A Construction of Knowledge Base for Personality Estimation based on Submitted Text Data in Twitter or Blogs

Noriyuki Okumura[1] and Manabu Okumura[2]
[1]*Department of Information Engineering, National Institute of Technology, Kagawa College, 551 Koda, Takuma-cho, Mitoyo Kagawa, Japan*
[2]*Precision and Intelligence Laboratory, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa, Japan*

Keywords: Personality Estimation, Twitter, Blog, Emotion Judgment.

Abstract: The personality that is estimated based on documents of blogs or tweets in Twitter can not agree in the sender's real personality. It is important that we recognize the difference between these estimated and real personalities. This paper constructs a knowledge-base for extracting the sender's virtual personality in customer-generated media. We focus on sender's emotions that are included in sender's posts for automatic personality estimation. We examined the correlation between the ratio of each emotion term (anger, sadness, fear, disappointment, regret, guilt, shame, pleasure, and ease) in all sentences of each participant and the values of NEO-FFI (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience) based on the experiment that human subjects who stayed in each sender's character answered to NEO-FFI.As an evaluation result, we find out that the sender's virtual personality is potentially-correlated with emotions in sender's posts.

1 INTRODUCTION

We use customer-generated media (CGM) such as blog, Twitter, Facebook, and so on to express our opinions or to record our experiences on a daily basis. We deepen exchange not only with acquaintances but also strangers through CGM. These computer-mediated communications (CMC) facilitate communication with users. On the other hands, CMCs encompass the trouble that is called flaming[1] because users of blogs are less than familiar with each other or readers of blogs misunderstand senders' posts. Our research aims to prevent this problem from occurring.

Senders' posts construct their virtual personality in CGM. The Virtual Personality of senders that are estimated based on senders' posts in CGM cause the flaming phenomenon because virtual personality is not always matching senders' real personality. We also wonder the problem that is called blogroach[2] on a daily basis. However, we focus on flaming as distinct from blogroach. Because the person who causes blogroach has his/her sights on many and unspecified persons. He/She also causes the flaming phenomenon, but senders of blogs, twitter, and so on can prevent this phenomenon if they mind their posts before they submit their posts.

What is the best will to avoid troubles in CGM? In 1995, Sally Hambridge propounded "Netiquette Guideline" that defined the etiquettes for E-mail, Mailing Lists, News services, and so on(Hambridge, 1995). We commonly study netiquettes in our first year in college in Japan. We users usually read blogs or Twitter with estimating the sender's personality based on his/her posts. They encompass troubles in CGM because they do not know how to behave suitably on the internet. Their behavior in CGM builds their virtual personality at the unconscious level.

This paper investigates the tendency between sender's real personality and sender's virtual personality estimated by readers of blogs. We users usually read blogs or Twitter with estimating the sender's personality based on his/her posts. If we do not recognize the difference between our real personality and our virtual personality, we cause the trouble such as flaming. In our research, we use the Big Five personality traits for extracting sender's real personality and sender's virtual personality. Therefore, they have to

[1]Flaming means flooded by comments. For example, we use this term such as "My blog is under flame."

[2]Blogroach is a coined term (Blog + cockroach). Blogroach blows into comment columns in unspecified blogs to encroach on the blogs.

Okumura, N. and Okumura, M..
A Construction of Knowledge Base for Personality Estimation based on Submitted Text Data in Twitter or Blogs.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 2: KEOD*, pages 418-423
ISBN: 978-989-758-158-8

know how users on the internet see their posts.

We obtained senders' real personality in a survey using Macromill[3]. Examinees of this questionnaire give a reply to 60 questions of NEO-FFI that is one of the Big Five personality traits and tell us their ID of Twitter and blogs. We obtained 205 valid responses in this survey. We construct a small-scale knowledge-base using 23 samples out of 205 valid samples.

