Journal of New Media
DOI:10.32604/jnm.2021.018923
Article

Tech Science Press

Research on Feature Extraction Method of Social Network Text

Zheng Zhang[*] and Shu Zhou

School of Computer & Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[*]Corresponding Author: Zheng Zhang. Email: 18362086511@163.com

Received: 26 March 2021; Accepted: 30 March 2021

Abstract: The development of various applications based on social network text is in full swing. Studying text features and classifications is of great value to extract important information. This paper mainly introduces the common feature selection algorithms and feature representation methods, and introduces the basic principles, advantages and disadvantages of SVM and KNN, and the evaluation indexes of classification algorithms. In the aspect of mutual information feature selection function, it describes its processing flow, shortcomings and optimization improvements. In view of its weakness in not balancing the positive and negative correlation characteristics, a balance weight attribute factor and feature difference factor are introduced to make up for its deficiency. The experimental stage mainly describes the specific process: the word segmentation processing, to disuse words, using various feature selection algorithms, including optimized mutual information, and weighted with TF-IDF. Under the two classification algorithms of SVM and KNN, we compare the merits and demerits of all the feature selection algorithms according to the evaluation index. Experiments show that the optimized mutual information feature selection has good performance and is better than KNN under the SVM classification algorithm. This experiment proves its validity.

Keywords: Social network text; mutual information; positive and negative correlation characteristics; SVM; KNN

1 Introduction

Today, the development of the Internet has gone through decades. The development of these decades has brought many changes to various places. These changes make it easier for people to communicate, and the distance between people is getting closer and closer. The overall scale of China's mobile Internet market is also increasing.

Recently, e-commerce, block-chain and artificial intelligence have developed rapidly. Among them, social networks are still showing an explosive popularity, such as domestic microblogs and foreign Twitter. These applications make information transmission very convenient, and at the same time, it contains a large amount of video information, which makes the text information content more colorful.

E-commerce, social networking platforms (such as Weibo, Twitter, Facebook), search engines, these entities have a sea of text data. Among them, e-commerce platform has a large number of commodity information, commodity evaluation. There is a large amount of comment information on social networking platforms. Different from e-commerce product information, these messages tend to have shorter content. The information we call social network text. Therefore, these social network texts play a very important role in communication processes, which have important research significance.

In the face of a large number of social network texts generated by these e-commerce platforms, social network platforms, etc., accurate extraction of important and reasonable information is the key. For example, for positive and negative social network text information, social network text classification occupies a higher position. The study of text classification in social networks is of great significance for the follow-up work, such as text description and text data mining. Based on the text characteristics of social network text, this paper studies and discusses the existing problems, and analyzes the accuracy of the influence of different information feature extraction on social network text.

What is text feature selection in social networks? Simply put, in a specific social network text feature selection system, the process of separately categorizing the specified social network text into several categories. The text feature selection of social networks has important research and application value for data mining and information processing. Social network text feature selection can process a variety of huge information quickly, and has more research value in various information processing fields. In the early text feature selection of social networks, there are mainly two types. The first is manual feature selection [1], which is mainly based on the existing rules and the characteristics of these rules for manual feature selection. The second is automatic feature selection [2–6]. Automatic feature selection mainly defines a number of rules with their own characteristics in each category, and actively divides the text with these characteristics into corresponding items.

2 Related Work

In the field of social network text feature selection and feature selection, the start is relatively early in foreign countries and a little later in China. The earliest start of foreign countries can be traced back to the 1950s, which can be divided into three major categories: the first category is the selection of automatic features of the text that can carry out correct design and planning research, and the second category is based on the first is the experimental test and research, and the third is the practical application scenario landing research based on automatic feature selection of text.

In 1979, the research achievements in the field of information retrieval are summarized, after that, until 1989, based on the characteristics of the manual writing rules to form the characteristics of the method in a very important part of social network text feature selection, until the ninety's, the network era emerge, social network text feature selection for the first time introduced the vector machine method based on linear kernel function. With the continuous update of this technology, more and more feature selection models and algorithms have been born, and these technologies have been widely used in today's information processing. Foreign scholars naturally stepped into this field and carried out in-depth research on automatic feature selection of network information resources. Recently, the newly emerging feature selection methods of social network text include the whole text classification feature selection method of social network text, multiple feature selection method and the improved feature selection method of classical feature selection method.

