Chinese Blog Classification Based on Text Classification and Multi-feature Integration

Jianzhuo Yan
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
yanjianzhuo@bjut.edu.cn

Suhua Yang, Liying Fang
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
yangsuhua86@126.com, fangliying@bjut.edu.cn

***Abstract*—The Chinese blog has become one of the most important sources of information in China. The content of Chinese blog varies widely, thus its classification is of great significance. The Chinese blog has the features of the title, straight matter, tags and user-defined types, and different features have different lengths. Traditional text classification method of the Chinese blog classification is not ideal. In this paper, the Chinese blog is classified by using a number of Chinese blog features in which traditional text classification technique and short text classification technique will be chosen according to the different length of features. In addition, the feature expansion method is adopted for sparse features of short text, and the features are integrated by linear training. Experimental results show that the proposed method improves the accuracy of classification.**

Keywords- text classification; Chinese blog classification; short text classification; feature expansion; multi-feature integration

I. INTRODUCTION

Chinese blog has become more and more popular in China. In recent years, with the rapid development of Chinese blog, the domains of scientific research and industry have been interested in Chinese blog. If we can make full use of the abundant Chinese blog resources and classify the Chinese blog correctly, it is of great practical and scientific significance to learn the development of internet, improve various internet services and enrich user's internet lives [1].

Chinese blog classification is the core and basis in personalized services. As only blogger's personalized information is well understood, the ideal of personalized services may be achieved. The Chinese blog which is different from the general text has the features of title, straight matter, tags and user-defined types. At present, there has been some related research in the Chinese blog classification; AiXin Sun points out that using tag for Chinese blog classification can improve the classification results [2]. The classification method is only for the whole blog not for the articles of the blog, so the classification granularity of this method is not detailed enough. Singh et al. [3] proposed a method of blog classification by combining the domain ontology, which improves the deficiencies of the traditional "word bag" model in the expression of semantic information. But that method does not combine the blog features, so the blog information can not be expressed well enough. Lin and Nenghai [4] proposed the classification of the multi-feature integration, but the method is lack of the analysis for the blog features. Obviously, the traditional method is not proper for each feature. The text content of title, tags and user-defined types is relatively short, so the description has weak signals. Only digging out more information for the short text, short text classification can be more correctly. Hyponymy relation between words is an important semantic relation, and extending short text feature vector by using the hyponymy relation between words can make the short text information richer.

Chinese blog has some features which have different lengths. For this reason, this paper uses the method of multi-feature integration for the different feature. In the classification process, the long feature and short feature use the traditional text classification technique and the short text classification technique, separately. Thus, the contents of various Chinese blog features can be fully expressed, and integrating the features through linear training. This subject has been widely applied in personalized search, advertisements automatically recommend, the construction of user community, and so on.

II. STATE OF THE ART

For the blog classification, the introduction section shows that many experts classify the Chinese blog by using traditional text classification method, but the result of classification is not very well. The blog has its features [5], and if the features are used in the classification, the result of classification will be more correct. Some papers use the Naive Bayes [6] and k-Nearest Neighbor algorithm (KNN) [7] for the classification. The model of Naive Bayes is a probability classification model based upon two assumptions. It requires that the probabilities of all words are independent and the class of the document has no correlation with its length, but the effect is unstable in practical application. KNN is a method based on lazy and required learning method, and the effect of classification is better. But the time of classification is nonlinear, and when the number of training text increases, the time of classification will sharply increase. Support Vector Machine (SVM) is a new machine learning method which is advanced by Vapnik [8] based on

 ISBN: 978-1-61208-202-8

statistical learning theory. It is similar to structure risk minimization principle [9], which has splendidly learning ability and only needs few samples for training a high-performance text classifier.

In this paper, SVM is used for text classification according to the blog features, and experiment results show that the classification is more credibility.

III. CHINESE BLOG CLASSIFICATION

The traditional model of Chinese blog classification includes two modules: pre-processing module and classification module. In this paper, the multi-feature fusion algorithm is added into the classification algorithm, so the module of feature integration is added into the classification module which combines the classification results of each feature. In the classification module, the traditional text classification technique and the short text classification technique are used according to the length of the feature. In addition, the algorithm of feature vector extension is adopted for the classification.

The framework of Chinese blog classification proposed in this paper is shown in Figure 1.

