How to Submit Proof Corrections Using Adobe Reader

Using Adobe Reader is the easiest way to submit your proposed amendments for your IGI Global proof. If you don't have Adobe Reader, you can download it for free at http://get.adobe.com/reader/. The comment functionality makes it simple for you, the contributor, to mark up the PDF. It also makes it simple for the IGI Global staff to understand exactly what you are requesting to ensure the most flawless end result possible.

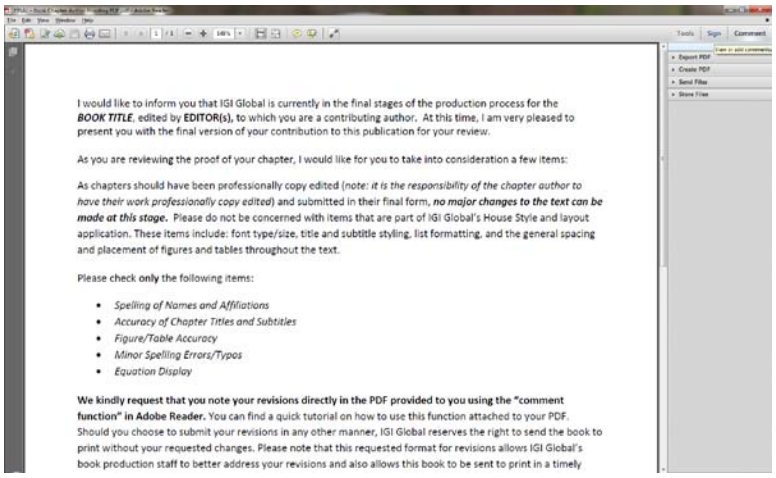
Please note, however, that at this point in the process the only things you should be checking for are:

Spelling of Names and Affiliations, Accuracy of Chapter Titles and Subtitles, Figure/Table Accuracy, Minor Spelling Errors/Typos, Equation Display

As chapters should have been professionally copy edited and submitted in their final form, please remember that ***no major changes to the text can be made at this stage*.**

Here is a quick step-by-step guide on using the comment functionality in Adobe Reader to submit your changes.

1. Select the **Comment** bar at the top of page to View or Add Comments. This will open the **Annotations** toolbar.



I would like to inform you that IGI Global is currently in the final stages of the production process for the ***BOOK TITLE***, edited by **EDITOR(s)**, to which you are a contributing author. At this time, I am very pleased to present you with the final version of your contribution to this publication for your review.

As you are reviewing the proof of your chapter, I would like for you to take into consideration a few items:

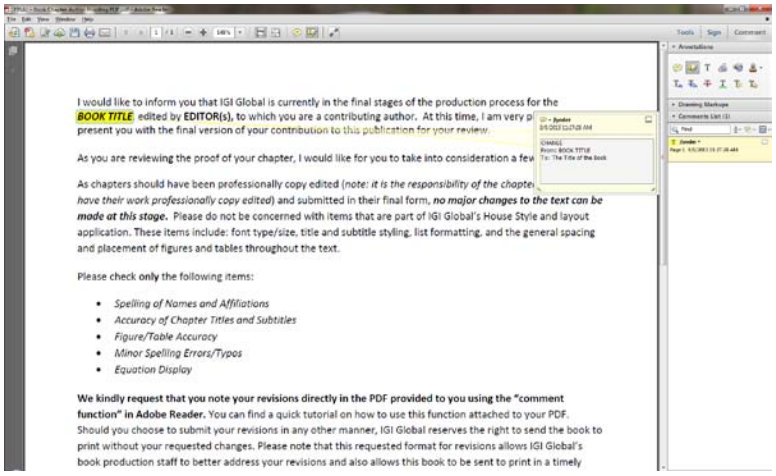
As chapters should have been professionally copy edited (*note: it is the responsibility of the chapter author to have their work professionally copy edited*) and submitted in their final form, ***no major changes to the text can be made at this stage.*** Please do not be concerned with items that are part of IGI Global's House Style and layout application. These items include: font type/size, title and subtitle styling, list formatting, and the general spacing and placement of figures and tables throughout the text.

Please check **only** the following items:

- *Spelling of Names and Affiliations*
- *Accuracy of Chapter Titles and Subtitles*
- *Figure/Table Accuracy*
- *Minor Spelling Errors/Typos*
- *Equation Display*

We kindly request that you note your revisions directly in the PDF provided to you using the "comment function" in Adobe Reader. You can find a quick tutorial on how to use this function attached to your PDF. Should you choose to submit your revisions in any other manner, IGI Global reserves the right to send the book to print without your requested changes. Please note that this requested format for revisions allows IGI Global's book production staff to better address your revisions and also allows this book to be sent to print in a timely

2. To note text that needs to be altered, like a subtitle or your affiliation, you may use the **Highlight Text** tool. Once the text is highlighted, right-click on the highlighted text and add your comment. Please be specific, and include what the text currently says and what you would like it to be changed to.



