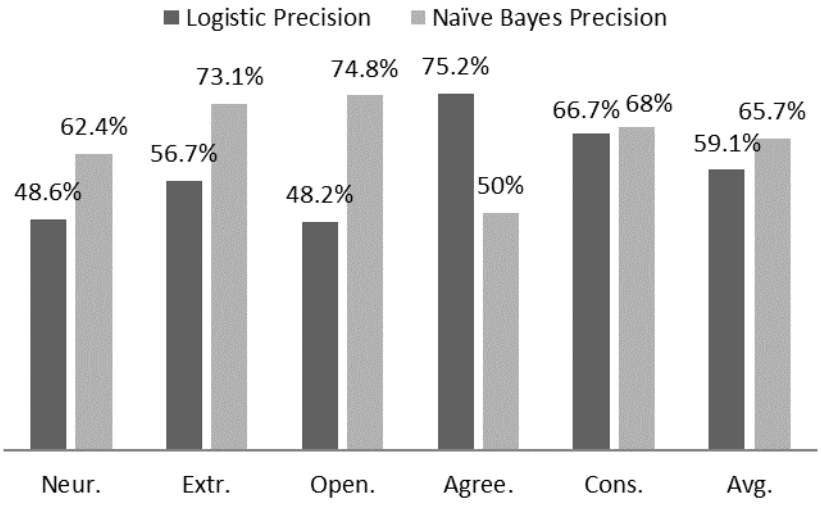

Fig. 3. Precision for each algorithm and personality dimensionalities

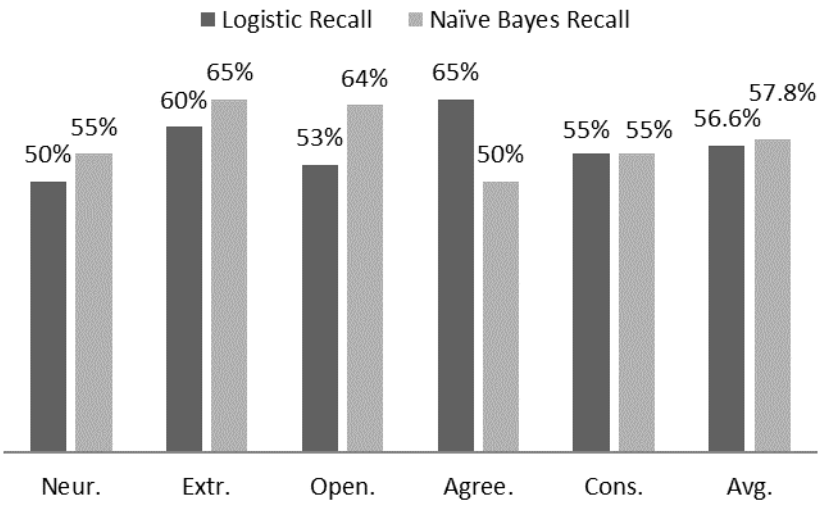

Fig. 4. Recall for each algorithm and personality dimensionalities

We can find that in general Naïve Bayes has better performance on precision, and two algorithms have similar performance on recall. Considering best result of two algorithms, we can find that neuroticism was the most difficult to predict and Openness and agreeableness are easy to compute. This is similar with the result in [5].

Although we have much more participants and consider more types of information on social networks than the research [5], we get a little bit lower prediction precision than their research. We believe part of the reason is Chinese semantic analysis is more difficult than the English semantic analysis, because Chinese semantic analysis needs to segment words according to complex language custom and it will generate error.

4 Conclusions and Future Work

This paper has shown that the Big-Five personality can be predicted by user generated information on Social media. Participants accomplished the standard Big-Five personality test and we collected their interaction behavior, user behavior and Weibo text information. Through computing the Pearson Correlation values between all three types of information and each personality dimensionality, we proves that there is a relationship between them and different personality dimensionality matches different types of features on Social media. We used two machine learning algorithms to predict scores of personality with extracted features, the mean precision of prediction of five personalities was 0.707.

In the future, we will continue recruit more participants and collect their information on Sina Weibo in order to get a larger dataset. And we will improve the

algorithms to predict personality. Our main research direction is to reveal the relation between use's personality and their strong ties friends' personality on social network. And if there is a relation, we can use it to get a better performance to predict user's personality.

References

1. Goldberg, L.: The structure of phenotypic personality traits. American Psychologist 48(1), 26 (1993)
2. Back, M., Stopfer, J., Vazire, S., Gaddis, S., Schmukle, S., Egloff, B., Gosling, S.: Facebook Profiles Reflect Actual Personality, Not Self-Idealization. Psychological Science 21(3), 372 (2010)
3. Burger, J.: Personality. Thomson Wadsworth, Belmont (2008)
4. Li, L., Li, A., Hao, B., Guan, Z., Zhu, T.: Predicting Active Users' Personality Based on Micro-Blogging Behaviors. PLoS ONE 9(1), e84997 (2014), doi:10.1371/journal.pone.0084997
5. Golbeck, J., Robles, C., Edmondson, M., Turner, K.: Predicting personality from Twitter. In: Proceedings of 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (Passat) and 2011 IEEE Third International Conference on Social Computing (Socialcom) (2011)
6. Golbeck, J., Robles, C., Turner, K.: Predicting personality with social meida. In: Proceedings of CHI 2011 Extended Abstracts on Human Factors in Computing Systems (2011)
7. Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J.: Our Twitter profiles, our selves: Predicting personality with Twitter. In: Proceedings of 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (Passat) and 2011 IEEE Third International Conference on Social Computing (Socialcom) (2011)
8. Bai, S., Hao, B., Li, A., et al.: Predicting Big Five Personality Traits of Microblog Users. In: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 1, pp. 501–508. IEEE (2013)
9. Rapoza, K.: China's Weibos vs US's Twitter: And the Winner Is? Forbes (May 17, 2011) (retrieved August 4, 2011)
10. Gosling, S.D., Augustine, A.A., Vazire, S.: Manifestations of personality in online social networks: Self-reported facebook-related behaviors and observable profile information. Cyberpsy-chology, Behavior, and Social Networking 14(9), 483–488 (2011)
11. Costa Jr., P.T., McCrae, R.R.: Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual. Psychological Assessment Resources, Odessa (1992)
12. Tupes, E., Christal, R.: Recurrent personality factors based on trait ratings. Journal of Personality 60(2), 225–251 (1992)
13. McCrae, R., John, O.: An introduction to the five-factor model and its applications. Journal of Personality 60(2), 175–215 (1992)
14. Digman, J.: Personality structure: Emergence of the five-factor model. Annual Review of Psychology 41(1), 417–440 (1990)
15. John, O.: The Big Five factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In: Handbook of Personality: Theory and Research, vol. 14, pp. 66–100 (1990)

16. McCrae, R.: Why I advocate the five-factor model: Joint factor analyses of the NEO-PI with other instruments. In: Personality Psychology: Recent Trends and Emerging Directions, pp. 237–245 (1989)
17. Schmitt, D., Allik, J., McCrae, R., Benet-Martinez, V.: The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. Journal of Cross-Cultural Psychology 38(2), 173 (2007)
18. Schrammel, J., Köffel, C., Tscheligi, M.: Personality traits, usage patterns and information disclosure in online communities. In: BCS HCI 2009: Proceedings of the 2009 British Computer Society Conference on Human-Computer Interaction, pp. 169–174. British Computer Society, Swinton (2009)
19. Pennebaker, J., Francis, M., Booth, R.: Linguistic inquiry and word count: LIWC 2001. Lawrence Erlbaum Associates, Mahway (2001)
20. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)