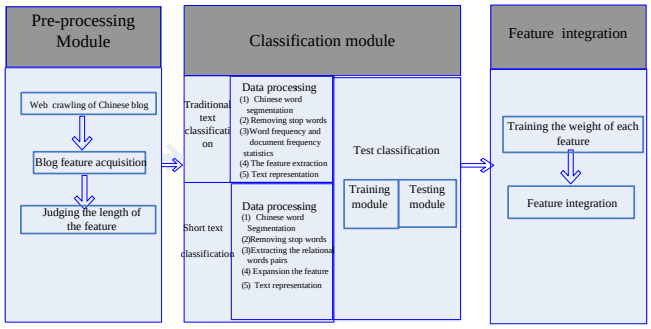

Figure 1. The framework of Chinese blog classification.

From Figure 1, the aims of the three modules are clear. The first module is to get the feature of the Chinese blog and judge the length of the feature, the second module is to classify the text for each feature of the Chinese blog, and the third module is to integrate the feature.

A. Pre-processing Module

Chinese blog pages are written in Hypertext Markup Language (HTML) which contains a wealth of information, and are semi-structured text files. In addition to plain text, the page also contains some labels and features. Before classification, the features of the straight matter, title, tags and user-defined types should be obtained. Then the content of the features are as regular texts, and the blog can be classified as the texts.

Specific steps include web crawling of Chinese blog, blog feature acquisition and judging the length of the feature, which are as follows:

1) Web crawling of Chinese blog: It aims to get the source code of the Chinese blog.

2) Blog feature acquisition: By using regular expressions to remove the label of source code of the Chinese blog, the features of the straight matter, title, tags and user-defined types are extracted.

3) Judging the length of the feature: The text which has less than 160 characters in length is considered as short text. The title, tags and user-defined types usually belong to the short texts, and the straight matter of the blog which usually has more than 160 characters is taken as the traditional text classified by traditional classification methods.

B. Traditional Text Classification

After text pre-processing, the feature which is judged as the traditional text is to execute data processing. Further data processing is used for further classification. The data processing includes Chinese word segmentation, removing stop words, word frequency and document frequency statistics, the feature extraction and text representation.

1) Data processing

The specific steps of data processing are as follows:

a) Chinese word segmentation: The Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS) is used for words segmentation.

b) Removing stop words: Create the list of terms which are filtered before the word frequency process started. The list includes mainly conjunctions, prepositions or pronouns.

c) Word frequency and document frequency statistics: Count word frequency for each word which appears in the text. The word frequency F is initialized as 1, and added 1 each time to count the document frequency of each category.

d) The feature extraction: By delete the words from the text which has no contribute or very little contribution to the entry category information and taking account of the large amount of information carried by nouns, verbs followed by adjectives and adverbs, this article frame is realized using only the noun.

e) Text representation: In this paper, the term frequency-inverse document frequency (TF-IDF) algorithm [10] is used as vector space model to represent the text. The method of vector space model representation is as follows: each Chinese blog text is represented as a n-dimensional vector (w_1, w_2, $w3$, ... , w_n), and the weight of each dimension in the vector of this text should correspond with the weight in this text.

Weight Set: $W = \{w_{i \mid i \in n}\}$

$$w_i = \frac{\sum_{i \in s} (w_i \times tf_i) \times \log(N / n_i)}{\sqrt{\sum_j ((\sum_{i \in n} w_i \times tf_i) \times \log(N / n_i))^2}} \quad (1)$$

where w_i is the corresponding weight of the i-key words, tf_i is the frequency of the i-key words in the page, N is the total number of text contained in the training set, n_i is the number of the text which contains the characteristics.

2) Text classification

The technique of text classification is mainly based on statistical theory and machine learning, such as Naive Bayes, KNN and SVM. The model of Naive Bayes is a probability classification model based upon two assumptions. It requires that the probabilities of all words are independent and the class of the document has no correlation with its length, but the effect is unstable in practical application. KNN is a method based on lazy and required learning method. The effect of classification is better, but the time of classification is nonlinear, and when the number of training text increases, the time of classification will sharp increase. SVM is a new machine learning method advanced by Vapnik according to statistical learning theory. It is similar to structure risk minimization principle, which has splendidly learning ability and only needs few samples for training a high-performance text classifier [11]. The input vector X is mapped to a high-dimensional feature space Z by nonlinear mapping, in which the optimal separating hyperplane is structured. SVM classification function is similar to neural network in form. The output is a linear combination of intermediate nodes, each intermediate node corresponds to a support vector, and the dot product is operated between vectors. The expression of the SVM function for classification of non-linear optimal separating surface is as follow:

$$f(z)=\sum_{supvector} a_i^* y_i \varphi(z)\varphi(z)+b^*=\sum_{supvector} a_i^* y_i k(z,z)+b^* \quad (2)$$

Therefore, adopting the kernel function can avoid the high-dimensional feature space for complex operations. The process can be expressed as follows: First, map the input vector X into a high-dimensional Hilbert space H. The kernel function has different forms, and different kernel functions will form different algorithms. In general, the commonly used kernel function has three kinds: Polynomial kernel function, Radial basis function, and Neural network kernel function.