2 RELATED WORK

The personality tests are mainly divided into three genres that are a questionnaire, a projection, and a performance test. Our study estimates the senders' personality using a questionnaire method. Questionnaire methods consist of Big Five personality traits(Costa and MacCrae, 1992), Yatabe-Guilford(Y-G) Personality Inventory, Egogram and so on. We are especially interested in Big Five(NEO-FFI) in this paper. NEO-FFI(NEO-Five Factor Inventory) is a personality test that provide five inventories; Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. This paper calculates these five factors using 60 questions.

Yarkoni(Yarkoni, 2010) proposed a personality estimation method based on tweets using Big Five. He argued that the method can estimate sender's real personality based on tweets. He state that the method can roughly estimate sender's personality based on about 50 tweets. In addition, the method can make detailed analysis based on 200 tweets. Researchers advance the study of personality estimation based on CGM, however, they do not focus on the virtual personality on CGM. Jin(Jin, 2013) reported about virtual identity in Twitter as similar as virtual personality, however, her viewpoint does not equal to point of our observation.

A research team in Pennsylvania exhibit Five-Labs[4] as an application of personality estimation using Big Five. Figure 1 shows the author's personality estimation using Five-Labs.

Five-Labs estimates the author's personality that he does not seem out going and he is an emotional disturbance man. Figure 2 shows comparison with the author's real personality. Figure 2 also shows that author's real personality does not equal to the result of personality estimation using Five-Labs.

Sumner et al. (Chris Sumner and Park, 2012) are interested in Dark Triad traits. The Dark Triad is a group of three personality traits: narcissism,

[3]http://monitor.macromill.com

[4]http://labs.five.com, Five-Labs is not available now because APIs of Facebook do not work correctly.

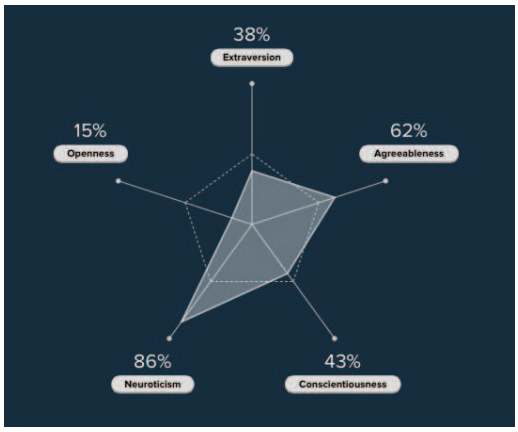

Figure 1: Personality estimation result using Five-Labs.

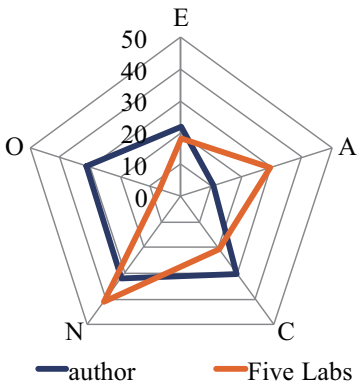

Figure 2: Difference between author's real personality and personality estimation result using Five-Labs.

Machiavellianism and psychopathy. These traits have the potential to cause antisocial personality disorder. These are the dark side of our personality. Therefore, we should consider Sumner's claim to estimate virtual personality.

Qiu et al. (Qiu et al., 2012) analyzed tweets using Big Five. They focused on the personality "expression" and "perception". Especially, they examined differences in gender, age, and ethnicity. Our experiment (Section 3.2) shows the difference between male and female as pointed out in a paper by Qiu et al.

The psychologists are interested in personality estimation in psychology research. Kraut et al. proposed the Rich get richer model(Kraut et al., 2002). This model argues that aggressive persons get more aggressive and the persons who is passive become more passive through communication on the internet. This argue associates with our purpose that we extract the difference between sender's real personality and sender's virtual personality on the internet.

