KSS'2013 NINGBO

Proceedings

Knowledge Creation Towards Emergency Management

Shouyang Wang, Yoshiteru Nakamori and Weiliang Jin (eds.)

JAIST Press

ISBN: 978-4-903092-36-2

Exploration on Posts of Tianya Club and Preliminary Results

Jindong Chen **Xijin Tang**

Institute of Systems Science, Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing, 100190 P.R. China
{j.chen,xjtang}@amss.ac.cn

Abstract

To explore the tacit knowledge of posts of Tianya Club, an attempt is made to identify the risk categories of posts on Tianya Club with support vector machine (SVM). The workflow of posts classification using SVM is provided at first. Then, similarities analysis of posts on Tianya Club and multi-class classification of posts using SVM with different training set are implemented. The similarity analysis reveals the difficulty in multi-class classification of posts on Tianya; and the predictive results indicate that SVM could be applied to posts multi-class classification, but still need further exploitation.

Keywords: text classification, SVM, multi-class classification, posts, Tianya Club

1 General instructions

Tianya Zatan board is one of most popular and influential board of Tianya Club, which is a famous Internet forum in China, and provides BBS, blog, micro-blog and photo album services etc. [1]. The posts on Tianya Zatan cover the hot and sensitive topics of society. Analyzing the posts is a good means to monitor the status of social risk. Posts classification plays an important role in the analyzing work, but this mission is impossible to be handled only by human. Text classification is to automatically assign predefined categories to free text documents [2], and posts classification is a branch of text classification, hence the methods of text classification could be tested to automatically identifying risk categories of the posts on Tianya Zatan.

Text classification utilizes learning methods to assign predefined categories labels to new documents based on the likelihood suggested by a trained set of labels and documents [3]. Generally, two main procedures affect the accuracy of text classification: text representation and classifier construction. On the one hand, text representation includes feature extraction and feature selection [4], feature extraction is to present the text documents into clear word format, and feature selection is to select a subset of feature from the original documents through some strategies, such as term frequency inverse document frequency (TF*IDF) [5], information gain (IG), term frequency etc. On the other hand, classifier construction is to build classifier through machine learning methods with training samples. Many research works have been presented on machine learning and their effectiveness in text classification field. The machine learning methods which can be divided into three classes: supervised, unsupervised and semi-supervised, while supervised methods have shown advantages in text classification. The representative supervised machine learning methods for text classification are neural network [6], support vector machine (SVM) [3, 7], etc.

However, posts classification is a bit different from previous text classification, although each board defined the scope and topics, the new posts usually contain many new and different contents, the length of posts varied dramatically, and the distribution of posts in different categories are unbalanced. Furthermore, the training samples are required to train the classifier, but the training samples are limited, because the categories of posts were labeled by human beings, but the amount of posts in one day is always much more than 400, so it is uneasy to label all the posts, hence we only obtained two months posts labeled: December 2011 and January 2012. All these factors hindered the development of classifier for posts on Tianya Club.

Therefore, in posts classification area, no similar research work has been presented on this topic. This is a new area of text classification, and no mature strategy can deal with this issue. Hence, following previous mature strategies dealt with text classification, a preliminary result of posts

categorization on Tianya Zatan has been present in this paper.

The rest of this paper is organized as follows. Section 2 presents the procedure of web documents representation and classifier construction; the feature extraction including word segment and stop word elimination, the feature selection method is TF*IDF, and the machine learning method is SVM. The results of similarities analysis and text categorization are presented in Section 3. Finally, conclusion and further research plan are given in Section 4.

2 The process of posts classification using SVM

The process of posts classification using SVM is presented in Figure 1. As mentioned before, feature extraction, feature selection is the first part of posts classification. After the feature selection, similarity analysis is provided to show the difficulty in posts classification. The introduction of SVM method and category membership score method is followed in this section.