The choice of kernel function has little effect on the accuracy of classification. But polynomial classifier can be applied for the low-dimensional, high-dimensional, large sample, small sample and so on. It is applicable, and has a wider domain of convergence, parameter easy to control, etc. Thus, this paper chooses polynomial classifier as a kernel function.

C. Short Text Classification

Traditional text classification method can not be applied in the short text classification very well. The correlation between words and categories is measured when extracting the feature, but the short text has less number of words and weak information, which leads to a serious shortage of short text feature and makes the traditional classifier not accurately classify the text [12]. In this paper, the feature expansion method is used to rich the content of the short text.

The data processing of short text includes Chinese word segmentation, removing stop words, the extraction of relational words pairs, the feature extraction, feature expansion and text representation. The methods of Chinese word segmentation, removing stop words, text representation and the feature extraction are the same as the methods of data processing for the traditional text classification, and the different processes are the extraction of relational words pairs and feature expansion.

1) Extracting the collection of the feature words pairs by HowNet

HowNet is an on-line common-sense knowledge based on unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. HowNet uses the knowledge representation language to describe the concept, and the words of the knowledge representation language are called as “Sememe” which is the smallest unit of the concept [13].

In this paper, the hyponymy strength of the relational words (A, B) is measured by the semantic distance of sememe.

$$Degree(A,B)=\frac{2\partial}{\partial+d};\ 0<\partial<=1 \quad (3)$$

where ∂ is an adjustable parameter, d is the distance of the Sememe in the Sememe hierarchical tree.

When d is greater than three, the semantic distance of sememe is far, and thus the hyponymy strength is determined as zero.

The collection of the feature words pairs is acquired by the relational words pairs which are extracted from the training corpus and feature items, according to the calculation of hyponymy strength to get the relational words pairs. The threshold of hyponymy strength is set as C, so it needs to meet the following formula:

$$Degree(w_1,w_2)>C \quad (4)$$

After the filtering, we can get the feature words pairs.

2) Expansion the feature of the test corpus

Expansion the feature vector of the test corpus by the relational words pairs which are extracted from the training corpus and feature items, which are described as follows:

Feature vector expansion is show in Figure 2.

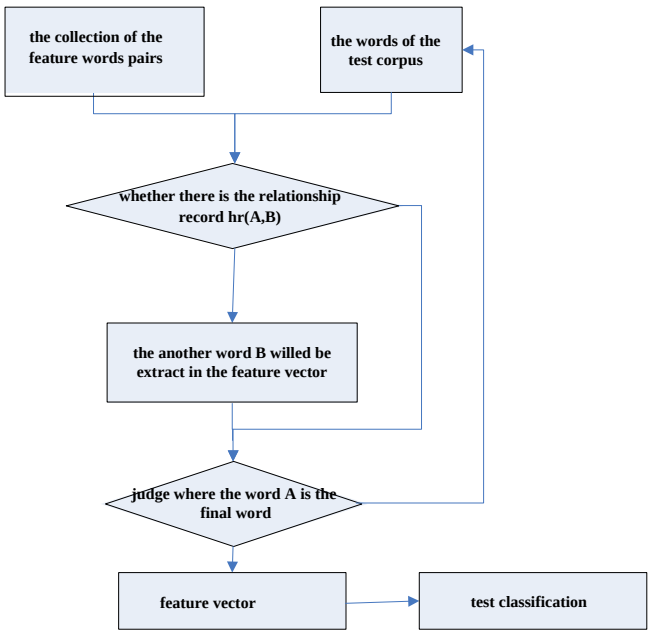

Figure 2. Feature vector expansion.

Step1: Judge whether there is the relationship record hr(A,B) by the collection of the feature words pairs and the words of test corpus.

Step2: If there is a relationship record, another word B will be extracted in the feature vector, otherwise go to step3.

Step3: Judge where the word A is the final word. If it is the final word, output the feature vector for text classification, otherwise input the next test word.

After the data processing, the text classification is the same as the traditional text classification.

D. Multi-feature Integration

The importance of each feature for classification is unknown, so we need to get the best weight of each feature. Depending on the different length of each feature, the text is classified by the traditional text classification and short text classification to train the different weight of each feature.

1) Training the weight of each feature

In this paper, the weight of each feature is obtained by linear training which is as follows:

Step1: The classification result of categories which are got by the training of feature t_i is denoted as $\vec{P}_{ti} = (P_{i1}, P_{i2}, ..., P_{in})$.