Murao(Murao, 2014) estimates the "City Personality" based on tweets that persons who lives in a certain area posted. The evaluation makes clear that

City Personalities are different in each area. He evaluates six areas: New York, Los Angels, Chicago, Salt Lake City, London and Oxford. City Personality of each area has different trend. This is an important result for estimating virtual personality in CGM because senders background information directly affect his/her posts.

3 EXPERIMENTS

This paper shows Examinees read all tweets and all posts in each blog. Examinees read all tweets and all posts in each blog that does not include comments on the blog. They answer NEO-FFI questionnaire by putting themselves in the mind of Twitter or blog users.

We also investigate the correlation between sender's virtual personality and emotions that are obtained from sender's posts using Emotion Judgment System(Seiji Tsuchiya, 2009). We extract emotions in sender's posts based on the knowledge base that is used in Emotion Judgment System.

3.1 NEO-FFI (NEO Five Factor Inventory)

We have a questionnaire using Macromill to obtain sender's real personality who use Twitter and some kind of blogs. This questionnaire targets at users who have both Twitter's and Blog's ID. The examinees answers these IDs and 60 questionnaires of NEO-FFI. The number of examinees is 483, however, there are some advertising ID or Twitter's "Bot" ID. We remove these incorrect IDs and obtain 203 valid responses.

Examinees answered Five Factor Inventory using NEO-FFI. Five Factor Inventory consist of following factors. Each factor is represented by values(0 to 48).

- E: Extraversion
- A: Agreeableness
- C: Conscientiousness
- N: Neuroticism
- O: Openness to Experience

In this paper, we investigate 23 examinees' results out of 205 examinees as a sample because it takes human subjects a while to read all of tweets and posts in examinee's blog.

3.2 Personality Estimation based on Posts of Twitter or Blogs

It is difficult that senders know how their posts influence readers in CGM. Readers cannot understand their behavior correctly as if senders intentionally behave as a certain character.

This paper quantifies senders' virtual personality based on tweets and posts in blogs by five human subjects (4 males and 1 female) using NEO-FFI. We extract the difference between senders' real personality and senders' virtual personality in CGM. Human subjects read all documents in 23 blogs until December 31, 2014 and all tweets from January 1, 2011 to December 31, 2014. We set no limitation for human subjects to answer NEO-FFI questionnaires.

3.3 Correlation between Estimated Values of NEO-FFI and Emotions Judgment

This paper aims to estimate the virtual personality automatically. For this reason, it is necessary to construct automatic estimation method of the virtual personality. Personality partly consist of their emotions. We constructed Emotion Judgment system based on a sentence using Concept-base and knowledge bases. We have to validate the correlation between each value of NEO-FFI and the ratio of emotions of each sentence in their tweets or posts.

Existing Emotion Judgment System refer to their own knowledge base for emotions judgment. The system can analyze well-formed sentences, however, the system cannot correctly judge chatty sentences such as tweets. We evaluate the system as exploratory experiment, the system can answer only 10% of sentences in all documents. Therefore, this paper uses a ratio of the total number of each emotion in each sentence and the total number of sentences in all documents.

4 RESULTS

In this section, we show the results of NEO-FFI estimation and correlation between the results of emotions judgment and NEO-FFI estimation.

4.1 NEO-FFI

Figure 3 and figure 4 show the sample of virtual personality estimation based on tweets and posts in blog. Dash lines show the estimation results of each human

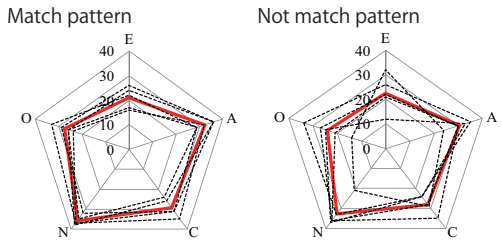

Figure 3: the result of virtual personality estimation by human subjects based on Twitter.

