Improving user profile with personality traits predicted from social media content

Conference Paper · October 2013

DOI: 10.1145/2507157.2507219

6 authors, including:

Rui Gao
Sun Yat-Sen University
294 PUBLICATIONS **5,996** CITATIONS

SEE PROFILE

Bibo Hao
IBM
25 PUBLICATIONS **311** CITATIONS

SEE PROFILE

Shuotian Bai
33
21 PUBLICATIONS **196** CITATIONS

SEE PROFILE

Ang Li
Beijing Forestry University
38 PUBLICATIONS **410** CITATIONS

SEE PROFILE

Improving User Profile with Personality Traits Predicted from Social Media Content

Rui Gao
gaorui11@mails.ucas.ac.cn

Bibo Hao
haobibo@gmail.com

Shuotian Bai
baishuotian10@mails.ucas.ac.cn

Lin Li
lilin10@mails.ucas.ac.cn

Ang Li
liang08@mails.ucas.ac.cn

Tingshao Zhu*
tszhu@psych.ac.cn

Institute of Psychology, University of Chinese Academy of Sciences
Beijing 100101, P.R. China

ABSTRACT

Existing studies indicate that there exists strong correlation between personality and personal preference, thus personality could potentially be used to build more personalized recommender system. Personality traits are mainly measured by psychological questionnaires, and it is hard to obtain personality traits of large amount of users in real-world scenes. In this paper, we propose a new approach to automatically identify personality traits with Social Media contents in Chinese language environments. Social Media content features were extracted from 1766 Sina micro blog users, and the predicting model is trained with machine learning algorithms. The experimental results demonstrate that users' personality traits could be predicted from Social Media contents with acceptable Pearson Correlation, which makes it possible to develop user profiles for recommender system. In future, user profiles with predicted personality traits would be used to enhance the performance of existing personalized recommendation systems.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Linguistic processing

Keywords

Text Analysis; User profile; Personality; Social Network; Social Media

1. INTRODUCTION

The Internet is growing dramatically and has become part of daily life. Recommender systems using personalized user profiles are widely used in e-commerce sites. Exiting researches about personalized recommendation system enhanced recommendation system performance by taking users' personal preference into consideration when generating user profiles [8]. Some studies have been performed for discovering the correlation between psychological traits(personality) and personal preference [11] [15] [18]. Results of these researches indicate that personality could affect the desire of online purchase, musical taste, etc. Improving user profile with personality traits could probably enhance personalized recommendation system performance. Social Media is one of the most popular kinds of websites, and users are generating large amount of text contents. Sina micro blog is the most popular micro blog in China, which is reported as having 300 million register users and over 100 million micro blog statuses were posted every day [1]. Personalizing Recommendation Systems based on social media usage behavior and content is among the top topics of the overall Internet area in both academia and business. However, user personality traits were mainly measured with psychological questionnaire. It is hardly possible to obtain personality traits of such large amount of users in real-world personalized recommendation system application scenes. In this paper, we propose a new approach to automatically develop user profile (personality traits) with Social Media contents in Chinese language environments. Sogou Cell Lexicons and Simple Chinese dictionaries for LIWC were built for psychological traits analysis in text contents.

Personality is a typical and enduring topic in psychology. The rapid growth of SNS(Social Network Sites) Usage provides a possible approach to automatically determine human personality.

Jon Oberlander found that classifying author personality from web blog text is promising [17]. Golbeck presented a promising method to predict user personality through the information on SNS in English language environment [6] [7].

Nass had demonstrated that personality factors could affect the desire of the users to buy products online [14]. Personality was found to be related to musical taste too [18]. Many other studies have been done for discovering the correlation between personality and personal preference [15]. China will become the biggest market of e-commerce in the world within a few years. Both Chinese SNS and e-commerce websites will significantly benefit from knowing users' personality and provide personalized service.

In this study, a method by which user big five personality traits can be accurately determined through text features extracted from Sina Micro blog status was demonstrated. To

*Author for correspondence.

RecSys'13, October 12–16, 2013, Hong Kong, China.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2409-0/13/10 ...$15.00.
http://dx.doi.org/10.1145/2507157.2507219.

the best of our knowledge, this paper may probably make the first attempt to determine personality from micro blog text features in Chinese language environment for recommendation system user profile enhancement.

