IJAMML 4:1 (2016) 43-63 March 2016
ISSN: 2394-2258
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/ijamml_7100121623

IDENTIFICATION METHOD OF CHARACTER TRAITS BASED ON WECHAT DATA

Qinfeng Li[a,b], Xiaofeng Zhou[b], Aihua Gu[b] and Zonghua Li[b]

[a]Department of Basic, Jinling Institute of Technology, P. R. China
[b]Hohai University, College of Computer and Information, P. R. China

Abstract

In recent years, WeChat App is used widely as a convenient communication tool, however most attention to WeChat is paid on the commercial value, few to the academic research. Moreover, there are very few studies about identifying adolescent character traits from WeChat data. For the above, a method of determining character traits is proposed. We discuss the characteristics of data from status, content and neighbour of account data, and extract features. Then Naive Bayes (NB), Support Vector Machine (SVM), K Nearest Neighbour (KNN), and Particle Swarm Optimization (PSO) algorithms are used to classify them according to their contributions to traits. Experimental results show that the means of classification based on data features can effectively identify character traits.

Keywords: feature extraction, classification algorithms, dataset character traits.

*Corresponding author.
E-mail address: haier20022@126.com (Qinfeng Li).

2010 Mathematics Subject Classification: 97R20.
Submitted by Li Li.
Project supported by the Foundation of modern educational technology research of Jiangsu Province (No. 2015-R-42631).
Received January 13, 2016; Revised February 15, 2016

1. Overview

WeChat is an app developed by the Tencent. It can send short information, text, voice, picture, and video to friends. With the rapid development of mobile communication and network technology, WeChat grows more powerful gradually, and it is widely used in recent years. Adolescent are important users because of their curiosity about new things and their acceptance, so WeChat has become very important communication way. However, young people are unsophisticated and poor self-discipline, they are exerted a tremendous influence.

On one hand, many researches focus on WeChat's commercial value, functional modules, communication modes and features, software applications and implementations, network architecture, etc. However, there are few researches paying attention to character traits based on WeChat. On the other hand, adolescent's dependence on WeChat and cellphone, as well as their rebellious against concerns from seniority, make researches become an effective way of identifying their characters and guiding them to the favorable development.

In this paper, we study the original data in WeChat app from some adolescent. We propose a method of character recognition based on the feature analysis. The work includes: We study the characters of accounts data from status, content and neighbours, and build the original feature dataset. Then we test and evaluate the effect of our method in variety algorithms. In the paper, we only distinguish introvert with extrovert.

The remainder of the paper is organized as follows. Section 2 reviews some related work. Section 3 describes the framework of our method. Section 4 presents analysis and extraction progress of feature set. In Section 5, the experimental results are discussed. Conclusions are made in Section 6.

2. Related Works

WeChat is developed in recent years as a convenient communication way. High interest is paid since it comes out, however the academic discussion and researches are few. In inquiry system of known online, we find some related works focused on WeChat's characteristics of the propagation [1], mechanism of transmission [2-4], presentation of commercial marketing [5], and functional mode [6]. A few works on challenges and developing ways about the ideological and political education, discussed from functional modules [7]. There is little works on research about the thinking way and character traits, let alone systematic analysis character traits from WeChat data.

WeChat's evaluations and expectations abroad are high. It is considered to a killer application, because if one uses, all of his friends will use. In order to contact and communication with friends, he needs to use continuously. However, the foreign study and attention to WeChat are focused on business. Academic research are precious little [8, 9]. Only several papers are published in foreign journals or meetings, and they are finished by Chinese researchers. These studies are about framework [10, 11], function modules [12, 13], behaviours of users [14] and application in ESP training [15], market [16, 1], latent evidences, etc [17-19]. These papers discuss mainly about communicational way, design and implementation of application software, function analysis, network architecture, etc. They reflect no culture and psychological characteristics.

3. Framework of Character Recognition

The paper draws ideas of classification and feature extraction in data mining [20]. It integrates characteristics analysis into classification knowledge. Then it proposes the character recognition framework based on WeChat data. The framework is shown in Figure. 1.

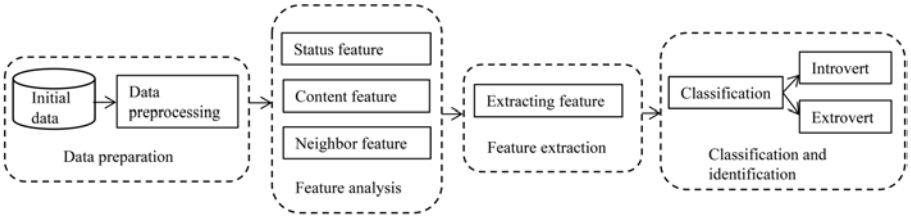


Figure 1. The framework of character recognition.

As can be seen from Figure 1, the frame breaks up into data preparation, feature analysis, feature extraction, classification and identification.

(1) Data preparation: The data collected from terminal units are treated with several processes, including eliminating or formalizing information of locations, transaction records, passwords, photo albums, and other privates.

