Computing Personality Trait Based on Multi-source

Zheng Chuanqin*
Network Information Center, Xiamen Medical College, Xiamen, Fujian, 361021, China
zcq@xmmc.edu.cn

Abstract. Personality computing is a hot research field recently, which mainly uses the individual’s traces on social platforms to understand, predict and analyze their behaviors, to make a certain accurate judgment of user personality types. The application of personality traits is of great significance to intelligent medical, personalized service customization, and other fields. The current personality computing research is mainly based on social media, the data source is limited, can not reflect the real situation. This paper collects college volunteers’ network access logs, the data is more comprehensive and complex, not limited to only the traces on social media. According to these network access logs, the calculation results show that the personality traits of users in different social media are not consistent, and the access preferences of network resources are not exactly the same as the personality traits reflected by social media. Therefore, it is more accurate to integrate multi-source network information to calculate an individual personality than a single data source.

1. Introduction
Personality trait is a method to identify personality traits by mining personality data. It plays an important role in many fields, such as intelligent medicine, personalized service customization and so on[1-3]. As the mainstream of personality Trait, self reporting scale has the following shortcomings: it needs to consume a lot of human and material resources; users have subjective factors or avoid the real situation, resulting in large errors, etc[4-5].

With the development of Internet technology and computer technology, many experts and researchers have studied personality computing by using explicit behavior data on the Internet, and achieved remarkable research results. It includes personality trait for social text analysis, personality trait for news comment analysis and personality trait for behavior data analysis of blog users[6-9].

However, most of these studies are limited to a single type of data, such as focusing on the open network data of social networking sites. These studies are bound to be biased: people are easily affected by the atmosphere, and then the emotional tendency of their published content are also affected[10-13].

We study the method of computing personality traits based on the network access data of multiple sources. We collected the network access logs of college volunteers for a period of time. Firstly, we classify the pages visited into different types of network sources, such as social media, news community, interest community, etc. Then, different weights are given to the classification of sources, such as social media are given higher weight, while some public pages such as school home page are given lower weight.
All text of each web page will be scanned, segmented and analyzed its emotional tendency. Thus a score of the psychological dimension of the web pages is calculated. And then all possible re-accessed web pages within the network access log are aggregated to get the personality tendency of the visitor.

We asked 122 volunteers to join our research and got their authorization to use their network access log. They also filled the self reporting scale, which will be the benchmark of personality tendency. We calculated their personality trait with the network access log based on multi-source. The results show that: the personality trait of visitors on different social media is not consistent; and the personality trait calculated by comprehensive multi-source records should be closer to the authenticity of visitors.

The following of this paper are organized as follows: Section 2 introduces some background knowledge and related research status of personality computing; section 3 gives the research methods we use in detail; section 4 is the experiment and result analysis; finally, the summary and prospect.

2. Background

2.1. Five factor personality model

Five factor personality model theory, also known as big five personality theory, describes an individual's personality from five dimensions: Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness. Table 1 shows the sub-dimension of Five factor personality model.

| Dimension     | Sub-Dimension                          |
|---------------|----------------------------------------|
| Neuroticism   | Anxiety, angry hostility, depression, self-awareness, impulse, vulnerability |
| Extraversion  | Enthusiasm, gregariousness, arbitrariness, activeness, seeking stimulation, positive emotion |
| Openness      | Fantasy, aesthetics, emotion, action, concept, value |
| Agreeableness | Trust, frankness, altruism, obedience, modesty, sensitivity |
| Conscientiousness | Ability, rules, due diligence achievement, self-discipline and prudence |

2.2. Personality trait with language characteristics

The relationship between language characteristics and personality of users in social networks is a hot topic. Schwartz et al. studies the relationship between two language features (n-grams phrases and theme words) and personality. On this basis, Schwartz et al. Used the differential language analysis (DLA) technology of open phrases and words in the latest research of PLoS One, and found that language features can also predict the big five personality[14]. Table 2 shows the relationship between personality factors and the language characteristics.