The rest of the paper is organized as following: Section 2 shows the related work. The method by which data corpus was collected, processed and experiment results will be shown in Section 3. Section 4 concludes this work and indicates an open space for future work.

2. RELATED WORK

2.1 Personality and Personal preferences

Musical tastes are related to personality [14] [19] [18], individuals with similar personality traits prefer to similar types of music. Individual's potential music purchase could be sorted with deals which had been made by other people who share the same personality. People prefer to use graphical user interfaces designed for their personality types [11]. When people review product comments in e-commercial web sites, their desires to buy the product improve significantly if the comments were written by people who have similar personality types [14]. In western, some successful attempts had been made to improving recommendation systems with information getting from social network [14].

Personality could be predicted from micro blog statuses and individuals' preferences are related to personality, providing us a new approach to improve the performances of existing recommendation systems. Otherwise, few study has been conducted in Chinese language environment. China will probably become the biggest market for e-commerce, so research on predicting personality and enhancing recommendation systems in Chinese language environment is essential. In this paper, a new approach was proposed to develop user profile (personality traits) for Recommendation Systems in Chinese language environment.

2.2 Language Usage and Personality

Studies had been done to investigate the relationship between Language Usage and Personality [12] [17] [16]. Linguistic cues in conversation and text were proved to be capable to recognize individuals' personality [12]. Web blog texts generated by individuals could be used for classifying author personality [17] [16], it indicates that text content in the internet environments is related with authors' personality as well as language they use in real world. One study based on the corpus of CRV1 and Enron Email data demonstrates that language features can be used to predict author's gender with accuracy of 85.1 percent [3]. In English language environment, a promising method to predict user personality through their information on SNS was presented by Golbeck [6] [7]. Micro blog status text features of 50 Twitter users' were extracted by LIWC(Linguistic Inquiry and Word Count) software. The correlation ship between text features and personality scores are significant [6]. A model for personality prediction was trained and tested with 279 Facebook users information, and the predicted personality scores and self-reported personality scores were strongly related [7].

All these studies indicate that predicting personality from Sina micro blog text content is feasible, and have potential

3. METHOD

Big five personality model is recently among the most regarded and most widely used personality theory [4] [5] [13]. Big five personality model has been proved effective in different languages [5]. In the Sina Micro blog application, users were invited to complete the 44-question Big Five Inventory [2] [10]. In this inventory, user personality is measured in five dimensions:

- Openness to Experience.
- Conscientiousness: responsible, organized, persevering.
- Extroversion: outgoing, amicable, assertive.
- Agreeableness: cooperative, helpful, nurturing.
- Neuroticism: anxious, insecure, sensitive.

Table 1 shows the score range of each dimension.

3.1 Data Collection

Original status texts of 1766 Sina micro blog users were analyzed in this study. An application was built based on Sina micro blog platform to collect status and administer Big-Five Inventory [8] to users. By calling Sina Application Programming Interface, the basic information (the number of published status, followers and friends) of 99,925,821 users was downloaded in 2012/4/18. These 99,925,821 users were the original user sample set of this text analysis.

Two kinds of users may be inappropriate to be taken into consideration in this text analysis. One kind is the inactive users. They published little original status, so their personality can hardly be reflected by text features of their status. The other kind is entertainment star users or advertisement users. Since almost all this kind of micro blog accounts are used for commercial targets. Thus their status should not be analyzed for personality related research.

All the 99,925,821 users were filtered by the following rules:

1. Users who did not publish any status in the last three months or published less than 512 statuses in total were excluded. In the original user sample set, the number of statuses of each user published is 136.6511 on average with a deviation of 788.8671. Users who published less than 512 (mean + 0.5 * standard deviation) statuses are probably inactive users. All the 99,925,821 users published 2.84 statuses per day in average. Users published less than 2.84 status per day in average were excluded, too.

2. Users who publish more than 40 statuses per day are highly likely to be entertainment star users or advertisement users. They were excluded too.

Then, original status texts of 1,953,485 active users were collected as user sample set for this text analysis. About 30,000 users were randomly picked, they were invited to accomplish widely used O. John's brief 44-item Big-Five Inventory [8] on our application based on Sina micro blog. 1766 volunteers accepted the invitation int the end and accomplished Big-Five Inventory. In this study, these 1766 volunteers' original status texts and Big-Five Inventory result were used as label sample to train predicting model and cross-validation.