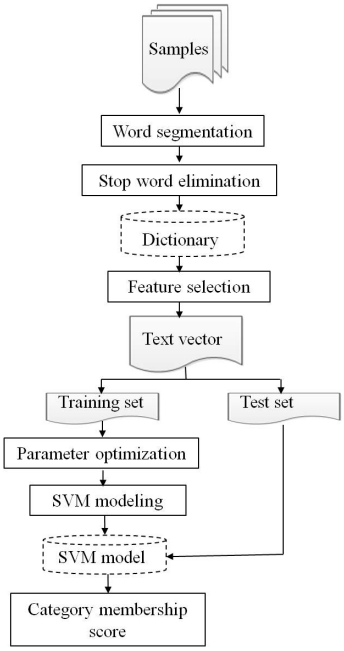

Figure 1. The process of posts classification using SVM

2.1 Feature extraction

In Figure 2, it can be found that feature extraction is the first step of posts classification, and including three parts: i) term segmentation, plain text is segmented into Chinese terms by MMSeg [8]; ii) stop words elimination, stop words form HIT (Harbin Institute of Technology) are applied [9], which contains 767 functional words in Chinese; iii) the remained terms constitute the initial structure of dictionary.

2.2 Feature selection

The feature selection is the second step of posts classification, which is to assign different weights to terms and generate the dictionary of salient terms. TF*IDF is evolved from IDF which is proposed by Sparck Jones with heuristic intuition that a query term which occurs in many documents is not a good discriminator [10], and should be given less weight than one which occurs in few documents. Eq.1 is the classical formula of TF*IDF used for feature selection,

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \qquad (1)$$

where $w_{i,j}$ is the weight for term i in post j, N is the number of posts in the collection, $tf_{i,j}$ is the term frequency of term i in post j and df_i is the post frequency of term i in the collection.

2.3 Similarity analysis

The similarity analysis is an important part of posts classification, which provides evidence whether post could be classified and how difficulty of this work. Up to now, there are many methods proposed for similarity analysis, such as cosine similarity function, Jaccard coefficient and Dice coefficient.

The cosine similarity function (CSF) is the most widely reported measure of vector similarity. The virtue of the CSF is its sensitivity to the relative importance of each word [11]. Through an example to illustrate CSF, assume:

n=number of unique words in the dictionary

$X = (x_1, \cdots x_n)$

where $x_i = \begin{cases} 1 & \text{if word } i \text{ is in the post } \mathrm{P}_x \\ 0 & \text{if word } i \text{ is not in the post } \mathrm{P}_x \end{cases}$

$Y = (y_1, \cdots y_n)$

where $y_i = \begin{cases} 1 & \text{if word } i \text{ is in the post } \mathrm{P}_y \\ 0 & \text{if word } i \text{ is not in the post } \mathrm{P}_y \end{cases}$

X and Y are defined as binary vector representations of the P_x and P_y respectively, denoting the

presence or absence of a word in either post. Given these definitions, therefore the CSF similarity function we implemented is

$$\mathrm{CSF} = \frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \quad (2)$$

2.4 Support vector machine

SVM is a relatively new learning approach introduced by Vapnik for solving two-class pattern recognition problem [12]. The method is originally defined over a vector space where the problem is to find a decision surface that "best" separates the data into two classes. For linearly separable space, the decision surface is a hyper plane which can be written as

$$\omega x + b = 0 \quad (3)$$

where x is an arbitrary objects to be classified; the vector ω and constant b are learned from a training set of linearly separable objects. SVM is equivalent to solve a linearly constrained quadratic programming problem as Eq.4; hence the solution of SVM is globally optimal.

$$\min_{\omega} \frac{1}{2}\|\omega\|^2 + C\sum_i \xi_i \quad (4)$$

With constraints

$$y_i(x_i\omega + b) \geq 1 - \xi_i \quad \xi_i \geq 0, \ \forall i \quad (5)$$

where ξ_i is non-negative slack variables.

For the linearly inseparable problem, kernel function [13] is used to derive the similarities in the original lower dimensional space.

Due to the good performance of SVM in binary classification, it has been expanded into multi-class area. Considering the multi-class classification issue in this field, the One-Against-One approach is adopted. Other methods for multi-class classification are also discussed such as error-correcting output codes, SVM decision tree, etc. [14].

