The Bloggers' Personality Traits Categorizing Algorithm Based on Text Features Analysis

Guohua Ou[1, a], Jingkai Li[1,b] , Junjia Guo[1,c], Zhaoquan Cai[2,d] and Mengping Lu[1]

[1]School of Software Engineering, South China University of Technology, Higher Education Mega Center, Guangzhou, China

[2]Huizhou University, Huizhou, Guangdong, China

[a]csghou@scut.edu.cn, [b]jkli@scut.edu.cn, [c]scut_junjia_guo@hotmail.com, [d]caizhaoquan@hzu.edu.cn

Keywords: Blog mining, text analysis, personality categorization.

Abstract. Nowadays, researches of blogs mining mainly concentrate on opinion mining, community mining, blogs recommendation system and so on, with little concentration on personalities mining. How to mine bloggers' personality accurately and effectively from the tremendous non-structural blog texts becomes a difficulty of blogs mining. This paper illustrates a research on categorizing bloggers' personalities based on the support vector machine(SVM), using the Big Five personality traits to categorize bloggers at Netease into two types of personalities, the extroverted personality and the introverted one, improving the results of categorization in the aspects of personality categorizing traits and approaches of traits selection and ultimately providing other researches about categorizing Chinese bloggers' personality traits with references.

Introduction

Personality is a stable but variable mental trait, reflecting someone's attitude and behaviors facing the reality. It plays a significant role in the fields of education, group management, communication and working so that it is more significant than intelligence to some extent [1]. Hence, mastering someone's personality trait can make deciders take some pertinent strategies to satisfy users' needs as well as obtain more profits for corporations [2].

At present, there are some achievements [3,4] from the researches of bloggers' personality mining abroad. However, no relevant achievements are found in China. Big Five is the most common personalities categorizing standard, which divides personalities into five major types, openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism.

This paper attempts to mine bloggers' traits [5] from their blogs and analysis the relationship between bloggers' personality traits and their blogs' contents, linguistic traits and behavior traits. The personalities discussed in the paper mainly involve two antitheses types of personalities, "extroverted" and "introverted". The extrovert means the person who spends interest on others or their events rather than himself or herself, while the introvert often avoids keeping in touch with others and mostly focuses on his or her own feelings, opinions and experiences [6,7].

This paper will discuss two types of personalities, "extroverted" and "introverted", and introduce the work of data preparation, traits categorization in the research, traits selection and the problem of categorizing personalities sequentially [8]. In this paper, the problem is considered as a dichotomy and the analysis of a related experiment has been done.

Traits Categorization

Data Description. The data from Netease Blog comprise tremendous personal data, involving bloggers' personalities. There are some inconsistencies between the types of personalities utilized by Netease and the types of Big Five. Only the "extroverted" or "introverted" personality traits [9] are the same part. Due to the reason, this paper only focuses on these personality traits to do the research.

A crawler program, coded by C#, has downloaded huge blog data, including bloggers' articles, lasting time of blogging, and the numbers of comments, pictures, hyperlinks in the blogs and the number of bloggers' friends. In this paper, only the data produced before Jun. 8th , 2011, the date when the program downloaded the data, has been used. When downloading, only the bloggers with "extroverted" or "introverted" traits have been reserved. The detailed numbers of downloaded blogs are displayed in Table 1.

Table 1. The corpus of bloggers' personality traits

Personality traits	No. of bloggers	No. of blogs	Disk space occupied by the blogs [Mb]
"extroverted"	1, 571	58, 908	89.6
"introverted"	1, 694	91, 359	177

Designing and Analyzing the Traits. Different from the traditional text categorization, it becomes more complex when facing the problem of categorizing personality traits [10]. Usually there is no direct connection between the bloggers' traits and the topics of their blogs. Hence, it is insufficient to analyze bloggers' personality traits merely from the topic of the contents in the blogs. From the aspects of blogs' topics, blogs' linguistic traits and behavior traits, this paper analyses the distinction of these traits from different people and finally finds the most appropriate traits which can be used to categorize bloggers' personality traits.