(2) Feature analysis: The process includes analyzing the feature of data respectively reflecting the status, content, neighbour of accounts.

(3) Feature extraction: The object of the process is to find feature set employing extraction techniques.

(4) Classification and identification: The process includes selecting better sort algorithms, determining the corresponding relationships of data characteristics and character traits, evaluating the performance of sort algorithms.

4. Data Characteristics

To study the relationship between data characteristics and character traits, as comprehensive and accurate as possible, the paper analyzes data separately from status, content and neighbours of accounts to build the feature set.

4.1. Feature of account status

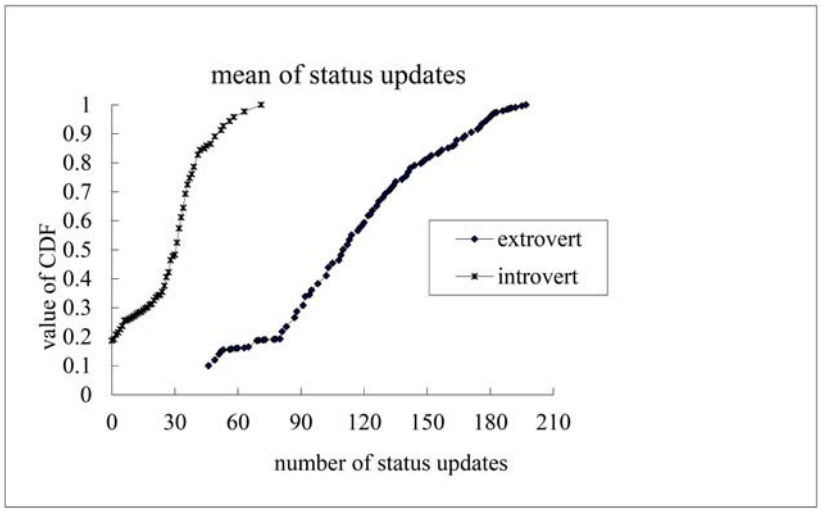
The status of account reflects the basic information in the field of updates of status, praises and comments, levels of speak, active friends, links in circle of friend, public accounts of concerned, transactions etc. Main information can significantly make the difference between character traits.

To extrovert users, updates of status information are of frequent occurrence. The relationships with others are close. Every message they released will attract a lot of friends to reply or comment. Every link they transmitted will be spread widely in a minute. Friends in list are added on their own initiative. They prefer to transmit links to share with others. However, to introvert, they prefer to collect links in case of need. To distinguish extrovert and introvert obviously, two new features are set as base of basic information about account. That is Activity and Rigor. These subjects are as follows:

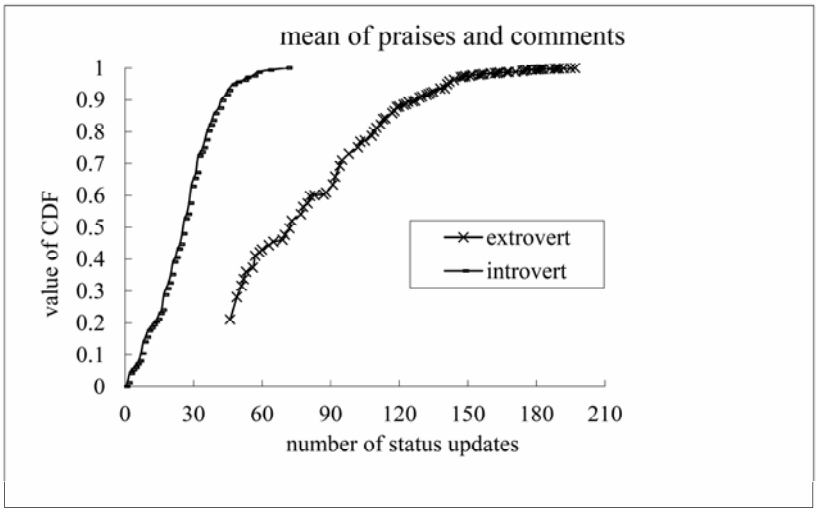
Activity: The interpersonal relationship and ability of expression are reflected in the number and frequency of praises, comments and status updates; in the proportion of friends active and on their own initiative; in the ratio of collect links and transmit links, of speak records and all records in chat group, of speak length and spacing interval. Generally speaking, the measure of Activity of the extrovert users is higher than the introverts.

Rigor: The safety consciousness is expressed as the frequency of transaction; as the proportion of hits of links or software collecting information; as the close rate of various security buttons. Generally speaking, the measure of Rigor of the extrovert users is lower than the introverts.

