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, blog, microblog and photo album services etc..” 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 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 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
Related Content

Recurrence Indicators for the Estimation of Characteristic Size and Frequency of Spatial Patterns
[www.igi-global.com/chapter/recurrence-indicators-estimation-characteristic-size/69463?camid=4v1a](www.igi-global.com/chapter/recurrence-indicators-estimation-characteristic-size/69463?camid=4v1a)

A Review on Evolutionary Prototype Selection: An Empirical Study of Performance and Efficiency
[www.igi-global.com/chapter/review-evolutionary-prototype-selection/38452?camid=4v1a](www.igi-global.com/chapter/review-evolutionary-prototype-selection/38452?camid=4v1a)
Optimized E-Government User Support Allocation and Its Influence on Citizens’ Adoption of E-Government: An Agent Based Approach
www.igi-global.com/article/optimized-government-user-support-allocation/80417?camid=4v1a

Ontology Alignment: State of the Art, Main Trends and Open Issues
Tatyana Ivanova (2012). *Systems Approaches to Knowledge Management, Transfer, and Resource Development* (pp. 147-167).
www.igi-global.com/chapter/ontology-alignment-state-art-main/68216?camid=4v1a