The technology of automatic text feature selection in social networks was first proposed by in 1981. It studied the relevant applications of document feature selection in computers, including automatic text feature selection technology and text feature selection search in social networks. After this, text automatic feature selection research technology developed rapidly in China, and a large number of scholars carried out systematic research on it. Due to the foreign research on automatic text feature selection started earlier, so the domestic scholar's researches of the original text in English words as the carrier specializes in, with the later Chinese scholars of relevant technologies such as feature selection algorithm of continuous improvement, social network also has been able to undertake automatic text feature selection, text in the end, formed a relatively systematic social network automatic text feature selection research techniques.

Zhu et al. developed an actionable text document feature selection system in 1986 [1]. The first Chinese-based text feature selection system appeared in 1995. Zhang et al. proposed in 1998 that the automatic feature selection system could design a larger correlation between the two based on the category features of the computer and all the features contained in the text [7].

Domestic scholars are not only to social network text feature selection system constantly upgraded and improved, the social networking social text feature selection algorithms are also carried on the thorough research. Up to now, automatic feature selection technology of social network text has gradually matured, and many research results have been widely applied.

Since English words are separated by Spaces, social network text needs word segmentation, so the preprocessing of social network text is critical.

3 Method

3.1 Description of Traditional Mutual Information Text Feature Selection Algorithm

Mutual Information feature selection (MI) mainly describes the existence of two signals in a single message. Different messages have different degrees of interdependent signals. Mutual Information mainly refers to the degree of interdependence of these signals. When facing the application scenario of text classification, the mutual information mainly faces the correlation between the key words to describe the features and the specific categories to which the classification belongs, as well as the correlation between the descriptions of the key words.

When there are two variables, we use their MI value to describe each other:

$$MI(t_i, c_j) = \log \frac{P(t_i \mid c_j)}{P(t_i)} = \log \frac{P(t_i \mid c_j)}{P(t_i) * P(C_j)} \quad (1)$$

When the denominator is the same as the numerator, MI = 0, which means that the specified key feature words are independent of (i.e., incompatible with) the selected category. When the numerator is less than the denominator, this is MI < 0, which means that the number of occurrences of a category is very small. So we can get a simple conclusion: the smaller the MI, the less likely it is that a keyword with a certain feature will be selected for that category.

Here we define the meanings of certain letters, as shown in Tab. 1.

Table 1: Document frequency meaning of various representatives

Alphabet	Does it contain feature t_i	Whether it belongs to category c_i
S	Yes	Yes
V	No	Yes
N	Yes	No
F(The total number of documents)	No	No

Therefore, MI formula of t_i and c_i can be obtained:

$$MI(t_i, c_j) = \log \frac{S * F}{(S+N)*(S+V)} \quad (2)$$

Specific problems with m categories of classification, mainly to calculate the MI of an eigenvalue t in each specific category. Represents the MI of an eigenvalue t in all data sets:

$$MI_{\max}(t) = \max_i MI(t, c_i) \quad (3)$$

The average value is:

$$MI_{avg}(t) = \sum_{i=1}^{m} P(c_i) MI(t, c_i) \quad (4)$$

3.2 Disadvantages of Mutual Information Feature Selection

The main disadvantages are as follows: (1) It is assumed that the specific correlation between a certain word and a certain classification is mainly calculated when classifying under a specific classification algorithm, so the influence of the occurrence frequency (word frequency) of a certain word is not considered when calculating MI; (2) When calculating its correlation, if there are many word segments and categories in the same document, the amount of computation will present a higher order of magnitude; (3) The different number of texts in different categories will also hinder the calculation of MI of a particular keyword, which is a tricky problem; (4) MI only considers the correlation degree between some feature keywords and some categories, but does not consider the correlation degree of these feature keywords.