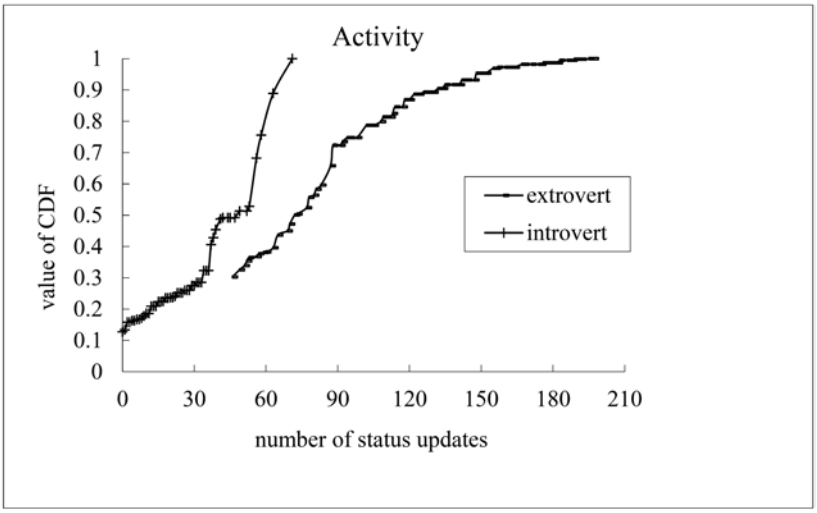
The partial characteristics of the status are described as the cumulative distribution function (CDF) curve in Figure 2 about the extroverts and introverts.



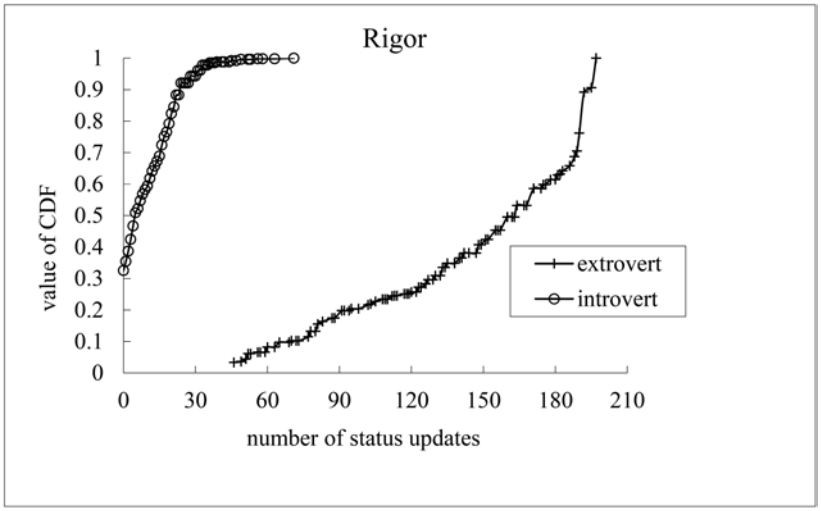
(a) Status update



(b) Praises and comments



(c) Activity



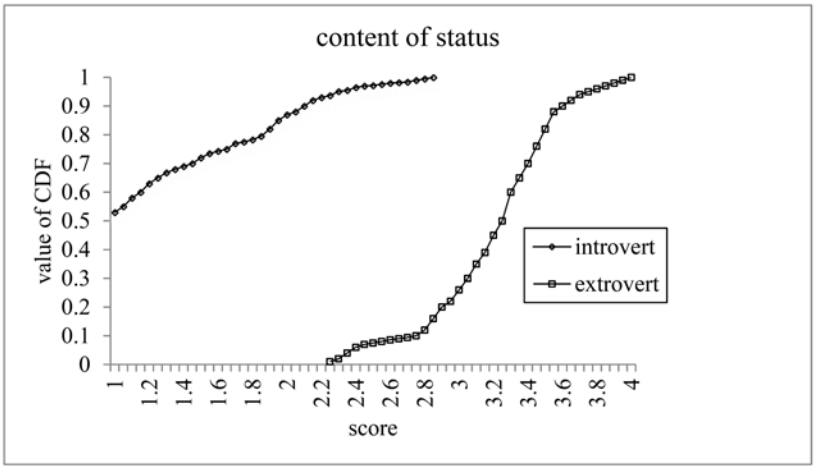
(d) Rigor

Figure 2. The cumulative distribution function curve of account status.

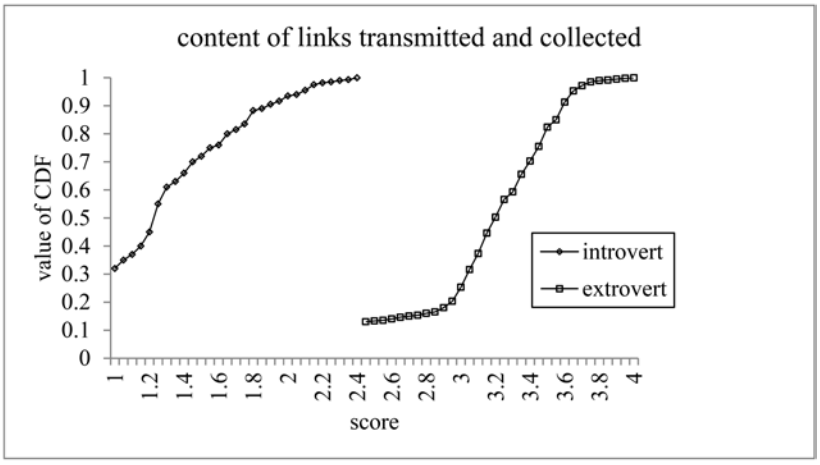
4.2. The feature of account content

The feature of account content refers to the theme of users concern and interest. It can be expressed as the category of public accounts concerned, of the content of users' status, of links transmitted and collected. From amount of data, it can be decided that those who prefer to concern beer and skittles mostly are extrovert, those who prefer to psychological or logical reasoning mostly are introvert. To represent the features of account contents better, here cites the 0 to 4 score rule to decide the score of content. Particularly, 4 says the extrovert, 3 the fairly extrovert, 2 the fairly introvert, 1 the introvert.