I would like to inform you that IGI Global is currently in the final stages of the production process for the ***BOOK TITLE***, edited by **EDITOR(s)**, to which you are a contributing author. At this time, I am very pl present you with the final version of your contribution to this publication for your review.

As you are reviewing the proof of your chapter, I would like for you to take into consideration a few

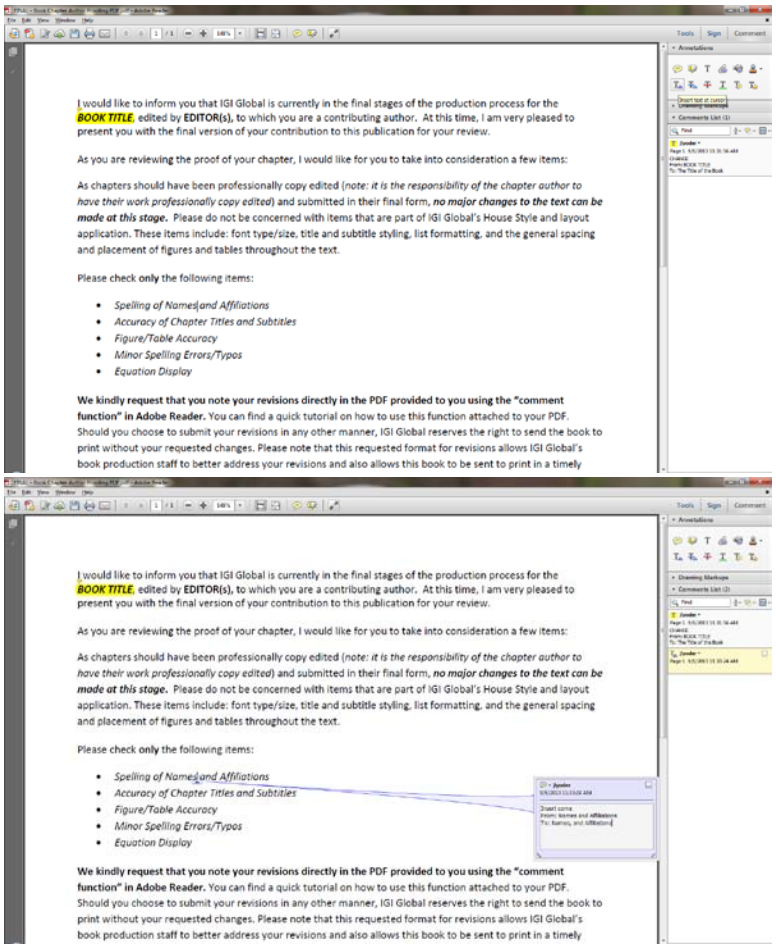
As chapters should have been professionally copy edited (*note: it is the responsibility of the chapte have their work professionally copy edited*) and submitted in their final form, ***no major changes to the text can be made at this stage.*** Please do not be concerned with items that are part of IGI Global's House Style and layout application. These items include: font type/size, title and subtitle styling, list formatting, and the general spacing and placement of figures and tables throughout the text.

Please check **only** the following items:

- *Spelling of Names and Affiliations*
- *Accuracy of Chapter Titles and Subtitles*
- *Figure/Table Accuracy*
- *Minor Spelling Errors/Typos*
- *Equation Display*

We kindly request that you note your revisions directly in the PDF provided to you using the "comment function" in Adobe Reader. You can find a quick tutorial on how to use this function attached to your PDF. Should you choose to submit your revisions in any other manner, IGI Global reserves the right to send the book to print without your requested changes. Please note that this requested format for revisions allows IGI Global's book production staff to better address your revisions and also allows this book to be sent to print in a timely

3. If you would like text inserted, like a missing coma or punctuation mark, please use the **Insert Text at Cursor** tool. Please make sure to include exactly what you want inserted in the comment box.



I would like to inform you that IGI Global is currently in the final stages of the production process for the ***BOOK TITLE***, edited by **EDITOR(s),** to which you are a contributing author. At this time, I am very pleased to present you with the final version of your contribution to this publication for your review.