In this paper, the characteristic words, which is the real words in Chinese linguistics, comprises nouns, verbs, adjectives, numerals, quantifiers, pronouns, words of orientation, words of place and words of time. The aim of using the characteristic words is to verify whether they can also be used to categorize personality traits. In addition, the words describing the same topic are classified into the same word category and from these word categories we can analysis the distinction from different people with different personality traits. According to HOWNET [3], we can obtain these word categories.

In this paper, the style trait is independent from the contents of articles, including function words, sentiment words, Internet glossaries, punctuations and N-tuple word sequences. Function words are made up of adverbs, prepositions, conjunctions, auxiliary words, interjections, mimetic words, modal particles, prefixes, suffixes, state words and distinguishing words. Sentiment words are the words which can express bloggers' feelings, such as "happiness" and "sadness". Internet glossaries are the popular words and expressions on the Internet and they can be gained from "the approach of detecting unlisted words and expressions" in the Chapter "Algorithm Design", like "Gei Li"("awesome" in Chinese) and "Shen Ma"("everything" or "what"). The N-tuple word sequence means a sequence with N consecutive words in an article, which can reflect bloggers' writing styles and behaviors to some degree. Bloggers' personality traits categorization [10] says that it is much more effective than characteristic words and it is more appropriate for the categorizing tasks irrelevant to the topics and contents so that it is introduced into this paper.

Algorithm Design

The Extraction of Text Features. Combined with the features of blog contents, this paper brings out an approach of detecting unlisted words and expressions. The matching relationships between bloggers and blog tags are shown in Fig. 1.

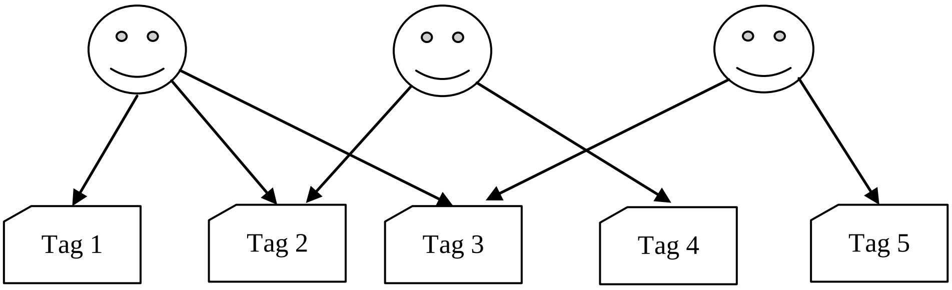
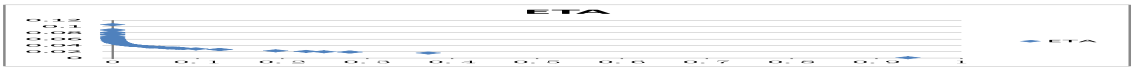


Fig.1. The matching relationships between bloggers and blog tags

As an important component of blog contents, tags usually contain topic words and expressions of blogs. Unlisted words and expressions are included as well. Tags have abundant information of articles and they are widely used to categorize or cluster blogs. Therefore, we use the tag-statistical approach to detect unlisted words and expressions and its steps are shown as follows.

(1) Collect the set of tags and count the times of the tags used in a blog.

(2) According to the formulas Eq. 1 and Eq. 2:



(1)

(2)

The tag_i means the reliability of unlisted words or expressions and the $\text{Count}(\text{tag}_i)$ means the overall times a blogger uses the tag. The $P_j(\text{tag}_i)$ represents the probability of the j-th blogger using the tag. This formula shows that the more times a tag appears and are utilized by several bloggers, the more possibility that the tag is an unlisted word or expression.

(3) Order all the tags by their reliabilities, reserve the tags with higher reliability and consider these tags as unlisted words or expressions.

Traits Selection. Similar with the traditional text categorization, the basic idea of traits selection is selecting the traits can not only represent this type of samples but also distinguish this type of samples from other types, which means the pertinence of traits and types and the apparent difference among different types. This aims to reduce the size of the traits set and improve the accuracy of categorizing.

