Research on Network Public Opinion Text Representation Strategy for Subject Classification——Taking Sina Weibo as an Example

Longjia Jia[1, a], and Kun Hou[2, b]

[1]School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China

[2]School of Computer Science and Information Technology, Northeast Normal University, Changchun 130024, China

[a]jialongjia@nenu.edu.cn; [b]houk431@nenu.edu.cn

Keywords: Internet public opinion security; Theme classification; Text representation strategy; Machine learning

Abstract: In this paper, we propose a text representation strategy, which solves the problem that term weights of Sina Weibo topic classification research are not suitable and the model explanatory is not strong. In the proposed document representation strategy, term weighting vector is constructed by taking pre-selection prediction. On training set, the effectiveness of term weighting vector is evaluated by cross-validation, and term weighting vector corresponding to the best evaluation result is selected as term weighting vector of test set. Compared with traditional *W-Max*, *D-Max* and *D-TMax* methods, the proposed method increases 4.25%, 5.03% and 7.10% respectively in $MicroF_1$. In classification of public opinion topics, the proposed method can construct a more explicit term weighting vector for data set. It can enhance the interpretability of the model, and improve the classification performance.

1. Introduction

With the rapid development of the Internet on a global scale, online media has been recognized as the "fourth media" after newspapers, radio, and television. Online media has become an amplifier for the distribution of ideas and cultural information and public opinion [1-3]. College students are the most active and sensitive group of netizens who reflect the social hotspot phenomenon. It is easy to express their views on social hot issues through the Internet. College student groups have both a strong sense of civic responsibility and the natural advantages of organizational actions. When they encounter certain sensitive hot topics, they are easily motivated by their social responsibility and national sentiments, which in turn triggers large-scale online public opinion. As an integral part of social public opinion, college network public opinion reflects and influences the formation and development of social public opinion to a certain extent. Therefore, the analysis and research on the public opinion security of universities has broad application prospects and very important practical and practical significance.

The classification of college public opinion topics based on Sina Weibo data is a special document classification task with very short training sets and test sets. For the task of constructing the classifier, one of the most basic tasks is the document representation. The current researchers use the vector space model representation method. However, there are few researches on the representation strategy of vector space model. At present, the global strategy is generally adopted. The global strategy has the problem that term weight is not accurate, which leads to the problem that the model is not very explanatory. Aiming at the problem of college public opinion security classification, a general representation strategy is urgently needed. In order to solve the above problem, this paper studies two common representation strategies, local strategies and global strategies. At the same time, the three representation strategies proposed by Younghoong Ko are discussed and analyzed [4]. Since we have not found any similar research work, the research content of this paper is of great significance and value in the classification of college public opinion security topics. A representation strategy is proposed, we name it supervised document

DOI: 10.25236/iciss.2019.058

representation strategy (*SDRS*). By comparing the effect of term weighting vector on each category on the training set classification, combined with the characteristics of the category distribution, a feature weighting vector suitable for the current data set can be constructed. The experimental results show that the proposed method can effectively improve the classification performance.

The remainder of this paper is organized as follows. We briefly review several document representation strategies and introduces our proposed strategy for supervised term weighting schemes.in Section 2. We show experimental results in Section 3, and finally, we draw conclusions in Section 4.

2. Methodology

In vector space model (*VSM*), a document is represented as a vector in term spaces, such as $d=\{t_1, t_2, \ldots, t_n\}$, where n is the total number of features. The value of t_i between [0,1] represents how much the term t_i contributes to the semantics of document d. The terms in VSM are extracted from training set. They can be words, phrases, or n-grams, etc. The value of the feature represents the extent to which the feature contributes to the semantics of the document. The feature weighting method can be used to weight the elements in the vector to clarify their contribution in the classification, thereby enhancing the interpretability of the model and improving the system classification performance.

Most studies indicate that on the same dataset, the supervised feature weighting method produces a classification effect that is generally superior to the unsupervised feature weighting method [5-7]. Most of the current research methods do not elaborate on document representation of test sets. There are two main strategies in the current research, local strategy and global strategy. Global policy is defined as Eq. 1.

$$TW(t) = \max_{i=1}^{|C|} TW(t, c_i) \tag{1}$$

In Eq. 1, *TW(t)* is the final weight of a term *t*; *TW(t,c$_i$)* is weight of term *t* in category c_i obtained with supervised term weighting methods. |*C*| is the number of categories. In the process of initial representation, a test document can be represented as |*C*| different vectors. After using appropriate selection policy, it can be represented as one vector which well describes the document. Global policy selects the maximum term value among all categories for each term. Although this method is effective in some cases, but not sure if it has the ability to select the most effective term weighting vector for current test samples [8,9].

The Sina Weibo document studied in this paper has the following particularities compared to normal text documents. First, the Weibo document contains fewer words, and the amount of information contained in the classification is less. Second, the stop words in Sina Weibo documents are higher than normal text documents. Third, Sina Weibo documents have special symbol expressions. Fourth, Sina Weibo documents are very short. Due to the large amount of data set, the feature matrix will be extremely sparse.