As you are reviewing the proof of your chapter, I would like for you to take into consideration a few items:

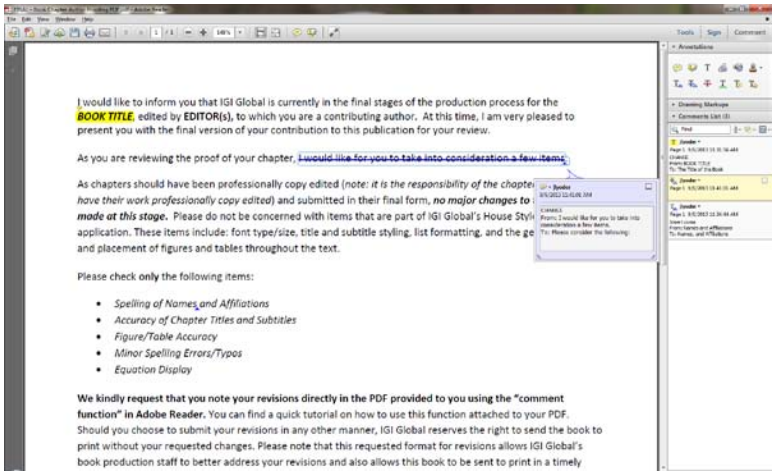
As chapters should have been professionally copy edited (*note: it is the responsibility of the chapter author to have their work professionally copy edited*) and submitted in their final form, ***no major changes to the text can be made at this stage.*** Please do not be concerned with items that are part of IGI Global's House Style and layout application. These items include: font type/size, title and subtitle styling, list formatting, and the general spacing and placement of figures and tables throughout the text.

Please check **only** the following items:

- *Spelling of Names and Affiliations*
- *Accuracy of Chapter Titles and Subtitles*
- *Figure/Table Accuracy*
- *Minor Spelling Errors/Typos*
- *Equation Display*

We kindly request that you note your revisions directly in the PDF provided to you using the "comment function" in Adobe Reader. You can find a quick tutorial on how to use this function attached to your PDF. Should you choose to submit your revisions in any other manner, IGI Global reserves the right to send the book to print without your requested changes. Please note that this requested format for revisions allows IGI Global's book production staff to better address your revisions and also allows this book to be sent to print in a timely

4. If you would like text removed, such as an erroneous duplicate word or punctuation mark, please use the **Add Note to Replace Text** tool and state specifically what you would like removed.



I would like to inform you that IGI Global is currently in the final stages of the production process for the ***BOOK TITLE***, edited by **EDITOR(s),** to which you are a contributing author. At this time, I am very pleased to present you with the final version of your contribution to this publication for your review.

As you are reviewing the proof of your chapter, I would like for you to take into consideration a few items:

As chapters should have been professionally copy edited (*note: it is the responsibility of the chapter author to have their work professionally copy edited*) and submitted in their final form, ***no major changes to the text can be made at this stage.*** Please do not be concerned with items that are part of IGI Global's House Style and layout application. These items include: font type/size, title and subtitle styling, list formatting, and the general spacing and placement of figures and tables throughout the text.

Please check **only** the following items:

- *Spelling of Names and Affiliations*
- *Accuracy of Chapter Titles and Subtitles*
- *Figure/Table Accuracy*
- *Minor Spelling Errors/Typos*
- *Equation Display*

We kindly request that you note your revisions directly in the PDF provided to you using the "comment function" in Adobe Reader. You can find a quick tutorial on how to use this function attached to your PDF. Should you choose to submit your revisions in any other manner, IGI Global reserves the right to send the book to print without your requested changes. Please note that this requested format for revisions allows IGI Global's book production staff to better address your revisions and also allows this book to be sent to print in a timely

IGI PUBLISHING
WWW.IGI-GLOBAL.COM

International Journal of Knowledge and Systems Science

January-March 2014, Vol. 5, No. 1

Table of Contents

Research Articles

The *International Journal of Knowledge and Systems Science* is indexed or listed in the following: ACM Digital Library; Bacon's Media Directory; Cabell's Directories; DBLP; Google Scholar; INSPEC; JournalTOCs; Library & Information Science Abstracts (LISA); MediaFinder; The Standard Periodical Directory; Ulrich's Periodicals Directory

Exploring Societal Risk Classification of the Posts of Tianya Club

Jindong Chen, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, P.R. China

Xijin Tang, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, P.R. China

ABSTRACT

To identify the societal risk category of the posts of Tianya Club, several studies are carried out toward the posts of Tianya Club. With 2-month manually risk labeled new posts published during December of 2011 to January of 2012, statistical analysis of posts is conducted at first. Later, similarity analysis of posts from one risk category, different risk categories and published on different days are implemented. Finally, multi-class classification of posts using support vector machine (SVM) with different training set is tested. The statistical analysis and similarity analysis reveals the difficulties in multi-class classification of the posts of Tianya Club. The multi-class predictive results indicate that SVM could be applied to multi-class classification of posts, but still need further exploitation.

Keywords: Multi-Class Classification, Posts, Similarity Analysis, Statistical Analysis, Tianya Club

INTRODUCTION

"Tianya Zatan board is one of the most popular and influential board of Tianya Club, which is a famous Internet forum in China, and provides BBS, blogs, micro-blogs and photo album services etc.. [1]" The posts of Tianya Zatan board cover the hot and sensitive topics of society. Analyzing the posts is a good means to monitor the status of societal risk (Tang, 2013). From previous studies (Tang, 2013), it is shown that the risk intensity of each category is varying, hence risk classification of posts plays an important role in the analyzing work, but this mission is impossible to be handled only by humans. As it can be found, the contents of the posts of Tianya Club are mainly textual information; only a minority of posts is attached with pictures or other media information, then text classification is the first choice to classify the posts of Tianya Zatan board.

