Correlation analysis and prediction of personality traits using graphic data collections

Vitaly Fralenko[1]*, *Vyacheslav* Khachumov[2], *Mikhail* Khachumov[3]

[1]Ailamazyan Program Systems Institute of Russian Academy of Sciences, 152020, Pereslavl-Zalessky, Russia
[2] RUDN University, 117198, Moscow, Russia
[3]Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, 119333, Moscow, Russia

Abstract. The questions of building mechanisms for identifying patterns and building modern tools for analyzing data from social networks are considered. It is proposed to apply modern methods of web pages' automatic analysis, testing hypotheses about the presence of correlation links, automatic classification of graphic information using the apparatus of artificial neural networks. The presence of correlation between personality traits of the "Big Five" is investigated. Strong fluctuations in the values of personality traits were revealed depending on various types for groups of people. The problem of predicting the personality traits of the Internet user by the images posted by him is investigated, artificial neural networks are used as a tool. Two series of experiments were carried out, in the first series, a convolutional neural network, trained on the images and results of the NEO-FFI questionnaire, was used to predict personality traits. The sequential use of convolution and subsampling in the convolution network leads to the so-called increase in the level of features: if the first layer extracts local features from the image, then subsequent layers extract common features that are called high-order features. In the second series of experiments, this type of artificial neural network was used to extract high-level features, which were then used to train a direct distribution network that performs forecasting. Thus, the more layers are used, the more features associated with personality traits are extracted from the images. For processing arrays of graphic information, the "Microsoft Cognitive Toolkit library" and the Nvidia Geforce GTX 1080 Ti graphics accelerator were used. The results of the experiments revealed those personality traits that are most correlated with the images posted by Internet users.

1 Introduction

The main idea of his work is to combine the advanced knowledge of psychologists about the personality traits of a person with modern information technologies based on tools and

* Corresponding author: alarmod@pereslavl.ru

artificial intelligence technologies.

The so-called “Big Five” (Tupes and Christal 1961) includes the following personality traits: “Openness” (openness to experience, intelligence), “Consciousness” (consciousness, self-awareness, conscientiousness), “Extraversion” (extraversion: energetic behavior, inclination to contacts), “Agreeableness” (kindness, pleasantness, ability to come to an agreement), “Neuroticism” (neuroticism, emotional instability, anxiety, low self-esteem).

The classic approach to the classification of traits is a psychological survey. However, with the development of information technology, collection of a wide variety of data has become available: information about the users’ natural behavior of social networks. A study of the network users’ behavior is carried out in the context of determining their personality traits and psychological state.

Recently, they often analyze images posted by authors on the Internet, including both portraits and sets of general-purpose photographs (Cristani et al. 2013; Liu et al. 2016). Research on graphic online content covers two broad areas: the images that people post and the images posted by other people that they like (click “like”).

The tools proposed in the work are used for automatic extraction and high-performance data processing aimed at solving two important scientific problems: checking the hypothesis about the presence of interpersonal correlation and predicting personality traits throughout the amount of information available.

2 Analysis of personality traits’ correlations

The following hypothesis was put forward: there is a significant (“noticeable”) correlation between personality traits from “Big Five” that determines their relationship. Testing the hypothesis required a study of the statistical validity of the correlation analysis results; generalization of results from various researchers obtained in various population groups (age, occupation, health status, success rate, etc.). To assess the strength of the correlation, the Robert Chaddock system (Chaddock 1925) was used.

Summarize the results of the scientific literature analysis of the proposed topic. From the research (Cristani et al. 2013), it follows that the “Agreeableness” trait has a “moderate” positive correlation with “Conscientiousness” (0.45647) and a “high” negative correlation with “Neuroticism” (-0.76511) according to collection of photographs. However, due to the low coefficient of experts consistency by Krippendorff (Krippendorff 2011), the obtained patterns cannot be considered reliable.