3.3 Improvement of Feature Selection Process for Mutual Information

3.3.1 Balance Weight Attribute Factor

In the case of ordinary text classification algorithms, those that show a useful classification effect are called Positive Correlation Characteristics (PCC), and vice versa, Negative Correlation Characteristics (NCC).NCC can help to extract text that is not needed for text classification, for which some specific feature items can be explained without much relation to the classification category.

Here, we propose a balance weight attribute factor α , mainly to adjust the relationship between NCC and PCC so that they reach a balance. At this time, in the case of PCC, MI is as follows:

$$MI(\mathrm{t},\mathrm{C}_i)^+ = \alpha * \log\frac{p(t\,|\,c_i)}{p(t)}, 0 < \alpha < 1 \tag{5}$$

In the case of NCC, MI is:

$$MI(\mathrm{t},\mathrm{C}_i)^- = (1-\alpha) * \log\frac{p(t\,|\,c_i)}{p(t)}, 0 < \alpha < 1 \tag{6}$$

The overall MI after improvement is:

$$MI(\mathrm{t},\mathrm{C}_i) = MI(\mathrm{t},\mathrm{C}_i)^+ - MI(\mathrm{t},\mathrm{C}_i)^- \tag{7}$$

3.3.2 Characteristic Difference Factor

When some features are under certain specific classification conditions, and at the same time meet a relatively average way, we can think that this feature and these specific classifications are presented as a high correlation. This can be better represented in the case of classification. These features are specifically expressed in discretized formulas as:

$$K_{ac} = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(tf_i(T_k) - \overline{tf(T_k)})^2}}{\overline{tf(T_k)}} \tag{8}$$

$\overline{tf(T_k)}$ is the uniform word statistical frequency of a word in each classification category, and $tf(T_k)$ is the probability of a word in a certain classification category. The larger the value, the better the classification. The formula for discretizing the degree of certain features within the classification is specifically expressed as:

$$K_{ic} = \frac{\sqrt{\frac{1}{m}\sum_{j=1}^{m}(tf_{ij}(T_k) - \overline{tf_i(T_k)})^2}}{\frac{m}{\sqrt{m-1}}\overline{tf_i(T_k)}} \tag{9}$$

The interpretation is similar, except for the statistical frequency in the document. The lower the value, the smaller the classification. Next, import the word frequency factor A, so that the characteristic difference factor can be expressed as:

$$P_r = \frac{K_{ac}}{K_{ic}} \tag{10}$$

So the whole formula becomes:

$$MI(\mathrm{t}, \mathrm{C}_i) = P_r * P(w) * (MI(\mathrm{t}, \mathrm{C}_i)^+ - MI(\mathrm{t}, \mathrm{C}_i)^-) \tag{11}$$

3.3.3 Text Processing Process Improvement

When processing text with improved mutual information, the whole process is shown in Fig. 1.

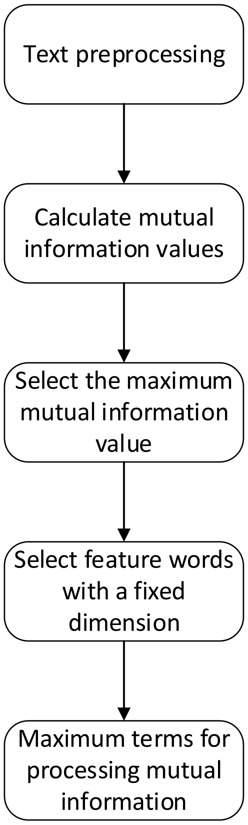
Text preprocessing
Calculate mutual information values
Select the maximum mutual information value
Select feature words with a fixed dimension
Maximum terms for processing mutual information

Figure 1: Schematic diagram of mutual information feature selection process

From the Fig. 1, it is the third point that we should focus on, which is to select the maximum mutual information value. So that is where the optimization comes in.

4 Experiment

4.1 Experimental Idea and Evaluation Index

The main experimental ideas are as follows: word segmentation is carried out first, then words are stopped, various feature selection algorithms are used, including optimized mutual information, and TF-IDF is used for weighting. According to the evaluation index, the advantages and disadvantages of all the feature selection algorithms are compared under SVM and KNN classification algorithms.