However, as it is found that risk classification of posts has several unique features from text classification, such as the dynamical variation of context, the limitation of training samples, the bad quality of corpuses, etc., which

DOI: 10.4018/ijkss.2014010104

make risk classification of posts more difficult than standard text classification discussed intensively by professionals. Furthermore, the difficulties confront in risk classification of posts, which will hinder progress in on-line societal risk perception research (Tang, 2013). Up to date, in the area of risk classification of posts, 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 effectively. Hence, following normal strategies applied to deal with text classification, risk classification of posts of Tianya Zatan board has been tested in this paper.

The basic principle of text classification is utilizing learning strategies to assign predefined categories labels to new documents based on the likelihood suggested by a trained set of labels and documents (Zhang, et al., 2007; Zhang, et al., 2008). Generally, two main procedures affect the accuracy of text classification: text representation and classifier construction. Text representation includes feature word extraction and feature word selection (Baharudin, et al., 2010), feature word extraction is to transfer the text documents into clear word vector; feature word selection is to select a subset of feature words from the original documents through some methods, such as term frequency inverse document frequency (TF*IDF) (Zhang, et al., 2011), information gain (IG), term frequency, etc.. Classifier construction is to build classifier through machine learning strategies using training samples. Many research works have been done on machine learning and their effectiveness in text classification field. The machine learning strategies which can be divided into three classes: supervised, unsupervised and semi-supervised (Huang, et al., 2006), while supervised methods have shown advantages in text classification. The representative supervised machine learning methods for text classification are neural network (Ruiz & Srinivasan, 2002), support vector machine (SVM) (Hu & Tang, 2013; Tong & Koller, 2002; Zhang, et al., 2008), etc.. In brief, mature procedure and strategies have been set up in text classification, and normal steps will be adopted in risk classification of posts.

As mentioned above, risk classification of posts is much more difficult than previous text classification; people may argue that the classification mission is impossible. Therefore, before classifier construction, to show and confirm the difficulties of risk classification of posts, similarity analysis of posts in one risk category are carried out; to describe the feasibility of posts classification, similarity analysis of posts between two risk categories are implemented; to consider the effect of time factor, similarity analysis of posts between published on different days are conducted.

The rest of this paper is organized as follows. Section 2 presents the procedure of web documents representation, similarity analysis and classifier construction. The results of statistical analysis, similarity analysis and text categorization are presented in Section 3. Finally, conclusion and further research plan are given in Section 4.

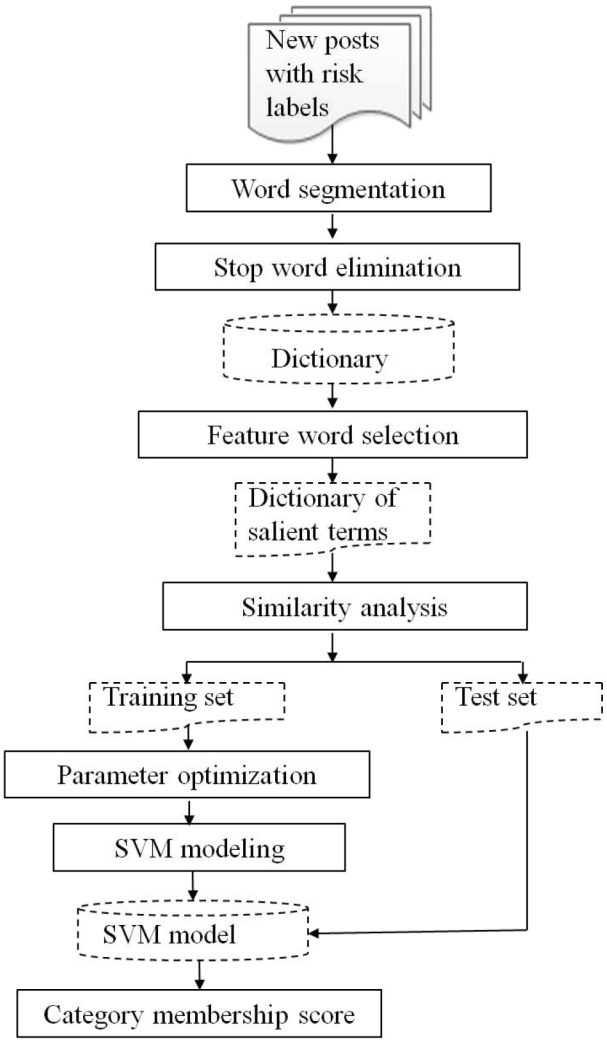
THE PROCESS OF SOCIETAL RISK CLASSIFICATION OF POSTS

The process of societal risk classification of posts is described in Figure 1. Feature word extraction, feature word selection is the first part of posts classification. After the feature word selection, the explanation the method of similarity analysis is presented in this section. And then the introduction of SVM strategy and category membership score method is followed.