There are several approaches of traits selection, like the information gain (IG). Japanese [11] has been applied into this approach and it has obtained an ideal result of categorization using only 200-dimension traits. So at first we use this approach directly to finish the traits selection and the experimental result shows that the accuracy of categorization is higher than the baseline. However, when doing traits selection with IG, it can introduce some irrelevant traits, especially when the selected trait set has been enlarged, the number of irrelevant traits will be increased which can lead to the loss of categorizing accuracy and the over-fitting. Fig.2 shows the pertinence of traits and personalities when doing traits selection and the difference between the trait and other people with different personalities. The horizontal axis means the difference and the vertical axis means the pertinence and they are based on ETA pertinent coefficients and ANOVA variance analysis. When the x-coordinate of a trait is greater than 0.05, it says that there is no apparent distinction between the "extroverted" and "introverted" people so that it cannot tell the difference between the "extroverted" and "introverted" bloggers. When the y-coordinate of a trait is less than 0.02, it says that there is little pertinent between the trait and someone's personality so that it cannot represent bloggers' personality traits. All the traits according with the two situations mentioned above are not suitable to categorize personalities. They can not only increase the expenditure of calculation but also decline the accuracy of categorization. From Fig. 2, we can find that when the number of traits is increased from 100 to 300, IG will introduce more irrelevant traits so that it is not suitable to be used into categorizing personalities in this paper.

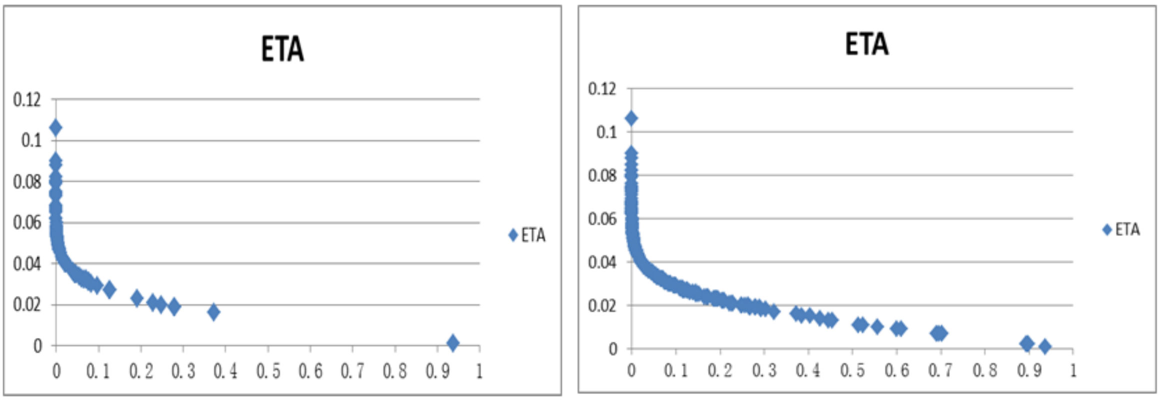


Fig.2. The pertinence of top-200 and top-2000 traits and personalities and the difference between the trait and other people with different personalities when using IG to do traits selection.

Because the categorization of personality traits deems a blogger's all blogs, rather than just one blog, as a sample, these traits occur in most of the samples and the frequencies of them differ from each other. However, IG only considers whether the traits have occurred without any concentration on their frequencies. Therefore, IG cannot screen out the best traits effectively.

Because of the weakness of IG, this paper proposes the approach of traits selection based on ETA pertinent coefficients and ANOVA variance analysis. This approach takes the frequencies of traits into consideration so that it is more efficient than IG.

The Calculation of the Trait Weights. In the text categorization, the classical formula of the trait weights uses TFIDF, which can achieve an ideal result. However, no ideal results are shown in our experiment of categorizing personality traits because this formula does not normalize the lengths of the samples. We propose a new approach of calculating the trait weights. In the approach, the trait weight is the ratio of the frequency of appearance of the traits to the overall length of the blogs in the sample. In addition, in order to eliminate the effects from the sizes of the weights, all the weights are mapped to the interval of [0, 1] and the mapping formula is shown in Eq. 3.

$$v' = \frac{v - \mathrm{Min}}{\mathrm{Max} - \mathrm{Min}} \tag{3}$$

In this formula, Max and Min represent the maximum and minimum of an attribute in all the samples and *v* represents the original value of the attribute.