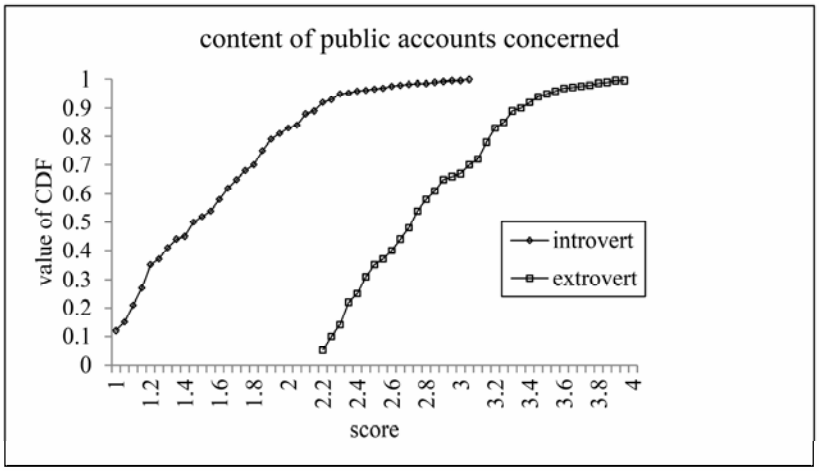
The partial characteristics of the account content are described as the cumulative distribution function curve in Figure 3 about the extroverts and introverts.



(a) Status



(b) Links



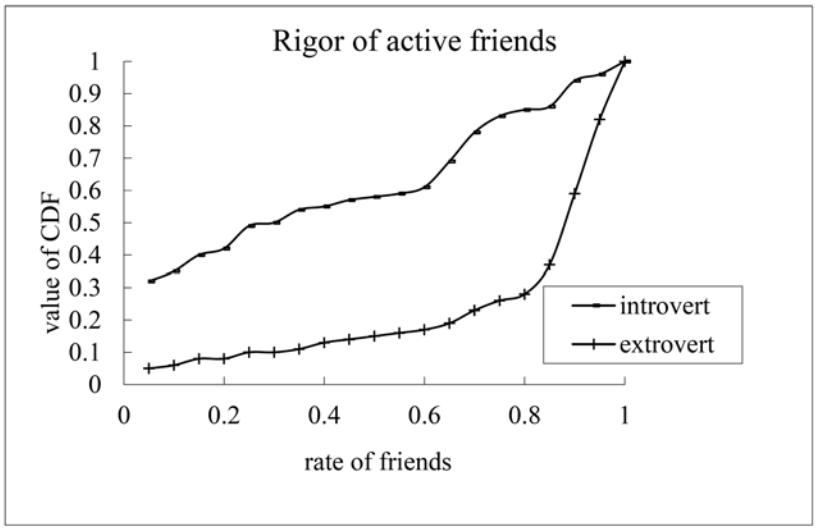
(c) Public account

Figure 3. The cumulative distribution function curve of account content.

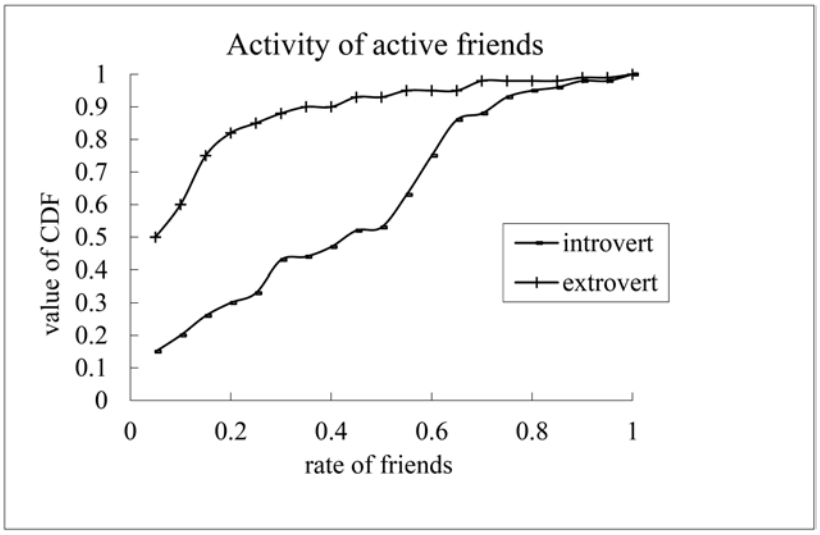
4.3. The feature of account neighbours

The feature of account neighbours expresses the account concerned and been concerned. It can be expressed as the number of active friends, as the links transmitted and been transmitted, as the measure of Activity and Rigor of friends. These features can effectively reflect the neighbour accounts. To a degree, these also reflect features of one's own account. Under normal field conditions, friends of the extrovert mostly are active, and get a high score of Activity.