Feature Word Extraction

In Figure 1, it can be found that feature word extraction is the first step of risk classification of posts, and including three parts: i) term segmentation, plain text is segmented into Chinese terms by ICTCLAS (Zhang, et al., 2003); ii) stop words elimination, stop words form HIT (Harbin Institute of Technology) are applied[2], which contains 767 functional words

Figure 1. The process of risk classification of posts using SVM



in Chinese; iii) the remained terms constitute the initial dictionary.

Feature Word Selection

The feature word selection is the second step of risk classification of posts, 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 Jones (1972) with heuristic intuition that a query term which occurs in many documents is not a good discriminator, and should be given less weight than one which occurs in few documents. Equation (1) is the classical formula of TF*IDF used for feature word selection:

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

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.

Similarity Analysis

Up to now, there are many methods proposed for similarity analysis, such as cosine similarity function, Jaccard coefficient and Dice coefficient, etc.. 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 (Salton, 1991). Through an example to illustrate CSF, assume: 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 each post, the CSF similarity function implemented is:

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

where n = number of unique words in the dictionary:

$$X = (x_1, \cdots x_n)$$

where:

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

$$Y = (y_1, \cdots y_n)$$

where:

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

Support Vector Machine

SVM is a relatively new learning approach introduced by Vapnik (2000) for solving two-class pattern recognition problem. The strategy 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:

$$\omega x + b = 0 \tag{3}$$

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 Equation (4); hence the solution of SVM is globally optimal:

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

$$s.t.\ y_i(x_i\omega + b) \geq 1 - \xi_i \quad \xi_i \geq 0, \forall i \tag{5}$$

where C is the penalty coefficient, ξ_i is non-negative slack variables.

For the linearly inseparable problem, kernel function (Aizerman, et al., 1964) 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. Many methods for multi-class classification of SVM are discussed before, such as error-correcting output codes, SVM decision tree, etc. (Crammer & Singer, 2002). Considering the multi-class classification issue in this field, the One-Against-One approach is adopted.

Category Membership Score

After feature word selection of text vectors, samples are fed into SVM training process. In

Table 1. The distribution of new posts in each risk category published in December 2011 and January 2012

Risk Categories	New Posts in December 2011	Ration	New Posts in January 2012	Ration
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	**12,125**	**100%**	**12,032**	**100%**

SVM training, the unbalance of the sample is among separate categories of samples. Moreover, the words of risk categories at different time and contexts vary 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 Equation (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 Equation (6) and the more the supporting voters as the second item, the more convinced that the sample belongs to this category. With category membership score, if membership score of classification result is under the chosen best-fit threshold, the classification result will be ignored.

RESULTS AND DISCUSSIONS

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

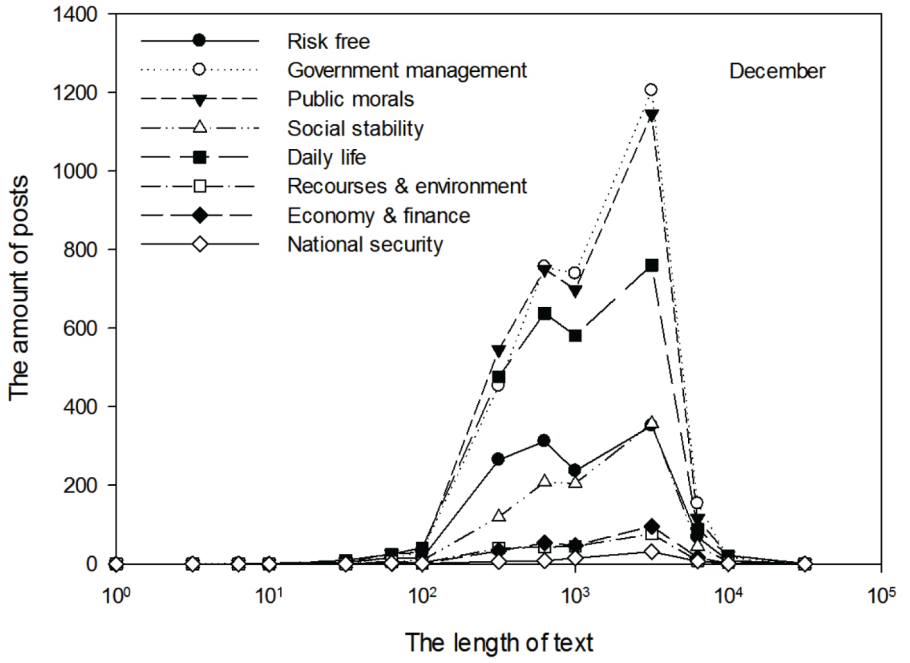
Statistical Properties of the Posts

The 2-month manually risk labeled posts are applied for statistical analysis. To get a general view of this 2-month data set, two characteristics of the posts are extracted: i) the distribution of the new posts in each risk category; ii) the distribution of the length of new posts in each risk category.