Categorization of Personality Traits

The blogs' behavior traits include hyperlinks, pictures, comments, frequencies of blogging, average lengths of blogs, numbers of friends, numbers of emoticons and so forth. The statistical data, using the means and the standard error, from the "extroverted" and "introverted" samples are shown in Table 2.

Table 2 .The comparison of blog behaviors from the "extroverted" and "introverted" bloggers

Blog Behaviors	Extroverted	Introverted
The average number of words from the blogs	509.64±19.83	**595.20±18.28**
The average length of the blogs	1788.25±56.30	**2037.79±54.47**
The number of blogs released per month	1.12±0.043	**1.84±0.089**
The length of time blog accounts been registered	**34.93±0.37**	33.58±0.37
The number of friends	**122.18±17.24**	98.91±9.78
The average number of comments of the blogs	**3.982±0.191**	3.488±0.194
The average number of emoticons of the blogs	**0.3255±0.0249**	0.2111±0.0175
The average number of hyperlinks of the blogs	1.1445±0.1806	**1.6862±0.1184**
The average number of pictures of the blogs	2.4633±0.1334	**2.4929±0.1145**

From Table 2 we can draw the conclusion that extroverted bloggers' average number of words and average length of the blogs are much larger than the introverted ones' and their average number of blogs released per month are larger as well. There are two major reasons. On the one hand, introverted bloggers does not good at keeping in touch with others directly and they are willing to noting down everything happened in their lives and releasing it in their blogs. On the other hand, introverted bloggers have fewer friends and it is hard for them to seek and communicate with someone who is suitable. The latter reason can be verified by the average number of bloggers' friends listed in Table 2. Extroverted bloggers usually registered their blog accounts earlier than introverted ones, for they are sensitive to something in fashion. Also, extroverted bloggers enjoy leaving some comments to communicate and interact with other bloggers. They are much more active and they enjoy using emoticons but they use hyperlinks and pictures less.

We have divided all the data into 10 equal parts randomly and taken one part in turn as a test sample as well as the rest as training samples to finish the experiment of categorizing personality traits and discover the effects of the categorizing results from different trait sets [12,13]. The classifier utilizes SVM and the kernel functions utilize the linear kernel. The accuracy of categorization is shown in Table 3.

Table 3. The accuracy of categorizing personality traits (The baseline results from the calculation of the ratio of big word categories, which is 51.9%)

Traits	Baseline	IG	ANOVA ETA(>=0.04)	ANOVA ETA(>=0.05)	ANOVA ETA(>=0.06)
Real words	51.9%	61.3%	63.4%	**64.0%**	63.2%
Function words	51.9%	58.5%	**62.5%**	61.7%	60.6%
Sentiment words	51.9%	59.3%	**64.3%**	62.6%	58.4%
Word categories	51.9%	**59.8%**	59.2%	59.7%	59.5%
Behavior traits	51.9%	52.4%	**52.6%**	52.3%	51.9%
1-POS	51.9%	60.3%	60.1%	60.2%	**60.7%**
2-POS	51.9%	59.9%	**64.1%**	61.1%	60.5%
3-POS	51.9%	60.4%	**72.4%**	68.8%	64.2%

From Table 3 we can conclude that, for all the trait sets, IG and the new approach of traits selection mentioned in the paper have a much higher accuracy than the baseline and the accuracy of the new approach is higher than IG as well. In all the sets, the trait of 3-POS has the best categorization result and its accuracy is 72.4%. The following traits are the sentiment words and 2-POS. Nevertheless, different from the researches abroad, function words are not as effective as real ones, maybe due to the distinction between Chinese and English. The number of behavior traits is 10 in all and the algorithm in this paper has improved the accuracy of the categorization apparently.

Conclusion

On the basis of Big Five, this paper shows a new approach about categorizing bloggers' personality traits, including “extroverted” and “introverted” personalities. Compared with the traditional text categorization, it is more difficult when categorizing personality traits. This paper mainly focuses on how different trait sets and the approach of trait selection can affect the result of categorization and proposes the new approach of traits selection to improve the accuracy of categorizing personality traits dramatically. Compared with the past researches, the data can be unprepared and the tags can be unattached in advance so that the expenditure can be cut down dramatically. In addition, this paper exploits all the bloggers' blogs to categorize their personalities rather than only one blog from each blogger so that we can obtain more information from the bloggers in the data. Nowadays, few researches about categorizing Chinese bloggers' personality traits have been done, so we can provide other researches about categorizing Chinese bloggers' personality traits with references.