Step2: If the vector $\vec{a}$ is the vector of the different weight of each feature, the classification weights of the text can be gotten by (5):

$$\vec{f} = \vec{P} \cdot \vec{a} \tag{5}$$

Step3: By labeling the type for the text of the training corpus, we can get an equation group about $\vec{a}$ using (5) and the value of the vector $\vec{a}$ by linear regression method.

Step4: Solve the average of M training texts, so the final vector of the different weight of each feature can be obtained:

$$\vec{a} = \frac{1}{M}\sum_{i=1}^{M}\vec{a}_i \tag{6}$$

2)Feature integration

Feature integration is calculated as:

$$\overrightarrow{f} = \sum_{i=1}^{m} a_{ti}\overrightarrow{P_{ti}} \tag{7}$$

where $\vec{P}_{ti} = (P_{i1}, P_{i2}, ..., P_{in})$ is the classification result of categories which is got by the training of feature t_i, $\vec{a}$ is the vector of the different weight of each feature, m is the number of features of the blog article, n is the number of categories.

The classification result is the category which has the highest score in the vector

IV. Experiments and Result Analysis

In the experiment, after the experimental data is collected, three group experiments are made and the results are recorded and, analyzed.

A. Experimental Data

The content of http://blog.sina.com.cn/ is as the reference materials to get the name of the category, and Chinese blog category is the following eight categories: news, sports, finance, entertainment, shopping, reading, travel, and military.

2400 Chinese blog pages as the training data for each category are downloaded from the website of http://blog.sina.com.cn/, and the testing data is 200 Chinese blog pages for each category which are downloaded from the website of http://blog.sina.com.cn/.

B. Experimental Results

There are three experiments. The first experiment only has the traditional text classification method, the second one uses the traditional text classification method and the algorithm of multi-feature integration, and the third one combines the traditional text classification method, the short text classification method and the algorithm of multi-feature integration. Performance evaluation of Chinese blog classification mainly includes the accuracy rate (P), recall rate(R) and $F1$.

$$F1 = \frac{2 \times P \times R}{P + R} \tag{8}$$

TABLE I. THE RESULT OF BOLG CLASSIFICATION WHICH ONLY HAS THE TRADITIONAL TEXT CLASSIFICATION METHOD

Blog category	Training corpus/ Testing corpus	*P* (%)	*R* (%)	*F*1 (%)
news	2400/200	84.2	83.0	83.6
sports	2400/200	86.1	85.2	85.6
finance	2400/200	80.3	84.7	82.4
entertainment	2400/200	88.6	87.3	87.9
shopping	2400/200	83.7	84.5	84.1
reading	2400/200	87.1	85.6	86.3
travel	2400/200	88.4	86.3	87.3
military	2400/200	85.6	86.7	86.1

TABLE II. THE RESULT OF CHINESE BLOG CLASSIFICATION WHICH HAS TRADITIONAL TEXT CLASSIFICATION METHOD AND THE ALGORITHM OF MULTI-FEATURE INTEGRATION.

Blog category	Training corpus/ Testing corpus	*P* (%)	*R* (%)	*F*1 (%)
news	2400/200	85.0	83.4	84.2
sports	2400/200	87.2	87.6	87.4
finance	2400/200	82.1	85.2	83.6
entertainment	2400/200	89.2	88.8	89.0
shopping	2400/200	84.6	85.8	85.2
reading	2400/200	89.2	86.7	87.9
travel	2400/200	89.1	87.2	88.1
military	2400/200	88.7	85.7	87.7

 ISBN: 978-1-61208-202-8

TABLE III. THE RESULT OF CHINESE BLOG CLASSIFICATION WHICH HAS THE TRADITIONAL TEXT CLASSIFICATION METHOD, THE SHORT TEXT CLASSIFICATION METHOD AND THE ALGORITHM OF MULTI-FEATURE INTEGRATION.

Blog category	Training corpus/ Testing corpus	P(%)	R(%)	F1(%)
news	2400/200	87.7	89.3	88.5
sports	2400/200	89.1	88.2	88.6
finance	2400/200	85.5	89.1	87.3
entertainment	2400/200	90.2	90.9	90.5
shopping	2400/200	87.6	88.1	87.8
reading	2400/200	90.2	89.7	89.9
travel	2400/200	93.1	90.6	91.8
military	2400/200	92.2	89.7	90.9

Accuracy rate and recall rate reflect two different aspects of classification quality, while a comprehensive evaluation index of the two aspects is the *F*1 value which is shown in Figure 3.