| Personality Factor | Language characteristics                                                                 |
|--------------------|------------------------------------------------------------------------------------------|
| Neuroticism        | Total words, negative emotions, physiological processes, listening, feeling, religion, exclamation, anxiety, eating, swearing, anxiety, health, brackets, |
| Extraversion       | Question mark, bracket, perceptual process, social activity, family, health, SMS average number of words, friend, diet, consent, exclamation point, |
| Openness           | exclamation mark, bracket, physiological process, cause and effect, certainty, body, work, money, family, anger, hearing, religion, death |
| Agreeableness      | Second person, exclamation, emotional process, positive emotion, physiological process, one cause and effect, one food, one money |
| Conscientiousness  | Second person, one auxiliary verb, one future situation, one negotiation, colon, exclamation mark, one negative emotion, one positive emotion, one anger, action, space, time, one swearing, feeling, one cognitive mechanism, one perceptual process, social activity, work, family, one death, one look, relativity |
3. Personality Trait on Multi-source

3.1. Procedure of the personality Trait based on Multi-source

Figure 1 shows the procedure of the personality Trait based on multi-source. At first, the network access log is cleaned to remove those invalid accessing record, such as URLs of non-web-page. Remaining URLs is collected into a valid pages log.

A crawling tool is used to crawl all URL contained in valid pages log. The crawler will capture web page text content, and integrate all text form same website into a large text.

Using Jieba word segmentation tool and SC-LIWC psychology dictionary, the frequency and proportion of language emotional features in sc-liwc are counted as the feature of browsing web page text content.

Thus, personality tendency of each website can be calculated. Those websites will be classified in advance. Each website will be assigned a weight according to its effect on indicate personality tendency. At last, we will calculate the weighted average of personality tendency of all website, which is the final result of the person’s personality tendency.

3.2. Website Classification and Getting final result

Different websites can reflect different personality trait. First of all, the functions of websites are different. Some websites belong to social media, and users often discuss, express their opinions or even argue on certain topics, so they can clearly show users’ emotional tendencies. And some websites are functional, such as operating their own account on the bank page and shopping on the E-commerce website. These operations generally have no emotion, so they can't express the user's emotion tendency obviously. And some websites, users mainly browse the operation, read the content they are interested in, but will not express their opinions, such as news websites, or other interested websites. These pages can reflect some potential emotions of users, but they are not very obvious.

Therefore, we can classify the websites and give a weight to each kind of websites to identify the ability of these websites to reflect users' emotions.

We make a simple classification of some of the most common websites in the network access log. Table 3 shows the website classification and their weights. Commonly, the website that can express emotion tendency clearly has higher weight.

According to the actual data, the most visited web pages belongs to expressive class, and the RMSE of the their personality trait and benchmark is the smallest. Thus, we assign the weight of expressive class to be the highest. However, web pages of reading class and hybrid class have also a lot of access, and the RMSEs are also small. Moreover, the RMES on certain factor of personality trait is smaller than expressive class. Therefore we assign them a certain weight. The web pages of
functional class are not accessed much frequently. Their RMSE of personal trait with benchmark is much higher. The weight of functional class is much smaller than others.

| Class      | Typical website          | Weight |
|------------|--------------------------|--------|
| Expressive | Weibo, Weixin, Zhihu, QQ, Tieba, ... | 60%    |
| Reading    | Sina, Sohu, 163, qidian,... | 20%    |
| Mixed      | Tieba, Music, Games,...   | 9%     |
| Functional | Taobao, Jingdong, Banks,... | 1%     |

Table 3. Website classification and their weights.