2.5 Category membership score

After feature selection of text vectors, samples are fed into SVM training process. In SVM training, the sample unbalance is among separate categories of samples. Moreover, categories of words at different time and different contexts are varied dramatically. Therefore, classification accuracy of SVM will be disturbed, and a category membership score is applied to enhance the classification accuracy. The category membership score is computed by Eq.6.

$$score = \frac{\sum S_i}{2 * k} + \frac{k}{2 * n} \quad (6)$$

where k is the number of voters supporting a certain category; n is the number of categories; S_i is the score of each supporting voter. As multi-class classification problem in SVM can be treated as multiple binary classification problem, and C_2^n voters as classifiers in bipartition is computed. The rule of the category membership score is: as to one test sample, the bigger the score of voter as the first item in Eq.6 and the more the supporting voters as the second item, the more convinced the judgment that the sample belongs to this category is decided. With category membership score, membership score of classification result is under the chosen best-fit threshold is ignored.

3 Results and discussions

This section includes three parts: i) statistics analysis of samples; ii) the results of similarity analysis; iii) the results of multi-class classification using different training set.

3.1 Statistics of the sample data

Due to more than 10 thousands new posts per month on Tianya Zatan, and the category of post is labeled manually which is a tough job, hence we only obtained two months labeled posts at present. The number of new posts of separate risk category for these two months is presented on Table 1.

From the Table 1, it can be found that post distribution of 8 categories is unbalanced. The new posts on Tianya Zatan mainly concentrate on government management, public morals and daily life. The unbalanced distribution of samples will decrease the accuracy of classifier, because it is hard for classifier to learn the feature of the category with fewer samples.

Table 1. The number of new posts of each category for December 2011and January 2012

Categories	December 2011	January 2012
Risk free	1282(10.6%)	2046(17.0%)
Government management	3372(27.8%)	1809(15.0%)
Public morals	3336(27.5%)	3730(31.0%)
Social stability	953(7.9%)	1013(8.4%)
Daily life	2641(21.8%)	3063(25.5%)
Recourses & environment	222(1.8%)	147(1.2%)
Economy & finance	247(2.0%)	133(1.1%)
National security	71(0.6%)	91(0.8%)
Total	**12124**	**12032**

Table 2. Similarity analysis in same category

Categories	Minimum	Maximum	Average	Standard error
Government management	0	0.564	0.031	0.003
Public morals	0	0.645	0.026	0.002
Social stability	0	0.678	0.037	0.003
Daily life	0	0.859	0.045	0.006
Recourses & environment	0	0.668	0.057	0.007
Economy & finance	0	0.800	0.041	0.005
National security	0	0.790	0.091	0.013

3.2 The results of similarity analysis

After feature extraction and feature selection, similarity analysis is carried out. The main object of this part is to show how difficulty of the posts on Tianya Zatan to be classified, and the time factor whether need to be considered in classification. Three kinds of similarity are analyzed in this part: i) the similarities of posts in same category; ii) the similarities of posts between different categories; iii) the similarities between different days.

Due to the big volume of posts in some categories, at present it is hard to calculate the similarities of all posts. Therefore, only the posts of December 2012 is considered in this section, and 100 posts are randomly selected from each category, if the number of post in one category is less than 100, the maximum number is used in this analysis. Eq.2 is used for similarity calculation, the minimum, maximum, average and standard error of similarity in same category were calculated, and results are presented in Table 2.

From the results of Table 2, it can be found that the similarities of posts in same category are less than 0.1, which means the similarities in same category are low, and the standard errors of categories keep small. From all these results, it can be said that classification of posts on Tianyan Zatan is difficult.

If argued that why such low similarity could be classified into one category, here we present some examples to explain this issue. As mentioned before, all the posts can be classified into 8 categories: one risk free category and seven risk categories. Seven risk categories usually include several types of risk respectively, and the difference between the content of these risks are great. Zheng, Shi and Li constructed a framework of societal risk indicators including 7 categories and 30 sub categories based on word association tests [15], and 2 qualitative meta-synthesis supporting technologies: CorMap and iView, were applied to help grouping the associated words into clusters and detect the main hazards [16], Table 3 lists the societal risk resulted from that study. To explain why posts in the same category shared such low similarity, two types of posts in the same sub category with different content are presented in Table 3. Through Table 3, we try to bring out more examples to illustrate the reason of posts shared such low similarities in the same category.