In research (Arshava and Amineva 2011), which investigated a group of patients with diabetes mellitus, it follows that the sign “Agreeableness” “noticeably” positively correlates with “Conscientiousness” (0.511), which is consistent with previous results. At the same time, a “noticeable” positive correlation (0.518) occurs between the “Extraversion” and “Neuroticism” traits, which is not consistent with the negative correlation between them obtained in the previous study (-0.25016). A study (Biesanz and West 2004) (peer groups and parents) showed a “weak” correlation while maintaining the trend of “Extraversion” and “Neuroticism” (0.18). In the work (Liu et al. 2016) (image analysis and test polls) it can be seen that the results of the studies using two different approaches are generally consistent. Neuroticism is “moderately” negatively correlated with “Conscientiousness”, “Extraversion” and “Agreableness” in both cases. "Conscientiousness" and "Extraversion" have a "weak" positive correlation. When extracting personality traits using images, “Conscientiousness” and “Agreableness” have a “weak” correlation, however, when using test polls, we see that the correlation is “moderate”. A "weak" correlation between "Openness" and "Extraversion" is observed only in the case of test polls. In the study (Oz 2015) (future teachers of the English language), there is a “moderate” correlation of the “Openness” and “Consciousness” traits (0.403), a noticeable correlation of the “Openness”

and "Agreeableness" traits (0.507). The results presented in (Klimstra et al. 2013) show that social roles of adults can influence the correlation of personality traits for adolescents.

A generalization of the obtained data is presented in Table 1, which contains the averaged correlation values from the considered works.

Table 1. Averaged correlation between personality traits

	O	C	E	A	N
O	1	0.2198	0.1810	0.1540	0.1201
C	0.2198	1	0.1036	**0.3212**	-0.0737
E	0.1810	0.1036	1	0.2125	-0.0189
A	0.1540	**0.3212**	0.2125	1	**-0.1730**
N	0.1201	-0.0737	-0.0189	**-0.1730**	1

It should be noted that the presence of a high correlation between the features obtained in a number of works cannot be accepted and interpreted without analyzing the degree of expert psychologists' coordination, who conducted the research. The integration of all data into a single table naturally reduces the correlation relationships, however, some trends remain. The most strongly correlated are "Consciousness" and "Agreeableness", which corresponds to the results of most works we have analyzed. Separately, it is worth noting the presence of a "weak" negative correlation between "Agreeableness" and "Neuroticism". In general, "noticeable" and "high" correlations are observed only for particular population groups. Thus, the hypothesis put forward is confirmed only in special cases.

3 The use of artificial neural networks for personality traits' prediction

The purpose of the study is to train neural networks to predict the personality traits of Internet users, using images from the "profiles" of these users in "VKontakte" social network. The created software is based on the Microsoft Cognitive Toolkit library (Meints 2019). During experiments, information from the "profiles" of 84 Internet users was used. Half of the data from users was taken as a training sample, and half as a test sample. According to the "NEO-FFI" questionnaire (Costa and McCrae 1992), directories with user images were assigned an information vector from the estimated values of five personality traits, normalized from 0 to 48. The database is described by the following characteristics: number of images in the training set: 4322; the number of images in the test sample: 4196; average number of images per user in the training set: 103; average number of images per user in a test sample: 100; minimum number of images per user: 5; maximum number of images per user: 170.

Main features and results of an experiment with one convolutional neural network

The first experiment is investigated the possibility of using a convolutional neural network (CNN) as the main tool for assessing the relationship of images posted by users with their personality traits. In the second experiment, the possibility of using pre-trained CNN to extract high-level features was tested; the hidden layer output preceding the fully connected one was used as the source of these data. Convolutional networks for extracting high-level features were trained on the basis of the "ImageNet 1k" annotated image database (Russakovsky et al. 2015).

The main characteristics of the first experiment:

- size of generated input images: 128x128, 256x256 and 512x512;
- the number of convolution layers: 1 5;
- the number of feature cards in the convolution layers: 3 90;
- convolution window size: 3x3;

- sub-sampling step: 2;
- the size of the sub-sampling window that selects the maximum (max-pooling): 3x3;
- the number of images used in one training iteration: 82 412.