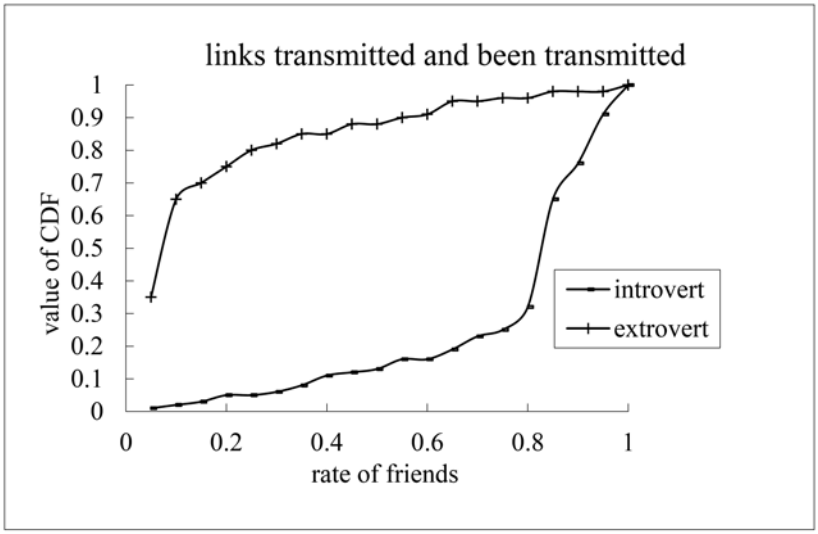
The partial characteristics of the account neighbours are described as the cumulative distribution function curve in Figure 4 about the extroverts and introverts.



(a) Rigor



(b) Activity



(c) Links

Figure 4. The cumulative distribution function curve of account neighbours.

The paper analysis the information of account from the features of account status, account contents and account neighbours. It extracts 33 features to illustrate the difference between the extrovert and the introvert personality. Due to the diversity of selected features, the range of characteristics has large differences. It will make the accuracy of classification decline if use features without processing. Therefore, all the characteristics must be normalized before use. The detail is as follows: If feature value is F, the normalized value is $F_0 = lg(F + 1)$.

The 33 features are as follows:

(1) The features of account status attributes are including:

Number of friends: $F_1 = Yn(u)$.

Number of status update: $F_2 = Zg(u)$.

Average of status update: $F_3 = F_2 / T_{\max} - T_{\min}$, where $|T_{\max} - T_{\min}|$ denotes the span time of date.

Frequency of status update: $F_4 = \begin{cases} 1 - 1/F_3 & F_3 \neq 0 \\ 0 & F_3 = 0 \end{cases}$.

Number of praises and comments: $F_5 = Pz(u)$.

Number of friends to praises and comments: $F_6 = PYn(u)$.

Average of praises and comments: $F_7 = F_5 / F_2$.

Frequency of praises and comments: $F_8 = \begin{cases} 1 - 1/F_7 & F_7 \neq 0 \\ 0 & F_7 = 0 \end{cases}$.

Number of responses to praises and comments: $F_9 = RPYn(u)$.

Number of public account concerned: $F_{10} = Gz(u)$.

Total number of speakers in a chat group: $F_{11} = Yzn(u)$.

Number of statements: $F_{12} = Yn(u, i)$, where i denotes the number of the i-th particular statement.

Levels of speakers: $F_{13} = Sd(u)$.

Length of statements: $F_{14} = Yl(u, i) = \frac{1}{F_{12}} \sum_{i=1}^{F_{12}} St(u, i)$, where $St(u, i)$ denotes the time of the i-th particular statement.

Time of statement last: $F_{15} = Yg(u, i) = T_{i1} - T_{i2}$, where T_{i1}, T_{i2}, respectively denotes the beginning and end time of the i-th particular statement.

Number of transmitted links: $F_{16} = Tl(u)$.

Number of collected links: $F_{17} = Cl(u)$.

Number of transactions: $F_{18} = En(u)$.

Rate of unsafe links: $F_{19} = USlp(u) / USl(u)$, where $USlp(u)$, $USl(u)$, respectively denote the clicks and totals of unsafe links.

Close rate of security buttons: $F_{20} = SNnp(u) / SNn(u)$, where $SNnp(u)$, $SNn(u)$, respectively denote the number of security buttons shut down and totals.

Number of friends added on one's own initiative: $F_{21} = NFAo(u)$.

Activity: $F_{22} = A(u) = \frac{1}{9}\sum a_j F_j$, $j = 1, 4, 6, 8, 13, 15, 16, 17, 21$, where a_j is the weighting factor.