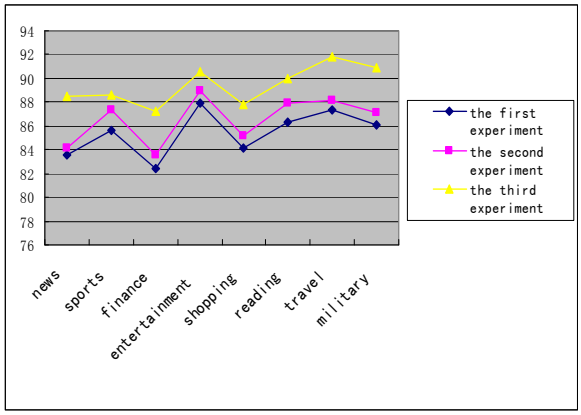

Figure 3. The comparison of comprehensive index *F*1 value

The results are clearly showed in the three tables, the compare results are expressly showed in the figure 2, and the result analysis will be introduced in the Section C.

The running times of the experiments are shown in the Table IV.

TABLE IV. THE RUNNING TIMES OF THE EXPERIMENTS

The times	The first experiment	The second experiment	The third experiment
total time	56min	66min	70min

C. *Result Analysis*

Comparing the fist experiment and the second experiment, it proves that the algorithm of multi-feature integration makes the Chinese blog classification more effective. The comparison of the second experiment and the third experiment shows that adding the short text classification makes the Chinese blog classification improved.

From the table IV, we can see that the time of the first experiment is shorter than that of the second experiment, and the time of the second experiment is shorter than that of the third experiment.

V. CONCLUSION AND FUTURE WORK

This paper presents a Chinese blog classification method which is based on the algorithm text classification and feature integration. Experimental results show that Chinese blog classification method proposed in this paper makes the classification accuracy of classification improved.

On the other hand, the blogger is the author of the blog. The blogger's interest directly affects the blog category, and the user's interest model will be structured to assist blog classification and more work about the classification will be done to perfect the blog classification.

ACKNOWLEDGMENT

The experiment used Institute of Computer Technology segmentation system interface, the date of HowNet and the data of [14], a special thanks here.

REFERENCES

[1] Hui Y., Bin Y., Xu Z., Chunguang Z., Zhe W., and Zhou C. Community discovery and sentiment mining for Chinese blog [C]. Fuzzy Systems and Knowledge Discovery. 2010, pp. 1740-1745.

[2] Aixin S., Suryanto M. A., and Liu Y. Blog classification using tags :an empirical study [C]. International Conference on Asia-Pacific Digital Libraries. 2007, pp. 307-316.

[3] Singh A. K and Joshi R. C. Semantic tagging and classification of blogs [C]. 2010 International Conference on Computer and ommunication Technology. 2010, pp. 455-459.

[4] Mai L and Nenghai Y. Multi-feature Fusion Method for Blog Post Classification [J]. Journal of Chinese Computer Systems. 2010, pp. 1129-1132.

[5] Luli C. Chinese Weblog Pages Classification Based on Folksonomy [J]. Computer engineering. 2009, pp. 50-52.

[6] Wajeed M. A. Building clusters with distributed features for text classification using KNN [C], in 2012 International Conference on Computer Communication and Informatics. 2012, pp. 583-605.

[7] Lewis D. Naive(Bayes)at forty:The independence assumption in information retrieval [C]. Lecture Notes in Computer Science. Heidelberg:Springer-Verlag, 1998, pp. 4-15

[8] Vapink V. Statistical Learning Theory [M]. New York:Spromger, 1998.

[9] Vapnik V.The Nature of Statistical Learning Theory [M]. New Yorlc:Springer, 1995.

[10] Qian Z., Mingsheng Z., and Min H. Study on Feature Selection in Chinese Text Categorization [J]. Journal of Chinese Information processing. 2004, pp. 17-23.

[11] Qiang N., Zhixiao W., Dai C., and Shixiong X. Web Document Classification Based on SVM [J]. Microelectronics & Computer. 2006, pp. 102-104.

[12] Yahui N., Xinghua F., and Yu W. Short Text Classification Based on Domain Word Ontology [J]. Computer Science. 2009, pp. 142-145.

 ISBN: 978-1-61208-202-8

[13] Huiqing C and Shiping L. A Taxonomic Relation Extraction Method Based on HowNet and Bootstrapping [C]. 2009 Communication theory and new technology development-The Fourteenth National Youth Conference on communication. 2009, pp. 102-108.

[14] http://blog.sina.com.cn/.

 ISBN: 978-1-61208-202-8