After calculating the personality traits of a person’s website visit records, we can use Equation (1) to get the average score of personality model for each class of website. While the \( n_{\text{access}_k} \) is the times that the user access the kth website and \( \sum n_{\text{access}_k} \) is total user access times for certain class websites. \( i \) means the ith dimension of five factor personality model, \( j \) means the jth class of websites. Thus, \( p_{\text{sub}_score}_{i,j} \) is the score of the ith dimension on jth class website,

\[
p_{\text{sub}_score}_{i,j} = \frac{\sum (n_{\text{access}_k} \times p_{\text{sub}_score}_{i,k})}{\sum n_{\text{access}_k}}
\]

(1)

Equation (2) can be used for last result of personality traits by weighted average of \( p_{\text{sub}_score}_{i,j} \). While \( p_{\text{score}_i} \) is the score of the ith dimension of five factor personality model, \( p_{\text{weight}_{i,j}} \) is the weight of jth website on ith dimension and \( p_{\text{sub}_score}_{i,j} \) means the score of ith dimension which is calculated from the jth class of website.

\[
p_{\text{score}_i} = \sum_{j=1}^{m} p_{\text{weight}_{i,j}} \times p_{\text{sub}_score}_{i,j}
\]

(2)

4. Experiment

We collect 122 volunteers’ network access log and their self reporting scale. The time span of network access log is 30 days. Those data are used to calculate their personality traits.

In order to compare the credibility of personality portraits obtained in different ways. We use root mean square error (RMSE) as the evaluation index. The calculation of RMSE refers to Equation (3). While \( n \) is the number of person to calculate their personality trait, \( p_{\text{score}_i} \) is the score of ith factor of his personality trait, \( p_{\text{score}_ib} \) is the benchmark of ith factor.

\[
RMSE = \sqrt{\frac{1}{n} \sum (p_{\text{score}_i} - p_{\text{score}_ib})^2}
\]

(3)

4.1. Personality traits of a volunteer

Table 4 shows the personality traits of one volunteer on different website (only some typical website, websites with few visits are not listed). Obviously, the personality traits given by different websites are very different.

| Data Source | Access Times | Extr. RMSE | Agr. RMSE | Cons. RMSE | Neu. RMSE | Open. RMSE |
|-------------|--------------|------------|-----------|------------|-----------|------------|
| Weibo       | 1344         | .307       | .792      | .635       | .753      | .164       |
| Weixin      | 1100         | .812       | .634      | .712       | .106      | .121       |
| Zhihu       | 822          | .217       | .189      | .335       | .386      | .463       |
| Sina        | 230          | .205       | .273      | .342       | .176      | .277       |
| QQ Music    | 72           | .507       | .297      | .205       | .586      | .161       |
| GamesZone   | 22           | .324       | .258      | .427       | .587      | .232       |
| Taobao      | 44           | .433       | .272      | .513       | .678      | .263       |
| Jingdong    | 23           | .704       | .699      | .713       | .825      | .621       |
Table 5 shows the final result of the personality trait of the chosen volunteer which is calculated by Equation (1) and Equation (2). According to Table 5, we can see that by synthesizing the data of a visitor on various websites, we can correct the deviation caused by the website atmosphere and make the comprehensive personality trait closer to the actual situation.

| Website type | Extr. RMSE | Agr. RMSE | Cons. RMSE | Neu. RMSE | Open. RMSE |
|--------------|------------|-----------|------------|-----------|------------|
| Expressive   | .454       | .587      | .585       | .443      | .224       |
| Reading      | .201       | .253      | .332       | .152      | .269       |
| Mixed        | .464       | .288      | .257       | .586      | .178       |
| Functional   | .526       | .419      | .582       | .728      | .386       |
| Final result | .359       | .433      | .446       | .356      | .208       |

4.2. Personality traits of a volunteer

Weibo is the most visited website in this experiment. Its RMSE value of personality trait is also the smallest in all separate website.

Figure 2 shows the comparison of the statistical data of personality traits based on multi-source and Weibo. Shown as the figure, the maximum RMSE value or the average RMSE value on each factors of personality traits based on multiple sources is smaller than that based on Weibo. And the minimal RMSE value of multiple sources is almost same with Weibo.