3.2 Feature extraction

Text features were extracted from 1766 user's Sina micro blog status, and then all the features vectors were analyzed

Table 1: Big Five Inventory scores range in five dimensions.

Dimension	Openness	Conscientiousness	extroversion	Agreeableness	Neuroticism
Range	[10,50]	[9,45]	[8,40]	[9,45]	[8,40]

to build a predicting model and 10-fold cross-validation with the algorithms Gaussian process, M5'Rules and Pace Regression. One hundred and sixty eight Text features were extracted in six perspectives:

1. Status statistics features (total number of statuses, total number of words in status, etc.);

2. Sentence-based features (average number of Chinese characters per sentence, declarative sentence proportion etc.);

3. Word-based features (number of emotion words, number of greeting words etc.);

4. Character-based features (number of commas, number of colons and number of 23 special characters etc.);

5. Sogou Industry lexicon features. Pinyin input method is the most popular Chinese input method in China, it takes 70% share of the whole Chinese netizens. Sogou Cell Lexicons are optional lexicons users can download from Sogou input method official web site. These lexicons are automatically generated by Sogou search engine or created by netizens. All the cell lexicons are uploaded on the Sogou official web site and rated by all Sogou input method users. Eighty eight official recommended and highest rating cell lexicons were downloaded and merged into 13 industry lexicons, such as, art lexicon, medical lexicon, network popular lexicon, etc. All the weibo statues in the data corpus were processed Chinese word segmentation using the latest version of ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System), and word count in each Sogou Industry Lexicons were added up as user text features;

6. LIWC(Linguistic Inquiry and Word Count) features. LIWC is text analysis software widely used by scholars [6] [7] [9]. It can determine the degree of language usage in 88 dimensions. Three versions of Simple Chinese dictionaries for LIWC were generated by master degree candidates, and a final version was developed by manually validating these three dictionaries according to CLIWC dictionary [9]. LIWC software was ran with the final version dictionary to calculate 88 micro blog status text features.

The above features are probably the most related text features to subject information and personality [3]. Chinese text analysis has a challenge: Chinese Word Segmentation. ICTCLAS is one of the most widely used Chinese words parser. Sogou Cell Lexicons and Simplified Chinese dictionaries for LIWC have been both imported into ICTCLAS as user dictionaries to improve recognition rate of parser.

3.3 Model Training and Evaluation

With randomly selected about 90% of 1766 Sina users, M5'Rules, Pace Regression and Gaussian process were applied to build predicting models. These models were used to predict the rest about 10% users' personality traits, and their performances were evaluated by Pearson correlation between predicted personality traits scores and Big-Five Inventory ratings. Pearson correlation is among the most major criteria of correlation ship in psychology, thus it is taken as evaluation criteria to measure the accuracy of the predicting model. It is widely accepted that Pearson correlation between 0.2 to 0.4 indicates weak correlation, 0.4 to 0.6 indicates mid correlation, and above 0.6 indicates strong or extreme correlation.

3.4 Results and Discussion

The results in Table 2 indicate that personality traits could be predicted through text features extracted from original micro blog status. Conscientiousness, Extroversion, and Openness to Experience have more significant correlation than Agreeableness and Neuroticism. That is consistent with many other studies [2] [14] [15]. In psychological point of view, the predicted personality has a bit weaker correlations with self-report personality than expertise observer ratings has [8]. The Pearson correlation between the predicted personality and self-report is about 0.4, which means that they are fairly correlated. Since personality provides some clues to identify new users' preferences, it might be an effective approach to alleviate cold start of recommendation system. Further works could be conducted to improve the correlations between predicted personality and self-reported personality. As predicted personality becomes more and more accurate, it could be used for personalized recommendation systems.

Some correlations between text features and personality dimensions are intuitively sensible. For instance, people using second person pronouns "you"("ni" and "nimen" in Chinese for single and plural) more often are more extraverted, and the more internet popular new words confers higher openness to experience. People with high scores in Neuroticism generated more words about religion and art. Other correlations between features and personality have not been interpreted yet. Exclamation mark usage is highly positively related to conscientiousness and more count of numbers usage and agreeableness are significantly negative related, which leaves an open space for future researches to confirm and explain.