The results of the distribution of new posts in December 2011 and January 2012 are presented in Table 1.

From the Table 1, it is shown that distribution of 8 categories is extremely unbalanced. The new posts on Tianya Zatan board mainly concentrate on government management, public morals and daily life, totally more than 75%. The unbalanced distribution of samples will

Figure 2. The distribution of the length of new posts in December 2011



affect the accuracy of classifier, because it is hard for the classifier to learn the feature of the category with fewer samples.

The results of the distribution of the length of new posts in December 2011 and January 2012 are presented in Figure 2 and Figure 3.

In Figure 2, it is described that the distributions of the length of new posts in 8 categories are similar. The length of new posts is varied dramatically in each risk category, from 10 to $3*10^4$; and the length of new posts on Tianya Zatan board mainly concentrates on $3*10^2$ to $3*10^3$. The length distribution of new posts will affect the accuracy of classifier, because short posts may not provide enough information for classification. The similar results are obtained toward new posts published in January 2012.

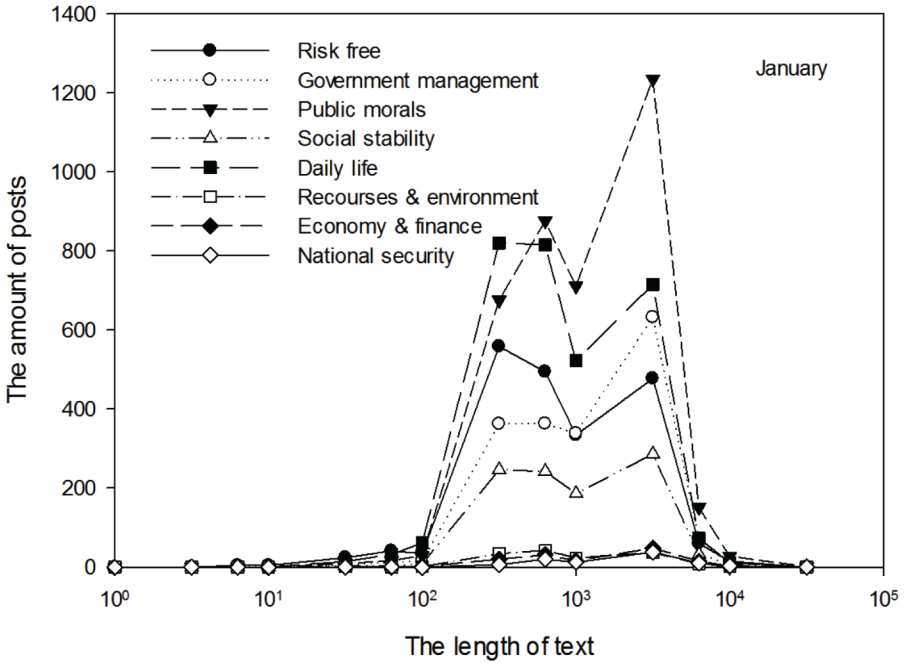
The Results of Similarity Analysis

After feature word extraction and feature word selection, similarity analysis is carried out. The main objective of this part is to show how difficult of the posts on Tianya Zatan board to be classified, and the time factor whether needs to be considered in classification. Three kinds of similarity are analyzed here: i) the similarities of new posts within one risk category; ii) the similarities of new posts between two risk categories; iii) the similarities between new posts published on different days.

For illustration, only the new posts of December 2012 are considered in this section, and 100 posts are randomly selected from each risk category. If the amount of post in one category is less than 100, the maximum number is used in this analysis. Equation (2) is used for similarity calculation, the maximum, average and standard error of similarity in one risk category are calculated with results presented in Table 2.

From Table 2, it can be found that the similarities of new posts in one risk category are all less than 0.1, which means the similarities in one risk category are low, while the standard errors of categories keep small. From all these results, we may say that risk classification of new posts on Tianya Zatan board is difficult.