The best results were shown by CNN with three layers with 9, 6 and 3 cards of signs in them; size of generated input images: 512x512 pixels; dropout layer with a probability of 0.1. The optimal size of the group of processed images (in one iteration of training): 100 images. A larger number of images worsened the accuracy of features prediction in the test sample, while a smaller number required the use of a significantly lower coefficient of learning speed and an increase in the number of learning eras. Improvements in the characteristics of the neural network in this problem formulation are no longer observed after 17 thousand eras of training. Two modes were investigated: with forecasting on one line and all five features together. As evaluative features, 1) the standard deviation (S.D.) was used, 2) the accuracy of identifying the minimum and maximum expressed personality traits (from 0 to 1).

The final results of the best CNN are presented in Tables 2-3.

Table 2. Separate traits' forecasting

	O	C	E	A	N	The average value of S.D. for individual traits
Training set selection, S.D. according to traits of "Big Five"	5.54	9.23	8.03	5.62	10.43	7.77
Testing set selection, S.D. according to traits of "Big Five"	5.98	9.17	8.40	6.69	10.45	8.14

Table 3. Joint forecasting of five personality traits, test set selection

	O	C	E	A	N
The average value of S.D. for individual traits «Big Five»:	6.17	8.93	7.72	6.44	10.45
The average value of S.D.:	7.94	The average value of S.D. for individual users:		7.72	
The accuracy of the selection for the most pronounced personality traits:	0.21	The accuracy of the selection for the least pronounced personality traits:		0.45	

Summarizing the data from Tables 2-3 and the experimental results, one can indicate the following patterns: the least pronounced personality trait (with the lowest "NEO FFI" coefficient) of the Internet user is consistently predicted better than the most pronounced, this feature manifested itself in all experiments without exception; configurations of neural networks with a small number of convolution layers (from one to two) do not provide sufficient accuracy for identifying personality traits; increasing the size of the input images sequentially increases the selection accuracy of relevant features and reduces S.D..

The main features and results of the experiment with two types of neural networks

The main characteristics of the second experiment:

- architecture of pre-trained CNN: "ResNet", "InceptionV3", "AlexNet" and "VGG";
- the number of features in the hidden CNN layer: 512 4096;
- the number of layers in the direct distribution network: 1 3;

- the number of neurons in the layers of the direct distribution network: 5 20000;
- the number of images used in one training iteration: from 82 to the entire training sample (4322 images).

The best result was obtained using attribute from the ResNet50 neural network with 2048 informative output signals of the hidden layer, and the entire training sample was used at each iteration. The time taken to complete 250 thousand training eras turned out to be an order of magnitude shorter than the training time of the neural network in the previous experiment. Direct distribution networks with one layer of neurons preceding the output layer of the neural network showed the best result; 50 neurons were enough for confident training. However, the overall quality of the resulting predictive tool has deteriorated, in particular, the accuracy of highlighting the most and least pronounced personality traits (see Tables 4-5). Nevertheless, the standard deviations are very close to those obtained in the first experiment.

Table 4. Separate forecasting of personality traits

	O	C	E	A	N	The average value of S.D. for individual traits
Training set selection. S.D. according to traits of "Big Five"	4.73	7.40	6.12	4.62	8.09	6.19
Testing set selection. S.D. according to traits of "Big Five"	5.62	8.84	8.06	6.30	10.35	7.83

Table 5. Joint forecasting of five personality traits, testing set selection

	O	C	E	A	N
The average value of S.D. for individual traits «Big Five»:	5.64	8.84	7.97	6.42	10.35
The average value of S.D.:	7.84	The average value of S.D. for individual users:		7.61	
The accuracy of the selection for the most pronounced personality traits:	0.19	The accuracy of the selection for the least pronounced personality traits:		0.33	

An experiment with a direct distribution neural network showed that an increase in the number of layers of a direct distribution network worsens the quality of the forecast; neural network begins "retraining".