4. CONCLUSIONS

Experimental results demonstrate that Big Five personality traits can be predicted from the users' Sina micro blog status. Personality was proved to be related with personal preferences in music tastes and graphic user interfaces [11] [14] [18]. Recommendation systems accuracy could probably be improved by taking users' personality traits into consideration. This work might provide a promising future to improve user profile with predicted personality traits, which could potentially improve the performances of existing recommendation systems.

On the other hand, while using LIWC2007 to process Chinese content, we found that it might not process Chinese

Table 2: The Evaluation Results on Predicted and Self-report Big Five Scores.

Key Data set	Number of training instances	Number of testing instances	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Correlation coefficient
A	1590	176	3.7559	4.83402	96.4699	96.26233	0.31079
C	1590	176	3.9956	4.91633	93.0128	92.1287	0.40601
E	1589	177	4.1850	5.16627	93.5224	93.42278	0.40312
N	1589	177	4.4448	5.57428	97.235	95.41558	0.31958
O	1590	176	4.5034	5.49945	94.4791	93.06926	0.37768

properly. We are developing a prototype system TextMind that is optimum for Simplified Chinese analysis. TextMind might could provide an all-in-one solution on Simplified Chinese processing for researchers in the future.

5. ACKNOWLEDGMENTS

The authors gratefully acknowledge the generous support from National High-tech R&D Program of China (2013AA01A606), NSFC (61070115), Institute of Psychology (113000C037), Key Research Program of CAS (KJZD-EW-L04), Strategic Priority Research Program (XDA06030800) and 100-Talent Project (Y2CX093006) from Chinese Academy of Sciences.

6. REFERENCES

[1] Sina weibo has more than 300 million users, over 100 million statuses are posted, 2012. http://news.xinhuanet.com/tech/2012-02/29/c_122769084.htm.

[2] V. Benet-Martinez and O. P. John. Los cinco grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in spanish and english. *Journal of Personality and Social Psychology*, 75(3):729–750, 1998.

[3] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[4] P. T. Costa and R. R. *Revised NEO Personality (NEO-PI-R) and NEO Five-Factor Inventory professional manual*. Psychological Assessment Resources, Odessa, FL, 1992.

[5] J. M. Digman. Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(1):417–440, 1990.

[6] J. Golbeck, M. E. C. R., and K. Turner. Predicting personality from twitter. In *2011 IEEE third international conference on andieee third international conference on social computing (socialcom)*, pages 149–156, Boston, 2011.

[7] J. Golbeck, C. Robles, and K. Turner. Predicting personality with social media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 253–262, New York, NY, USA, 2011. ACM.

[8] S. D. Gosling, P. J. Rentfrow, and W. B. S. Jr. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003.

[9] C.-L. Huang, C. K. Chung, N. Hui, Y.-C. Lin, S. Yi-Tai, B. C. P. Lam, W.-C. Chen, M. H. Bond, and J. W. Pennebaker. The development of the chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology*, 55:185–201, 2012.

[10] O. P. John, L. P. Naumann, and C. J. Soto. Paradigm shift to the integrative big-five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, and L. A. Pervin, editors, *Handbook of personality: Theory and research*, pages 114–158. Guilford Press, New York, NY, 2008.

[11] A. Karsvall. Personality preferences in graphical interface design. In *ACM International Conference Proceeding Series*, pages 217–218, 2002.

[12] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007.

[13] R. R. McCrae and O. P. An introduction to the five-factor model and its application. *Journal of Personality and Social Psychology*, 60(2):175–215, 1992.

[14] C. Nass and K. M. Lee. Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In *Computer Human Interaction*, pages 329–336, 2000.

[15] E. Neustadt, T. Chamorro-Premuzic, and A. Furnham. The relationship between personality traits, self-esteem, and attachment at work. *Journal of Individual Differences*, 27(4):208–217, 2006.

[16] S. Nowson and J. Oberlander. Identifying more bloggers Towards large scale personality classification of personal weblogs. In *International Conference on Weblogs and Social Media*, 2007.

[17] J. Oberlander and S. Nowson. Whose Thumb Is It Anyway? Classifying Author Personality from Weblog Text. In *Meeting of the Association for Computational Linguistics*, 2006.

[18] 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*, 84(6):1236–56, 2003.

[19] J. L. Zhou, Y. Fu, H. Lu, and C. J. Sun. From popularity to personality - a heuristic music recommendation method for niche market. *Journal of Computer Science and Technology*, 26(5):816–822, 2011.