For the classification of public opinion security topics based on Sina Weibo data, the existing text representation strategy is still valid or not. If they are effective, which kind of representation strategy can get the best results? This is the question we hope to solve in this study.

Besides local policy and global policy, Younghoong Ko proposed the following three solutions for this problem, i.e., *W-Max*, *D-Max* and *D-TMax*. They are described as follows.

W-Max: each term’s value of term weighting vector will be replaced by the maximum value of the corresponding dimension’s term weight in all categories. After comparing with global policy, we may find that they have the same idea.

D-Max: the sum of all term weights in each term weighting vector is first calculated and then one term weighting vector with the maximum sum value is selected as the document representation vector.

D-TMax: the sum of all term weights in each term weighting vector is calculated and then two term weighting vectors with the two largest sum values are selected. Then the term weighting vector is constructed by choosing the higher term weighting value from the selected two term weighting vectors for each corresponding dimension’s term weight.

The methods proposed in this paper have three main improvements compared to traditional methods such as *W-Max*, *D-Max* and *D-TMax*.

(1) The proposed method makes a breakthrough in text representation. For the supervised term weighting method, it is no longer based on the researcher's experience to use some text representation strategy, but intelligently construct term weighting vectors based on current data set.

(2) The proposed method introduces the idea of loop traversal. The alternative vector for reconstructing term weighting vector is no longer limited. According to the distribution of each category, a term weighting vector suitable for current data set can be constructed.

(3) The proposed method implements pre-selection prediction. When constructing the term weighting vector, the effect of term weighting vector is tested on training set in a similar way to cross-validation. Term weighting vector is constructed according to the relationship between documents, features and categories.

In the proposed method, by traversing the term weighting vectors generated by each class, we compare their weighting effects on the training set. The term weighting vector which produces the best effect on training set will be selected as the term weighting vector of test set. We summarize the main process is shown in Table 1.

Table 1 The proposed method

```
Algorithm 1: Supervised Document Representation Strategy based on ergodic ideas
Input:
  fea: feature matrix of training set
  gnd: a vector of labels for documents in training set
Output:
  selectedC: the most appropriate N value for current dataset
Local variables
  |C|: total number of categories;
  M: total number of features;
  termWeightingVec1: the set of |C| original term weighting vectors;
  $termWeightingVec1_i$: i-th vector of the original term weighting vectors;
  $termWeightingVec2_i$: i-th vector of the reconstructed term weighting vectors;
  $sumVec_i$: sum value of all terms in i-th term weighting vector;
  sortSum: sorted list of each sum values;
  $weightedFea_i$: the weighted fea by using $termWeightingVec2_i$;
  $MicroF_1^i$: result of 10-fold cross validation on $weightedFea_i$;
begin
      apply supervised term weighting method to fea, and get termWeightingVec1;
      for i = 1 to |C|
          for j = 1 to M
              compute $sumVec_i$ for $termWeightingVec1_i$;
          end for
      end for
      sort all $sumVec_i$, and get sortSum;
      for i = 1 to |C|
          for j = 1 to M
              for k=1 to i
                  construct $termWeightingVec2_i$ by the following ways. The j-th  dimension of each term
                  weighting vector in the selected k term weighting vectors is obtained, and the maximum
                  value will be selected as the j-th value of the $termWeightingVec2_i$;
              end for
          end for
        end for
      for i = 1 to |C|
          compute $weightedFea_i$;
      end for
      for i = 1 to |C|
          compute $MicroF_1^i$;
      end for
      record i corresponding to the maximum $MicroF_1^i$, and assign it to selectedC;
end
```

After the algorithm 1 is executed, we can get *selectedC*. Before weighting test set, select the top *selectedC* vector. According to above steps, the term weighting vector of test set is constructed.

3. Experimental Results

3.1. Data Corpora.

In order to verify the effectiveness of the proposed algorithm, this paper applies web crawler technology to capture 20,000 college microblog document data from Sina Weibo. Extract some microblog documents from the source data according to the following rules. First, the microblog document of the plain text type is selected. Second, the microblog document with more than 120 characters is selected. Through these two rules, a total of 13,079 Weibo documents are selected. According to the "2016 China University Government New Media Development Report", the top 10 types of campus students' microblogs are leisure and entertainment, humanities and art, science and technology, education, transportation services, news, reading, exercise fitness, public welfare, emotions. In the experiment, the above 10 categories are used as target categories. The extracted data is marked in following rules. All data are labeled twice by four people. The results of two labels are checked one by one. Documents that have same content but different category labels need to be screened and discussed separately. Discard Weibo documents that are difficult to identify categories. The annotated data set contains a total of 9,183 Weibo documents. According to statistics, the number of Weibo documents included in each category is shown in Table 2.