Table 3. The examples of posts shared low similarity in same category

Categories	Sub categories	Type 1	Type 2
Government management	Corruption & degeneration	"Chen Liangyu" corruption case	"Guo Meimei" event
	Governance ability	Violent demolition	Medically unqualified in civil service examination
	Legal system	Governance on-line rumors	Adjustment of the policies of real estate
	Social security & social welfare	The retirement policy	Minimum living standard
Public morals	Ethics & morality	Extramarital affair	Filial piety
	Faith & reputation	"Han Han" and "Fang Zhouzi" event	"Liu Xiang" 2012 Olympic Games
	General mood of society	Nanjing "Peng Yu" case	Wasting food culture
Daily life	Health		
	Education		
	Employment		
	Prices		
	Transportation		
	Food and medicine safety		
	Housing		
	Fake & shoddy goods		
Social stability	Serious epidemics		
	Poor-rich Gap		
	Safety at work	Aircraft accident	Coal mine tragedy
	Crimes & mass incidents		
	Issue concerning agriculture, farmer and rural area	Low price of vegetable hurts farmers	Urbanization construction
Economy & finance	Economy problems		
	Finance problems		
Recourses & environment	Natural disaster		
	Population	Migration of farmers	Family planning issue
	Energy shortage & environment pollution		
National security	Terrorism & cults		
	Taiwan Issue		
	Political stability	1989 Tiananmen Square Event	Cultural revolution
	National security and foreign relations		
	Very important events	2008 Beijing Olympic Games	2010 Shanghai EXPO

Table 4. Similarity analysis between different categories

Class	Public morals	Social stability	Daily life	Recourses & environment	Economy & finance	National security
Government management	0.0138	0.0167	0.0148	0.0103	0.0100	0.0084
Public morals		0.0159	0.0204	0.0122	0.0131	0.0127
Social stability			0.0207	0.0118	0.0130	0.0109
Daily life				0.0196	0.0140	0.0269
Recourses & environment					0.0092	0.0112
Economy & finance						0.0102

Table 5. Similarity analysis during December 10-17, 2011

Day	Dec.11	Dec.12	Dec.13	Dec.14	Dec.15	Dec.16	Dec.17
Dec.10	0.858	0.9015	0.8662	0.8814	0.8816	0.8807	0.8950
Dec.11		0.8627	0.8618	0.8548	0.8489	0.852	0.8681
Dec.12			0.881	0.8969	0.8998	0.8904	0.8994
Dec.13				0.8975	0.8809	0.8567	0.889
Dec.14					0.9011	0.8667	0.9009
Dec.15						0.8837	0.8934
Dec.16							0.8750

In Table 3, for sub category with clear area border, no confused example is present to illustrate the difficulty. It can be found that in one risk category, there are several sub categories; and in the same sub category, their contents can also be quite different, such as in corruption & denegation sub category: "Chen Liangyu" corruption case and "Guo Meimei" event, the posts of "Chen Liangyu" corruption mainly discuss the amount of money he embezzled and the justice of judgment; "Guo Meimei" event is related to the corruption of Red Cross, but many posts are gossip news of Guo Meimei. Therefore, these cases could explain why posts in the same category shared low similarity.

For different categories similarity analysis, the posts of December 2012 is considered in this section, only 100 posts are selected from each category, if the number of posts in one category is less than 100, all posts of that category are adopted. The similarities of posts between two categories are calculated first, Eq.2 is used for similarity calculation; and then the average values of all the similarities are computed, the results are presented in Table 4.

From the results of Table 4, the similarities between different categories are generally much lower than the similarities in same category, which is good news to our classification research; it also provides the evidence that posts on Tianya Zatan could be classified. However, the difference is still unobvious, the classification of posts on Tianya Zatan board is still uneasy.