Rigor: $F_{23} = UA(u) = \frac{1}{3}\sum a_j F_j$, $j = 18, 19, 20$.

(2) The features of the account content status attributes, including as follows:

Score of category of public account concerned: $F_{24} = Gzh(u) = \frac{1}{i}\sum a_i sim_p(u, u_i)$, where $sim_p(u, u_i)$ denotes the similarity of category of public account concerned and themes, themes u_i are food, entertainment, shopping, travel, logic, adventure, psychology, health and others.

Score of status content: $F_{25} = Zgh(u) = \frac{1}{i}\sum a_i sim_s(u, u_i)$, where $sim_s(u, u_i)$ denotes the similarity of status content and themes.

Score of links content transmitted and collected: $F_{26} = Ljh(u) = \frac{1}{i}\sum a_i sim_l(u, u_i)$, where $sim_l(u, u_i)$ denotes the similarity of links content and themes.

(3) The features of account neighbours, including as follows:

Number of active friends: $F_{27} = AY(u)$, where u is regarded as active friend just as his Activity $A(u) \geq 0.5$, his Rigor $UA(u) \leq 0.5$.

Rate of active friends: $F_{28} = F_{27} / F_1$.

Activity of friends: $F_{29} = YA(u, i)$, where $YA(u, i)$ denotes the Activity of the i-th friend. Rigor of friends: $F_{30} = YAU(u, i)$, where $YAU(u, i)$ denotes the Rigor of the i-th friend.

Average of links transmitted: $F_{31} = JSl(u) = Tl(u) / T_{\max} - T_{\min}$.

Number of links been transmitted: $F_{32} = ATl(u)$.

Average of links been transmitted: $F_{33} = JATl(u) = ATl(u) / T_{\max} - T_{\min}$.

5. Experimental Results and Analysis

5.1. Dataset

The raw data here are from the mobile terminals. The data have been de-noised before use. The characteristics are extracted with mining algorithms of latent semantic analysis (LSA), stored in the appropriate database. To reduce the run time and run space, we only adopt the last two months of data from the account records. Previous studies [21] have shown that section of data can reflect the overall characteristics to some extent.

The task of the paper is to distinguish between extrovert and introvert. When data are collected users have to make the self-assessment. Ultimately, we collected 536 users' data. The experiments are repeated with several algorithms. Results are compared with the self-assessments. Here data are used as training sets and test sets alternately.

5.2. Evaluation

Since we study users' characters, we only have users' self-evaluation results. We cannot decide whether those are consistent with facts. Therefore, we compare the self-assessments with the test results, and select the larger value as the actual result. Table 1 shows the confusion

matrix of the actual and the self-assessments. Precise rate P, recall rate R, accuracy rate A, and metric $F1$ are as the evaluation indicators. Their mathematical formulas are as follows:

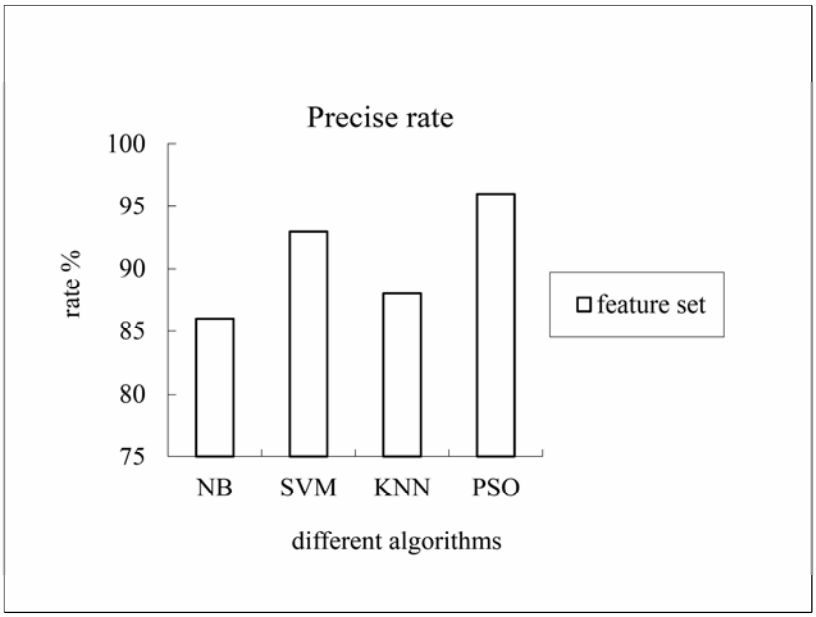
$$P = a / a + c,\ R = a / a + b,\ A = a + d / a + b + c + d,\ F1 = 2PR / P + R.$$