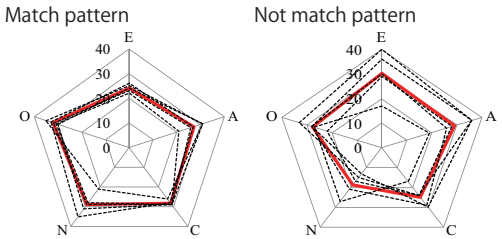

Figure 4: the result of virtual personality estimation by human subjects based on blog.

subjects. The red line shows the average of each estimation result. Match pattern means that all human subjects estimate similar personality. Not match pattern means that each estimation result is divided.

In the result of Twitter estimation, 16 persons out of 23 examinees are Match pattern. On the other hand, 12 persons out of 23 examinees are Match pattern in the experiment of blogs.

Figure 5 shows the comparison with sender's real personality and estimated virtual personality. Black line shows sender's real personality. Blue dash line shows the average of each subjects estimation based on blogs. Red dash line shows the average of each subjects estimation based on Twitter. 21 persons out of 23 examinees have same tendency like Figure 5.

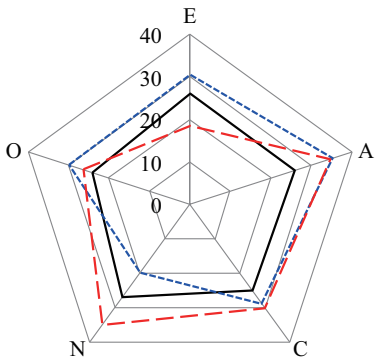

Figure 5: comparison with sender's real personality and estimated personality(Twitter, blog).

4.2 Correlation Analysis

Table 1 and table 2 show the result of correlation analysis between emotions judgment and each Five Factor. The bold character shows high correlation values in the table.

Table 1: Correlation between the output of Emotion Judgment System about Twitter and each factor of Five Factor Inventory.

	E	A	C	N	O
Anger	-0.03	-0.17	**0.62**	0.15	-0.13
Sadness	0.08	-0.10	0.22	-0.06	-0.21
Fear	-0.13	-0.17	**0.51**	0.08	-0.04
Disappointment	-0.15	**0.51**	**0.67**	**0.65**	0.20
Regret	0.14	0.29	0.49	0.25	0.32
Guilt	0.04	-0.33	0.31	-0.12	-0.23
Shame	0.16	-0.17	0.33	0.04	-0.06
Pleasure	0.24	0.27	0.35	0.20	0.06
Ease	0.10	0.26	**0.79**	0.43	-0.06

Table 2: Correlation between the output of Emotion Judgment System about Blogs and each factor of Five Factor Inventory.

	E	A	C	N	O
Anger	-0.18	-0.29	-0.25	0.10	-0.36
Sadness	0.01	0.12	0.16	0.03	-0.16
Fear	-0.27	-0.30	-0.22	0.17	-0.31
Disappointment	0.16	0.14	0.03	-0.20	-0.41
Regret	-0.03	0.08	-0.18	0.13	-0.08
Guilt	-0.34	-0.31	-0.20	0.29	-0.14
Shame	0.14	0.19	-0.02	-0.11	-0.21
Pleasure	0.10	0.21	-0.02	-0.04	-0.18
Ease	-0.07	-0.02	-0.15	0.18	-0.24

4.3 Knowledge Base

We constructed a knowledge-base for estimating sender's real personality based on sender's virtual personality. The knowledge-base has three fields: sender's virtual personality scores estimated from blogs, sender's virtual personality scores estimated from tweets, and sender's real personality scores using NEO-FFI (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience). Table 3 shows constructed knowledge-base. Table 1, table 2 are also defined as knowledge-base as positive sample for machine learning method.

5 DISCUSSION

In the experiment of human subjects, about 70% of examinees have same tendency based on Twitter.

Table 3: Constructed knowledge-base for virtual and real personality estimation.