To consider the influence of time factor, the posts in the period of December 10-17, 2011 are selected. All posts in one day were collected together to one vector; it is used to calculate the similarity of different days. The results are presented in Table 5.

From the results of Table 5, it can be found that the similarities between different days almost exceed 0.85, which means the variability of posts on Tianya Zatan in one week is unobvious, the posts mainly concentrated on several topics. Hence, the classifier of Tianya Zatan is unnecessary to be updated every day.

3.3 The results of classification

Based on the similarities analysis above, the classification research is tried in this part. Two experiments are designed in this part: i) the training set is 2 days' data; ii) the training set is 31 days' data.

Parameters setting: the kernel of SVM is radial basis kernel, other parameters of SVM are optimized online.

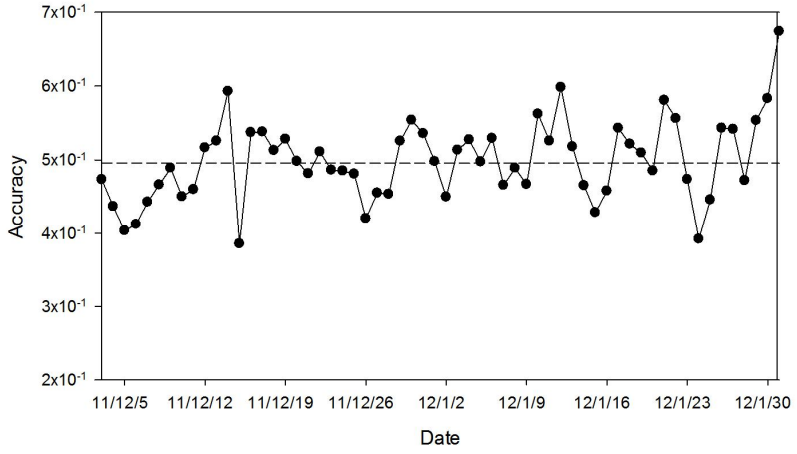

Figure 2. The predictive results of SVM based on 2 days' data

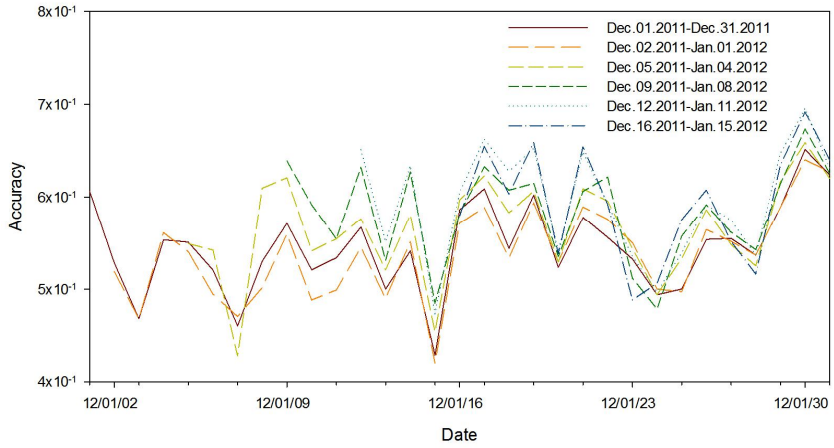

Figure 3. The predictive results of SVM based on 31 days' data

To measure the performance of SVM classification, a standard definition of accuracy is as shown in Eq.7 in this research.

$$Accuarcy = \frac{sum_{lsvm=lmanual}}{sum_{smaple}} \quad (7)$$

where $sum_{lsvm=lmanual}$ is the set of those posts that SVM outputs as manual label, sum_{smaple} is the set of posts in the test samples. The predictive results are presented in Figure 2 and Figure 3.

In Figure 2, it is shown that the predictive results of SVM using samples of two days before is unacceptable, the average predictive accuracy is less than 50%. Because SVM has shown excellent performance in many fields [3, 7], the predictive accuracy here is much lower than the results in other fields. However, as the similarities analysis presented above, the classification of posts on Tianya Zatan is more difficult than all those research mentioned above.