Table 1. The confusion matrix of the actual and the self-assessments

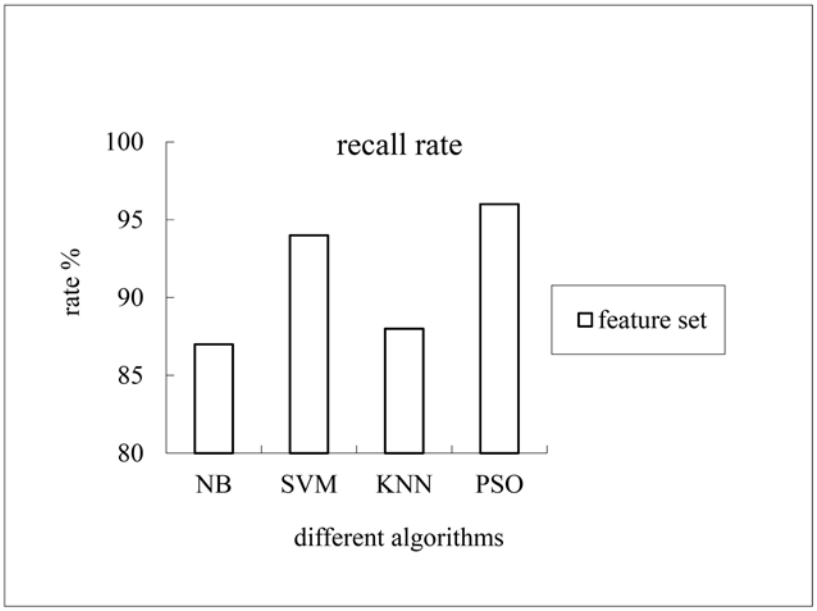
	self-assessment results	
actual results	extrovert	introvert
extrovert	a	b
introvert	c	d

5.3. Analysis

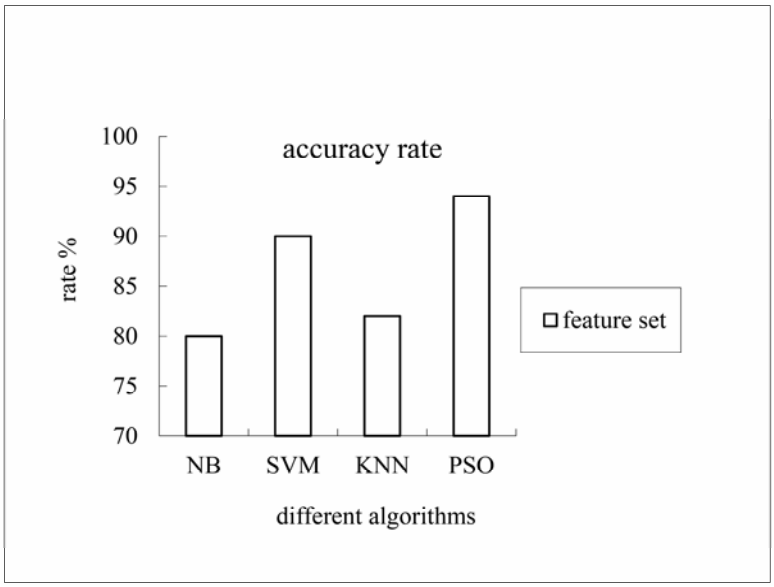
To judge the assessments of the feature subset, we do experiments on the feature set based on four classification algorithms of NB (Naive Bayes), SVM (Support Vector Machine), KNN (K Nearest Neighbour,) and PSO (Particle Swarm Optimization), using 10-fold cross-validation alternately. The classified results of character traits are shown in Figure 5.



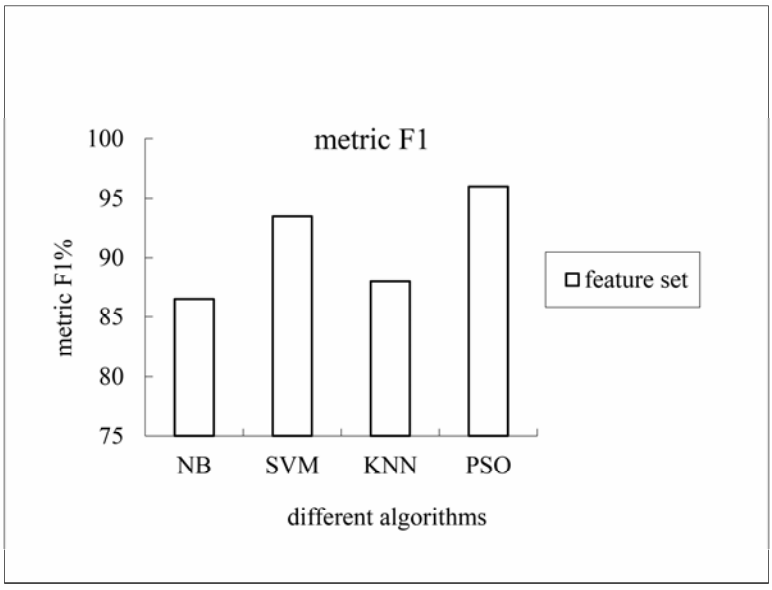
(a) Precise rate



(b) Recall rate



(c) Accuracy rate



(d) F1 metric

Figure 5. Comparison of different algorithms on the original feature set.