Estimated(Blog)	E	A	C	N	O
Examinee A	34	37	29	19	30
Examinee B	19	19	21	31	31
Examinee C	31	34	27	25	31
...	...	...	...	...	...
Estimated(Twitter)	E	A	C	N	O
Examinee A	22	27	29	30	24
Examinee B	22	29	32	32	27
Examinee C	30	28	30	30	20
...	...	...	...	...	...
Real Personality	E	A	C	N	O
Examinee A	25	21	24	28	33
Examinee B	27	21	34	28	30
Examinee C	20	24	35	31	23
...	...	...	...	...	...

However, in the case of blogs, about 50% of examinees have same tendency. The reason why each human subject answers different personality is mainly influenced by woman subject. Men generally read documents analytically, however, women generally read them by empathetic sight. This result argues the necessity to divide men's model and women's model for virtual personality estimation.

In the case of blog, sentences increase in number. In addition, senders arrange their verbs and objects before they submit their documents to the blog. This tendency is caused by the education of Netiquette that defined the behavior on the internet.

For example of author's personality estimation in Figure 2, the factor of Openness to Experience in estimated personality is lower than author's real personality. This is because we should not disclose our individual information without no discretion based on the education of Netiquette. This is the self-defense on the internet.

Thus, the behavior in CGM depends on the literacy education partly. Therefore, it is important for us to comprehend our virtual personality in CGM to prevent troubles such as flaming.

In the experiments of the analysis of correlation between emotions judgment and each Five Factor, Twitter's result(Table 1) has higher correlation values than blog's result(Table 2). We use Twitter with a light heart as compared to blogs because all tweets have 140 characters limitation. Tweets are generally submitted without wordsmith. Therefore, tweets have more emotional keyword than blog's documents.

In this paper, we organize all 23 examinees result as knowledge base for virtual personality estimation. However, only about 10% of valid responses are analyzed in this paper. We must evaluate the remainders.

6 CONCLUSIONS

In this paper, we investigated the tendency of virtual personality in CGM using NEO-FFI. The experiments of human subjects revealed that virtual personality was not match with examinee's real personality basically.

For automatically estimation of virtual personality, we compared with the values of each Five Factor to emotions judgment. Twitter's result showed higher correlation with emotions judgment.

This paper investigated small-scale estimation, we will analyze all of examinees. We also construct a method of estimating virtual and real personality using machine learning method (i.e. regression, support vector machines and clustering) based on the constructed knowledge-base as a future work.

ACKNOWLEDGEMENTS

This work was supported by KAKENHI 15K21592.

REFERENCES

Chris Sumner, Alison Byers, R. B. and Park, G. J. (2012). Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. *Proceedings of the 2012 11th International Conference on Machine Learning and Applications - Volume 02*, pages 386–393.

Costa, P. T. and MacCrae, R. R. (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual*. Psychological Assessment Resources.

Hambridge, S. (1995). Netiquette guidelines. *IETF RUN Network Working Group, RFC 1855*.

Jin, S.-A. A. (2013). Peeling back the multiple layers of twitters private disclosure onion: The roles of virtual identity discrepancy and personality traits in communication privacy management on twitter. *New Media & Society*, 15(6):813–833.

Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V., and Crawford, A. (2002). Internet paradox revisited. *Journal of Social Issues*, 58(1):49–74.

Murao, H. (2014). Personality estimation from {SNS} messages and its application to evaluating a city personality. *Procedia Technology*, 18:72 – 79. International workshop on Innovations in Information and Communication Science and Technology, {IICST} 2014, 3-5 September 2014, Warsaw, Poland.

Qiu, L., Lin, H., Ramsay, J., and Yang, F. (2012). You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 46(6):710 – 718.

Seiji Tsuchiya, Eriko Yoshimura, R. F. H. W. (2009). Emotion judgment based on relationship between speaker and sentential actor. *Knowledge-Based and Intelligent Information and Engineering Systems Lecture Notes in Computer Science*, 5771:62–69.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363 – 373.