Figure 3. The distribution of the length of new posts in January 2012



If argued that why such low similarity could be classified into one risk category, here we will explain the reasons to this issue. As mentioned before, all the posts can be classified into 8 categories: one risk free category and seven risk categories, which is proposed by Zheng, et al. (2009), they constructed a framework of societal risk indicators including seven categories and 30 sub categories based on word association tests, 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 (Tang, 2009), hence the difference between the content of these seven risk categories are great. Table 3 lists the societal risk resulted from that study. To further explain why posts in the one risk category share such low similarity, two types of posts in the same sub risk category are presented in Table 3, because the risk cat-

Table 2. Similarity analysis of new posts within one risk category

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

egory is composed of sub risk categories, low similarity in sub risk category will lead to low similarity in that category. Through Table 3, we try to bring out more examples to illustrate the reason of posts shared such low similarities in the same sub risk category.

In Table 3, for sub risk category with clear definition, no example is presented to illustrate the low similarity in these sub risk categories, such as all sub risk categories in daily life and economy & finance, and some sub risk categories in social stability, resources & environment and national security, but this does not mean that these sub risk categories should share high similarity, because many different events are contained in these sub risk categories, which only means the events or posts in these sub risk categories would be clear.

However, for sub risk category without clear definition in Table 3, two examples are presented to show the big difference of posts in one sub risk categories. In sub category of corruption & denegation: "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 trial; "Guo Meimei" event is related to the corruption of Red Cross, but many posts are gossip news of Guo Meimei. In governance ability, the posts on violent demolition mainly reveal these events and ask for help, and the posts of medically unqualified in civil service examination mainly complain the unfair policy and require for justice. In general mood of society, the posts of Nanjing "Peng Yu" case mainly argue who cheat in that case, and the posts of wasting food culture mainly blame the people who waste food and ask people to save food; and so many other examples in other sub risk categories. Therefore, from these cases, it is sensed that why posts in the one risk category shared low similarity.

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

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

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

The Results of Classification

Based on the similarity analysis in Table 5, 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 setting by online optimization. To measure the performance of SVM classification, a standard definition of accuracy is as shown in Equation (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 4 and Figure 5.

In Figure 4, 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

Table 3. Illustrations of posts with low similarity in the one risk category

Risk Categories	Sub Risk 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	Corruption of public morals
	Integrity & 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 major events	2008 Beijing Olympic Games	2010 Shanghai EXPO

Table 4. Similarity analysis between different risk categories

Risk Category	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

shown excellent performance in many fields (Hu & Tang, 2013; Zhang, et al., 2008), the predictive accuracy here is much lower than the results in other fields. However, as the similarity analysis presented above, the classification of posts on Tianya Zatan board 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 5, 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 2. 2011- January 1. 2012; the predictive results of the training set of December 5. 2011- January 4. 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 9, so the improvements are unclear; the predictive results of the training set of December 9. 2011- January 8. 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.

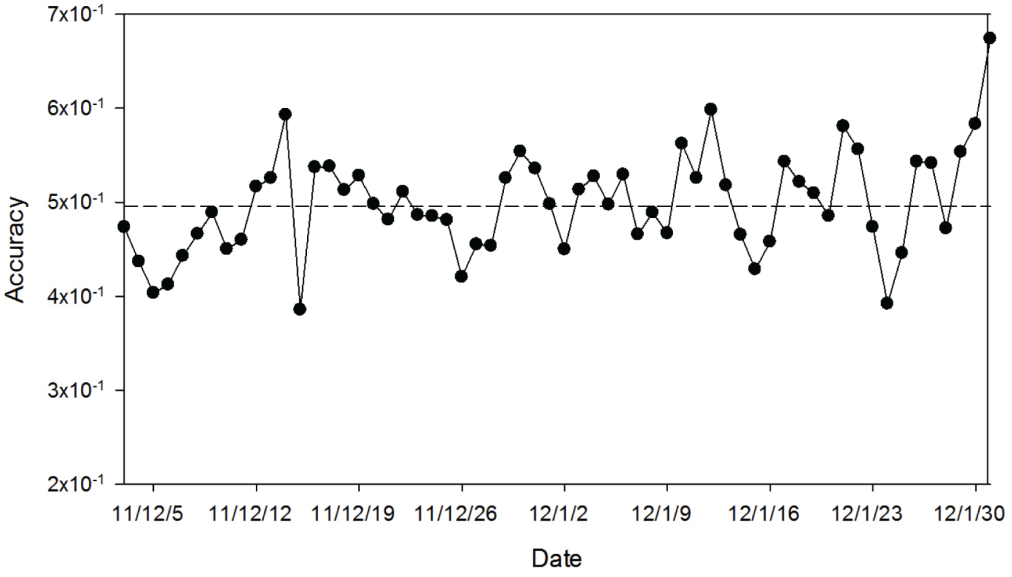
CONCLUSION

In this paper, we mainly analyze the statistical properties and similarities of posts in different risk categories on Tianya Zatan board, and then we follow the process of SVM to text classifica-