To improve the performance of SVM, the 31 days' data as training set is selected. As in Figure 3, it is presented that the predictive performance is improved, and the average predictive results are almost 60%. It also can be found that influence of time factor, along with the training set moving, the predictive accuracy is also increased. Furthermore, the predictive results of training set of whole December of 2011 are close to the training set of December 02. 2011- January 01. 2012; the predictive results of the training set of December 05. 2011- January 04. 2012 outputs similar results as the training set of whole December of 2011 in the first several days, and better results on December 08 and December 09, so the improvements are unclear; the predictive results of the training set of December 09. 2011- January 08. 2012 show obvious improvements than the training set of whole December of 2011. Hence, from these results, it can be said that the classifier can keep almost one week, and other simulations show similar results.

4. Conclusions

In this paper, we mainly analyze the similarities of posts in different categories on Tianya Zatan, and then we follow the process of SVM to text classification to automatically identify risk categories of posts. Two experiments with different training set are conducted. The results show that the training set with previous 31 days provides better performance, but the results are still unsatisfied.

Therefore, further study needs to be done. As the results shown, only depending on SVM, the predictive results cannot satisfy the practical requirement, even if with bigger training set. Hence, to decrease the burden of labeling by man power, multi-level method will be considered in this research: as it is found that the dictionary of each category is stable, a dictionary for each category could be built based on this case, and the similarity analysis could also be conducted as first level of classification, and then SVM classifier or other machine learning will be applied to posts classification.

Acknowledgment

This research is supported by National Basic Research Program of China under Grant No. 2010CB731405 and National Natural Science Foundation of China under Grant No.71171187.

References

[1] http://en.wikipedia.org/wiki/Tianya_Club

[2] Zhang W, Tang XJ, Yoshida T. Text classification with support vector machine and back propagation neural network. *In Computational Science* (ICCS 2007 proceedings, Part IV,Y. Shi, et al eds.), Lecture Notes in Computer Science, 4490, Springer-Verlag, 2007,150-157.

[3] Zhang W, Yoshida T, Tang XJ. Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, 2008, 21(8): 879-886.

[4] Baharudin B, Lee LH, Khan K. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 2010, 1(1): 4-20.

[5] Zhang W, Yoshida T, Tang XJ. A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, 2011, 38(3): 2758-2765.

[6] Ruiz ME, Srinivasan P. Hierarchical text categorization using neural networks. *Information Retrieval*, 2002, 5(1): 87-118.

[7] Hu Y, Tang XJ. Using support vector machine for classification of Baidu hot word. *In Knowledge Science, Engineering and Management* (KSEM2013, Dalian, China. M. Wang, et al eds.). Lecture Notes in Computer Science, 8041, Springer Berlin Heidelberg, 2013, 580-590.

[8] Tsai CH. MMSEG: A word identification system for mandarin Chinese text based on two variants of the maximum matching algorithm. (2000-03-12). http://technology.chtsai.org/mmseg.

[9] http://ir.hit.edu.cn/bbs/viewthread.php?tid=20

[10] Jones KS. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972, 28(1): 11-21.

[11] Salton G. Developments in automatic text retrieval. *Science*, 1991, 253(5023): 974-980.

[12] Vapnik V. The nature of statistical learning theory. New York: Springer, 2000.

[13] Aizerman M, Braverman E, Rozonoer L. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 1964, 25: 821-837.

[14] Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2002, 2: 265-292.

[15] Zheng R, Shi K and Li S. The influence factors and mechanism of societal risk perception. *Proceedings of the First International Conference on Complex Sciences: Theory and Application* (Shanghai, China, J. Zhou eds.). Springer Berlin Heidelberg, 2009: 2266-2275.

[16] Tang XJ. Qualitative meta-synthesis techniques for analysis of public opinions for in-depth study. *Proceedings of the First International Conference on Complex Sciences: Theory and Application* (Shanghai, China, J. Zhou eds.). Springer Berlin Heidelberg, 2009: 2266-2275.