Therefore, it can be considered that the personality trait based on multi-source is a more accurate and effective personality calculation method.

5. Conclusion And Future Work

In this paper, the personality computing of multi-source is studied. Through the collection of College volunteers’ network access logs, we can get more comprehensive network activity footprints of volunteers, not limited to social media activities. By analyzing the emotional tendency of web pages, we can get the personality characteristics of volunteers in different types of websites. The experiment shows that the personality traits of volunteers are different on different social media and different websites. The same social media has the same orientation to different people's personality traits. The results of this study show that the personality traits of individuals in a certain environment are affected by the environment.

The data of this paper comes from web pages. The web page is only a small part of the network footprint, more content such as mobile phones access logs are not collected. In addition, this study only analyzes the text content of the web page, and does not consider the image, video and other types of content. Therefore, more network behaviors and accompanying personality traits can not be extracted and analyzed. This is what we will do in the future.
References

[1] Stanek, S. , & Sabat, A. . (2021). The application of it tools in assessing employees' personality and motivation. European Research Studies Journal. 24(Special 1):689-707.

[2] Fiske, S. T. , Cuddy, A. , Glick, P. , & Xu, J. . (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology, 82(6):878-902.

[3] Bargh, J. A. , Gollwitzer, P. M. , Lee-Chai, A. , Barndollar, K. , & Tr?Tschel, R. . (2019). The automated will: nonconscious activation and pursuit of behavioral goals. Journal of personality and social psychology,81(6):1014-1027.

[4] Heeringen, C. V. , Audenaert, K. .., Laere, K. V. .., Dumont, F. .., Slegers, G. .., & Mertens, J. .., et al. (2019). Prefrontal 5-htr2a receptor binding index, hopelessness and personality characteristics in attempted suicide. Journal of affective disorders,74(2):149-158.

[5] Id, A. . (2019). Computational personality prediction based on digital footprint of a social media user - sciencedirect. Procedia Computer Science, 156:185-193.

[6] Marouf, A. A. , Hasan, M. K. , & Mahmud, H. . (2020). Comparative analysis of feature selection algorithms for computational personality prediction from social media. IEEE Transactions on Computational Social Systems, PP(99), 1-13.

[7] Nguyen, T. , Phung, D. , Hoai, M. , & Nguyen, T. H. . (2020). Structural and Functional Decomposition for Personality Image Captioning in a Communication Game. arXiv preprint arXiv:2011.08543.

[8] Ning, H. , Dhelim, S. , & Aung, N. . (2019). Personet: friend recommendation system based on big-five personality traits and hybrid filtering. IEEE Transactions on Computational Social Systems, 394-402.

[9] Padmanabhan, B. . (2020). Computational personality recognition and sentiment analysis of select novels of Cormac Mccarthy. IUP Journal of English Studies. 15(3): 91-102.

[10] Matz, S. C. , Segalin, C. , Stillwell, D. , SR Muller, & Bos, M. W. . (2019). Predicting the personal appeal of marketing images using computational methods. Journal of Consumer Psychology. 29(3): 370-390.

[11] Farnadi, G. , Sitaraman, G. , Sushmita, S. , F Celli, Kosinski, M. , & Stillwell, D. , et al. (2016). Computational personality recognition in social media. User modeling and user-adapted interaction, 26(2-3):109-142.

[12] Segalin, C. , Dong, S. C. , & Cristani, M. . (2017). Social profiling through image understanding: personality inference using convolutional neural networks. Computer Vision and Image Understanding, 156:34-50.

[13] Kandasamy, I. , & Smarandache, F. . (2017). Triple Refined Indeterminate Neutrosophic Sets for personality classification. In: 2016 IEEE Symposium Series on Computational Intelligence. Athens. pp. 1-8.

[14] Park, G. , Schwartz, H. A. , Eichstaedt, J. C. , Kern, M. L. , & Seligman, M. . (2014). Automatic personality assessment through social media language. Journal of Personality and Social Psychology, 108(6), 934-952.