As can be seen from Figure 5, the differences of different algorithms are little. All algorithms are effective, where the best is PSO. Its precise rate P is 96%, recall rate R is 96%, accuracy rate A is 94%, and metric $F1$ is 96%. So, the method of based on data feature analysis can effectively identify the character traits. The accuracy rates of the test results are different with the actual results. There are may be these factors that users do not understand themselves, users have double characters, classification algorithms are defective, numerical methods and combinations of feature extraction need improvement.

6. Conclusion

In this paper, we analysis WeChat data deeply and propose a new method based on data features to identify the user personality. It extracts features from account status, account content, account neighbours. Then, it employs different classification algorithms to evaluate the performance of the feature set. Experiments show that the method presented here can effectively identify the users' character traits. The next step of our work we will gradual refine the classification of characters to make a better appropriate assessment, improve method of extracting and combining features to obtain a better effect, and prepare for subsequent data work.

References

[1] Q. F. Cai, J. F. Guo, S. H. Yi, etc., Fund Managers' WeChat Behavior, IPO Underpricing, and Earnings Volatility (May 5, 2015). Available at SSRN:

http://ssrn.com/abstract=2607045

[2] L. L. Wu and X. J. Yang, Research on the services of university mobile library based on the Wechat public platform, Research on Library Science 61(51) (2013), 18-57.

[3] C. F. Ji and Z. F. Zhou, Problems and countermeasures of the WeChat public platforms applied in Project 211 University Libraries, Research on Library Science 17 (2014), 38-41.

[4] J. H. Xiao and L. H. Huang, Research on information service mode of the library based on WeChat, Journal of Modern Information 33(6) (2013), 55-57.

[5] H. Tong, Research on WeChat according to communication study and its influence, Chongqing Social Sciences 09 (2013), 61-66.

[6] D. H. Kuang, J. Q. Hao and L. X. Ke, The subject service marketing based on WeChat, Library Work and Study 09 (2014), 123-125.

[7] W. He and K. Liang, New Weapons of Ideological and Political Education in Universities – WeChat, SHS Web of Conferences, IFSRAP 2013 – The First International Forum on Studies of Rural Areas and Peasants.

http://dx.doi.org/10.1051/shsconf/20140604001

[8] H. L. Che and C. Yang, Examining WeChat users' motivations, trust, attitudes, and positive word-of-mouth: Evidence from China, Computers in Human Behavior 41 (2014), 104-111.

[9] K. Zhang, Mining data from Weibo to WeChat: A comparative case study of MOOC communities on social media in China, International Journal on E-Learning 14(3) (2015), 305-329.

[10] Y. T. Huang, X. Z. Lai, B. P. Dai, etc., Web-of-Things Framework for WeChat, Internet of Things, 2013 I (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing (2013), 1496-1500.

[11] S. H. Hua and H. Wei, Study on knowledge propagation in complex networks based on preferences, Taking WeChat as example, Abstract and Applied Analysis 2014 (2014), 543734.

http://dx.doi.org/10.1155/2014/543734

[12] H. Xia, Q. Wei and S. Y. Zhang, Research on Undergraduates Perception of WeChat Acceptance, in e-Business Engineering (ICEBE), 2014 IEEE 11th International Conference on, pp. 61-67, 5-7 Nov. 2014.

[13] L. M. Huang, Q. Y. Zhu, J. Ding etc., Design and realization of technology intelligence push based on WeChat, Applied Mechanics and Materials 631-632 (2014), 1119-1122.

[14] D. L. Song, Y. J. Wang and F. You, Study on WeChat User Behaviors of University Graduates, Digital Home (ICDH), 2014 5th International Conference on, pp. 353-360, 28-30 Nov. 2014.

[15] Z. W. Liu, A study on the application of WeChat in ESP training, Theory and Practice in Language Studies 4(12) (2014), 2549-2554.

[16] L. Yi, Introduction to the WeChat Marketing Advantages and Development Prospects, Bachelor's degree (UAS), Finland: Savonia-ammattikorkeakoulu.

http://www.theseus.fi/handle/10024/76332

[17] F. Gao and Y. Zhang, Analysis of WeChat on IPhone, 2nd International Symposium on Computer, Communication, Control and Automation (3CA 2013), 278-281.

[18] J. H. Xua, K. G. Qi, Z. Q. Song and Christopher Peter Clarke, Applications of mobile social media: WeChat among academic libraries in China, The Journal of Academic Librarianship 41(1) (2015), 21-30.

[19] C. Q. Wang, Research on mobile information service of the library based on WeChat, Applied Mechanics and Materials 701-702 (2014), 1008-1012.

[20] J. W. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, Third Edition, Beijing: China Machine Press, 2012.

[21] M. McCord and M. Chuah, Spam Detection on Twitter using Traditional Classifiers, Proceedings of the 8th International Conference on Autonomic and Trusted Computing, Piscataway, USA: IEEE Press, (2011), 175-186.

[22] M. Hall, E. Frank, G. Holmes etc., The WEKA data mining software: An update, SIGKDD Explorations 11(1) (2009), 10-18.

■