4.1.1 Word Segmentation Process

The first step of text classification preprocessing in social networks is mainly word segmentation preprocessing, because most text information consists of words at first, and many words constitute words.

Therefore, more information can be obtained directly through word segmentation to some extent, so word segmentation has become a necessary prerequisite and a primary task.

At present, the commonly used word segmentation methods are mainly divided into several: the first one is carried out in this way of statistics, through the statistics of the number of fields before and after the full text, calculate the mutual information between each paragraph and chapter, etc., and finally carry out word segmentation processing; The second method is based on the understanding of this way, mainly using the principles of lexical analysis and semantic analysis in the compilation principle to process the relevant text information, but this premise is to need enough basic text information to help understand; The third one is carried out in the way of matching, which mainly matches the main information in the text with the existing dictionary data, including forward and reverse matching and so on.

Here, we assume that a paragraph of text has been processed, and the word segmentation effect formed after processing is shown in Fig. 2.

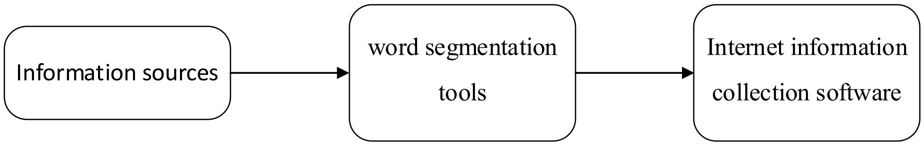

Figure 2: The result of a paragraph of text processed by word segmentation

4.1.2 Stop Words

The second step in the pre-processing of social network text classification is to remove the stop words. The main function of the stop words is to remove the interference of social network text, such as auxiliary words and conjunctions of the main feature information. Removing these meaningless stop words is conducive to better extraction of the feature information of the text. Moreover, the suspended words have the following characteristics in text information processing: (1) the suspended words have little correlation with the extracted feature information of the short text. Reducing these stops can increase the accuracy of feature extraction of text information to a certain extent. (2) Stopped words are generally keywords that constantly appear in text paragraphs. The classification ability of ceaseless stop words in text paragraphs is not improved enough. These stop words may not be based on higher classification information to a certain extent, and will increase the time loss of the algorithm.

To sum up, when these stop words appear in the text, we must try our best to remove them, so as to bring certain accuracy and efficiency to the later feature extraction.

4.1.3 Feature Weighted Representation of Text

Here, TF-IDF is selected for feature weighting. The feature weight calculation of TF-IDF mainly has two characteristics. The first characteristic is that if a feature appears more frequently in social network text, it will have more weight for overall classification or other feature extraction. The second feature is that if this feature is more frequent in all the social network texts to be extracted, it will be more difficult to distinguish all the social network texts. So in order to combine the advantages and disadvantages of these two characteristics, it is necessary to set up some special methods to calculate。

In the traditional feature weight calculation, there are many methods, among which the most famous is the TF * IDF weight calculation method, which is widely used and constantly developed.

Now, this paper mainly introduces the weight calculation method of TF * IDF, which mainly considers three calculation factors:

(1) Considering the occurrence frequency of words in the text of social network, TF is used here to represent them.

(2) Considering the inverse document frequency in social network text, IDF is used here to represent it; IDF mainly represents the data distribution of keywords or main feature words in all texts of social network texts.

(3) The normalization factor of social network text is introduced, and the problem of different text lengths exists in the text, which is mainly to eliminate such problems and carry out normalization processing.

The calculation formula of tf * idf is as follows:

$$W_{ik} = \frac{tf_{ik} * \log(\frac{N}{n_k} + 0.01)}{\sqrt{\sum [tf_{ik} * \log(\frac{N}{n_k} + 0.01)]^2}} \tag{12}$$

where, tf_{ik} represents the frequency of word tk appearing in social network text, and n_k represents the frequency of word tk appearing in all social network text training sets.