Table 2 Correspondence between category and document number

category	number of documents	category	number of documents
leisure and entertainment	1,528	news	1,275
humanities and art	826	reading	1,136
science and technology	772	exercise fitness	845
education	425	public welfare	395
transportation services	576	emotions	1,405

3.2. Learning Algorithms and Performance Evaluation.

To evaluate classification performance of the proposed method, we choose the promising learning algorithms in this study, i.e., *SVM* classifier [10]. Although other algorithms such as Decision Tree and Naive Bayes are also widely used, they are not included because the real number format of term weights could not be used except for the binary representation (see an exception in [10]).

In this paper, $MicroF_1$ is employed to measure the performance of the proposed method.

3.3. Results and Analysis.

This paper will use the *tf*rf* combined with *W-Max, D-Max, D-TMax* and the methods proposed in this paper to compare their results [11,12].

Compared with the three strategies, the *SDRS* method proposed in this paper obtains the optimal results. Table 3 shows the $MicroF_1$ measure result.

Table 3 $MicroF_1$ measure result

Method	$MicroF_1$
D-MAX	0.7836
D-TMAX	0.7629
SDRS	0.8339
W-MAX	0.7914

In order to show the performance of the proposed method, we list the result of optimal *selectedC* which are selected by *SDRS*. We also report the results of classification experiments with different parameters in Table 4.

Table 4 A comparison on MicroF1

selectedC	*MicroF$_1$*	*selectedC*	*MicroF$_1$*	*selectedC*	*MicroF$_1$*
1	0.7836	**5**	**0.8339**	9	0.7903
2	0.7629	6	0.8256	10	0.7914
3	0.7862	7	0.8224		
4	0.7871	8	0.8175		

By observing the results of the *MicroF$_1$* values in Table 4, it can be known that when selecting a text representation strategy for the microblog short document unbalanced data set, the traditional text representation strategy is not optimal choice. In contrast, the *SDRS* can obtain better classification results. The main reason is that Weibo documents contain few characteristic words compared to regular documents. Due to the large amount of data, the data set feature matrix is extremely sparse. The traditional text representation strategy cannot construct the appropriate term weighting vector. The *SDRS* uses pre-selection prediction to construct term weighting vector. Firstly, the cross-validation method is adopted on training set to fully evaluate the effect of current term weighting vectors. Then the term weighting vector corresponding to the best evaluation result is selected as the term weighting vector of test set.

4. Conclusion

With the rapid development of Weibo, there is an urgent need for the classification of Weibo documents. However, as a special short text document, each Weibo document contains fewer feature words. In this paper, we have studied several widely used text representation strategies. At the same time, we propose a new text representation strategy, which has obvious effects on classification of college public opinion topics. The innovations are as follows. By pre-selection prediction method constructs term weighting vector, which avoids the problem of poor classification results caused by traditional text representation strategies. The proposed method satisfies the practical requirements of text representation problem in the network public opinion classification, and can provide certain technical methods for the network public opinion analysis of universities.

Acknowledgements

This paper was sponsored by Philosophy and Social Science School Youth Fund Project of Northeast Normal University (Grant No. 19XQ029). We would like to thank the organization for the supports.

References

[1] Salloum S A, Mhamdi C, Al-Emran M, et al. Analysis and classification of Arabic newspapers' Facebook pages using text mining techniques[J]. International Journal of Information Technology and Language Studies, 2017, 1(2): 8-17.

[2] Arganda-Carreras I, Kaynig V, Rueden C, et al. Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification[J]. Bioinformatics, 2017, 33(15): 2424-2426.

[3] Akaichi J. Sentiment Classification: Facebook'Statuses Mining in the "Arabic Spring" Era[M]//Artificial Intelligence: Concepts, Methodologies, Tools, and Applications. IGI Global, 2017: 1858-1883.

[4] Ko, Youngjoong. "A study of term weighting schemes using class information for text classification." Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

[5] Cai, Deng, Xiaofei He, and Jiawei Han, Locally consistent concept factorization for document clustering. Knowledge and Data Engineering, IEEE Transactions on 23:902-913 (2011)

[6] Pereira, Rafael B., et al. "Categorizing feature selection methods for multi-label classification." Artificial Intelligence Review 49.1 (2018): 57-78.

[7] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27.

[8] Quan X, Wenyin L, Qiu B. Term weighting schemes for question categorization[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 33(5): 1009-1021 (2011)

[9] Miao, Yun-Qian, and Mohamed Kamel. "Pairwise optimized Rocchio algorithm for text categorization." Pattern Recognition Letters 32.2 (2011): 375-382.

[10] Yang, Yiming, An evaluation of statistical approaches to text categorization. Information retrieval 1:69-90 (1999)

[11] Lan M, Tan C L, Low H B. Proposing a new term weighting scheme for text categorization[C] AAAI, 6: 763-768 (2006)

[12] Lan M, Tan C L, Su J, et al. Supervised and traditional term weighting methods for automatic text categorization[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31(4): 721-735 (2009)