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

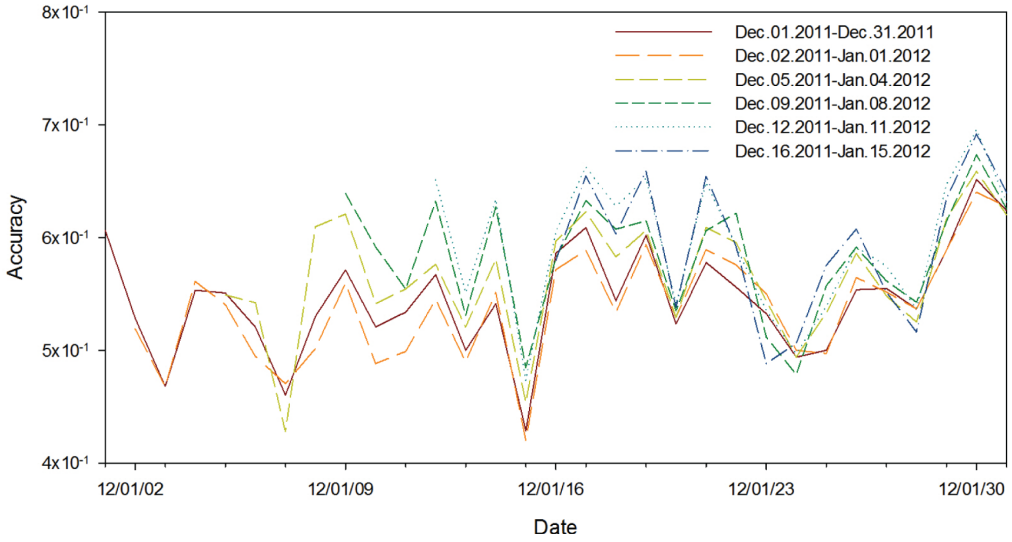
Figure 4. The predictive results of SVM based on 2 days' data



tion 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

Figure 5. The predictive results of SVM based on 31 days' data



this research: as it is found that the dictionary of each risk category is stable, a dictionary for each risk 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. The authors would like to thank Mr. Yongliang Zhao for his data collection work, Ms. Lina Cao for her data processing work, and the other members of our team. The original version of this paper was presented at the 14th International Symposium on Knowledge and Systems Sciences, Ningbo, October 25-27, 2013.

REFERENCES

Aizerman, A., Braverman, E. M., & Rozoner, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, *25*, 821–837.

Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*, *1*(1), 4–20. doi:10.4304/jait.1.1.4-20

Crammer, K., & Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, *2*, 265–292.

Hu, Y., & Tang, X. (2013). Using support vector machine for classification of Baidu Hot word. In M. Wang (Ed.), *Knowledge science, engineering and management* (Vol. 8041, pp. 580–590). Springer Berlin Heidelberg. doi:10.1007/978-3-642-39787-5_49

Huang, T.-M., Kecman, V., & Kopriva, I. (2006). *Kernel based algorithms for mining huge data sets: Supervised, semi-supervised, and unsupervised learning* (Vol. 17). Berlin, Germany: Springer.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *The Journal of Documentation*, *28*(1), 11–21. doi:10.1108/eb026526

Ruiz, M. E., & Srinivasan, P. (2002). Hierarchical text categorization using neural networks. *Information Retrieval*, *5*(1), 87–118. doi:10.1023/A:1012782908347

Salton, G. (1991). Developments in automatic text retrieval. *Science*, *253*(5023), 974–980. doi:10.1126/science.253.5023.974 PMID:17775340

Tang, X. (2009). *Qualitative meta-synthesis techniques for analysis of public opinions for in-depth study. Complex Sciences* (pp. 2338–2353). Springer.

Tang, X. (2013). Applying search words and BBS posts to societal risk perception and harmonious society measurement. In *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics.*

Tong, S., & Koller, D. (2002). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, *2*, 45–66.

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer. doi:10.1007/978-1-4757-3264-1

Zhang, H., Yu, H., Xiong, D., & Liu, Q. (2003). HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing* (Vol. 17).

Zhang, W., Tang, X., & Yoshida, T. (2007). Text classification with support vector machine and back propagation neural network. In Y. Shi, G. Albada, J. Dongarra & P. A. Sloot (Eds.), *Computational science (ICCS 2007)* (Vol. 4490, pp. 150-157). Springer Berlin Heidelberg.

Zhang, W., Yoshida, T., & Tang, X. (2008). Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, *21*(8), 879–886. doi:10.1016/j.knosys.2008.03.044

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758–2765. doi:10.1016/j.eswa.2010.08.066

Zheng, R., Shi, K., & Li, S. (2009). The influence factors and mechanism of societal risk perception. In J. Zhou (Ed.), *Complex sciences* (Vol. 5, pp. 2266–2275). Springer Berlin Heidelberg. doi:10.1007/978-3-642-02469-6_104

ENDNOTES

1 http://en.wikipedia.org/wiki/Tianya_Club
2 http://www.datatang.com/data/13281