4.2 Analysis of Experimental Results

4.2.1 Comparison of Feature Selection Algorithms

Here we calculate all kinds of text together, and then calculate their average value for comparison. SVM and KNN are selected as the two classification algorithms, and the feature functions listed above are selected for comparison. The specific results are shown in Tabs. 2 and 3.

Table 2: Performance of different feature selection algorithms under SVM classification algorithm

Feature selection algorithm	**Metric**		
	F(%)	**Recall(%)**	**Acc(%)**
Mutual Info	44.18	46.23	43.32
optimize Mutual Info	**55.24**	**54.78**	**53.45**
Gini index	42.38	44.78	41.21
Document Frequency	42.13	44.08	41.01
chi-square test	44.14	45.23	42.12
KLIC	44.21	46.28	43.48

As shown in Tab. 2, the difference between the original mutual information in the first row and the original other data is not very big. The improved mutual information feature selection algorithm is better than others in terms of F value, recall rate and accuracy. This data shows that the optimized mutual information is real and effective, and it is a feature selection algorithm that can be used in practical applications.

Table 3: Performance of different feature selection algorithms under KNN classification algorithm

Feature selection algorithm	**Metric**		
	F(%)	**Recall (%)**	**Acc (%)**
Mutual Info	42.38	44.19	41.22
optimize Mutual Info	**53.54**	**52.68**	**51.45**
Gini index	40.18	42.58	40.21
Document Frequency	40.03	42.08	40.01
chi-square test	42.04	43.13	40.02
KLIC	42.01	44.18	41.18

As shown in Tab. 3, the optimized mutual information under the KNN algorithm is still better than other feature selection algorithms. It can be seen from the F-value, recall rate and accuracy of the optimized mutual information (53.54/52.68/51.45 are the maximum). However, compared with SVM (55.24/55.78/53.45) in the same situation, SVM is better than KNN in various performance parameters.

5 Conclusion

This paper mainly introduces the common feature selection algorithms and feature representation methods. The basic principles, advantages and disadvantages of SVM and KNN are briefly introduced. Next, it specifically describes the processing process, shortcomings and optimization improvement of mutual information in the face of text. It mainly introduces a weight balance factor and feature difference factor to make up for the deficiency in the processing of positive and negative correlation. In the experimental stage, word segmentation is carried out first, and then words are stopped. Various feature selection algorithms are used, including optimized mutual information, and TF-IDF is used for weighting. Under SVM and KNN classification algorithms, the advantages and disadvantages of all feature selection algorithms are compared according to the evaluation indexes. The word segmentation is carried out first, and then the words are stopped. Various feature selection algorithms, including the optimized mutual information, are used for weighting with TF-IDF. According to the evaluation index, the advantages and disadvantages of all the feature selection algorithms are compared under SVM and KNN classification algorithms.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] L. J. Zhu, "MULTI-AGENT approach for distributed flexible manufacturing systems," *Journal of Shanghai Jiaotong University*, no. 2, pp. 78–82, 1998.

[2] R. Bekkerman, R. El-Yaniv, N. Tishby and Y. Winter, "Distributional word clusters *vs.* words for text categorization," *JMLR*, vol. 3, pp. 1183–1208, 2003.

[3] K. Bennett, J. B. Bi, M. Embrechts, C. Breneman and M. Song, "Dimensionality reduction via sparse support vector machines," *JMLR*, vol. 3, 1229–1243, 2003.

[4] C. Yu-Chin, E. Michael, Y. F. Han, H. Rosen and S. Yantis, "Decoding task-based attentional modulation during face categorization," *Journal of Cognitive Neuroscience*, vol. 23, no. 5, 2010.

[5] R. Caruana and V. R. de Sa, "Benefitting from the variables that variable selection discards," *JMLR*, vol. 3, pp. 1245–1264, 2003.

[6] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *JMLR*, vol. 3, pp. 1289–1306, 2003.

[7] M. Zhang, Kilimci and M. C. Ganiz, "Higher-order smoothing: a novel semantic smoothing method for text classification," *Journal of Computer Science & Technology*, vol. 29, no. 3, pp. 376–391, 2014.