4 Conclusion

The analysis showed the presence of both “weak” and “noticeable” correlation in the results of experimental studies, with a generally accepted level of statistical reliability. It is not currently possible to establish unambiguously the fact of correlation dependencies between personality traits based on the study’s results conducted by the authors and the available data from the reviewed publications due to the lack of information on the expert opinions’ consistency. Strong fluctuations in the values of personality traits are observed depending on various types of groups of people. A generalization of the results shows a strong positive correlation between “Consciousness” and “Agreeableness”.

In the experiments with predicting personality traits from images using a convolutional neural network, the best prediction indicators were obtained for the "Openness" and "Agreeableness" traits; the worst are for neuroticism; more accurate results are obtained by jointly predicting five personality traits at once. The use of two networks (convolutional and direct distribution) made it possible to significantly reduce the training time, however, the accuracy of selecting the most and least pronounced features decreased.

In the future, it is planned to create a unified information environment and software system, concentrating the expert knowledge of psychologists and specialists, methods of intellectual analysis. The system should provide simple and convenient means of navigation in the data stream due to an intelligent interface, including monitoring of personal characteristics.

The reported study was funded by RFBR, project number 18-29-22003.

References

1. I.F. Arshava & Ya.R. Amineva, *The study of correlations between the psychological characteristics of the personality of patients with type 2 diabetes*, Proceedings of V International Scientific and Practical Conference "Actual Problems of Personality Psychology", (2011) https://sibac.info/sites/default/files/files/2011_01_30_Psihologiya/30.01.2011.docx. Accessed 4 July 2019. (In Russian).
2. J.C. Biesanz, & S.G. West, *Towards understanding assessments of the Big Five: Multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer*, Journal of Personality, **72(4)**, 845–876, (2004). doi: 10.1111/j.0022-3506.2004.00282.x
3. R.E. Chaddock, *Principles and methods of statistics* (Houghton Mifflin Co.; 1st edition, 1925). doi: 10.1177/000271622612300150
4. P.T. Costa, & R.R. McCrae, *Normal personality assessment in clinical practice: The NEO Personality Inventory*, Psychological assessment, **4(1)**, 5–13 (1992). doi: 10.1037/1040-3590.4.1.5
5. M. Cristani, A. Vinciarelli, C. Segalin, & A. *Perina, Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis*, Proceedings of the 21st ACM International Conference on Multimedia, 213–222 (2013). doi: 10.1145/2502081.2502280
6. T.A. Klimstra, W. Bleidorn, J.B. Asendorpf, van Aken M.A.G., & Denissen J.J.A., *Correlated change of Big Five personality traits across the lifespan: a search for determinants*, Journal of Research in Personality, **47(6)**, 768–777 (2013). doi: 10.1016/j.jrp.2013.08.004
7. K. Krippendorff, *Computing Krippendorff's alpha-reliability*, *S*cholarlyCommons, (2011). Information on https://repository.upenn.edu/asc_papers/43. Accessed 4 July 2019
8. L. Liu, D. Preotiuc-Pietro, Z.R. Samani, M.E. Moghaddam, & L. Ungar, *Analyzing personality through social media profile picture choice*, Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016), 211–220 (2016)
9. W. Meints, *Deep Learning with Microsoft Cognitive Toolkit Quick Start Guide: A practical guide to building neural networks using Microsoft's open source deep learning framework*. (Packt Publishing Limited, Birmingham, 2019)
10. H. Oz, *Personality traits and ideal l2 self as predictors of academic achievement among prospective English teachers*, Proceedings of the 8th Annual International

Conference of Education, Research and Innovation (ICERI-2015), Seville, Spain, 5833–5841 (2015)
11. O. Russakovsky, J. Deng, H. Su, et al., *ImageNet large scale visual recognition challenge*, International Journal of Computer Vision, **115(3)**, 211–252 (2015). doi: 10.1007/s11263-015-0816-y
12. E.C. Tupes, & R.E. Christal, *Recurrent personality factors based on trait ratings*, USAF ASD technical report, 61–97 (1961).