Acknowledgement

This work is supported by National Natural Science Foundation of China (61170193, 61370185), Guangdong Natural Science Foundation (S2012010009865, S2012020011081, S2013010013432, S2013010015940), Science and Technology Planning Project of Guangdong Province

(2011B090400041, 2012B010100039, 2012B040305011, 2012B010100040), Guangdong ”Twelfth Five-Year Plan” education information technology research projects(11JXZ012), Guangdong Higher Education Discipline and Profession special fund projects(2013KJCX0174) and Science and Technology Planning Project of Huizhou City (2011P002, 2011g012, 2011P005, 2011P003, 2011g011).

References

[1] Yafeng Lu; Feng Wang; Maciejewski, R., Business Intelligence from Social Media: A Study from the VAST Box Office Challenge, Computer Graphics and Applications, IEEE, Year: 2014, Volume: 34, Issue: 5,Pages: 58 - 69

[2] Guangxia Li; Hoi, S.C.H.; Kuiyu Chang; Wenting Liu; Jain, R., Collaborative Online Multitask Learning, Knowledge and Data Engineering, IEEE Transactions on, Year: 2014, Volume: 26, Issue: 8,Pages: 1866 - 1876

[3] Changbo Wang; Zhao Xiao; Yuhua Liu; Yanru Xu; Aoying Zhou; Kang Zhang, SentiView: Sentiment Analysis and Visualization for Internet Popular Topics, Human-Machine Systems, IEEE Transactions on, Year: 2013, Volume: 43, Issue: 6, Pages: 620 - 630

[4] Ajitha, P.; Gunasekaran, G., Sentiment prediction based on valence and arousal using concept search engine, Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference on,Year: 2015,Pages: 1 - 5

[5] Jeyapriya, A.; Kanimozhi Selvi, C.S., Extracting aspects and mining opinions in product reviews using supervised learning algorithm, Electronics and Communication Systems (ICECS), 2015 2nd International Conference on, Year: 2015,Pages: 548 - 552

[6] Liu, Bin; Zhang, Jingyuan; Liu, Qiang; Li, Han; Zhang, Mingliang; Qiu, Rui; Zhao, Jingyang, Data Acquisition, Hot Issues and System of Microblog Mining,Network and Information Systems for Computers (ICNISC), 2015 International Conference on, Year: 2015,Pages: 116 - 119

[7] Spasojevic, Nemanja; Rao, Adithya, Identifying actionable messages on social media,Big Data (Big Data), 2015 IEEE International Conference on,Year: 2015,Pages: 2273 - 2281

[8] Yan Yan; Subramanian, R.; Ricci, E.; Lanz, O.; Sebe, N.,Evaluating Multi-task Learning for Multi-view Head-Pose Classification in Interactive Environments, Pattern Recognition (ICPR), 2014 22nd International Conference on,Year: 2014,Pages: 4182 - 4187

[9] Scott Nowson，Jon Oberlander. Identifying more bloggers：Towards large scale personality classification of personal weblogs. ICWSM’2007 Boulder, Colorado, USA. 2007

[10] Shlomo Argamon, Sushant Dhawle, Moshe Koppel, James W. Pennebaker. Lexical predictors of personality type. In Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America. 2005.

[11] Atsunori Minamikawa, Hiroyuki Yokoyama.Blog Tells What Kind of Personality You Have: Egogram Estimation from Japanese Weblog. CSCW 2011. Pages: 217 - 220

[12] Francisco Iacobelli, Alastair J. Gill, Scott Nowson, and Jon Oberlander. Large Scale Personality Classification of Bloggers. Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II, Pages: 568 - 577

[13] Alastair J. Gill，Scott Nowson，Jon Oberlander. What Are They Blogging About? Personality, Topic and Motivation in Blogs. Proceedings of the Third International ICWSM Conference. 2009. Pages: 18 - 25.