Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 166 (2020) 523–528

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Mechatronics and Intelligent Robotics (ICMIR-2019)

The Lao Text Classification Method Based on KNN

Zhuo Chen[1], Lan Jiang Zhou[2,*1], Xuan Da Li[4], Jia Nan Zhang [4], Wen Jie Huo[5]

[1]*School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China*
[2]*The Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, Yunnan 650500, China*
[3]*School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China*
[4]*Information Engineering University, Kunming team of the three schools 650500, China*
[5]*School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China*

Abstract.

Text categorization is a common application scenario in the NLP field, and has many applications in public opinion monitoring and news classification. At present, there are few classifications for Lao text, but classification-oriented methods are widely used in other languages. Faced with the situation, this paper proposes a KNN-based classification method of Lao news text. First of all, preprocessing and feature extraction of Lao news text, then adjusting the parameters through KNN classifier, finally processing in data normalization and data dimensionality reduction, thus improving the classification effect.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Mechatronics and Intelligent Robotics, ICMIR-2019.

Keywords: Lao News Category, KNN, Data Processing

1. Introduction

Text classification is a classic problem in the field of NLP, Initially, its related research was carried out using the expert rules of 1950. After 30 years, it developed into the use of knowledge engineering to establish an expert system, but the efficiency is very low.Since the 1990s, text classification methods have become more and more

[1] Corresponding Author. Tel.+(86) 13888651699
*E-mail: 915090822@qq.com

1877-0509
Peer-review under responsibility of the scientific committee of the 3rd International Conference on Mechatronics and Intelligent Robotics, ICMIR-2019.
10.1016/j.procs.2020.02.053

perfect. Common classification models are rule-based, probability-based, geometry-based, statistical model-based, the essence of which is text similarity measurement and computation[1]. However, The text categorization of Lao language is rare.

As a simple and effective text classification method without parameters, the KNN method has good accuracy and recall rate, which makes it one of the classifiers commonly used in text classification[2].Therefore, the improvement of the performance of KNN classification method is increasingly becoming a focus of attention.

2 Calculation Process of the Lao Text Classification

2.1 The Text Preprocessing

Data were collected from 2411 news articles in 7 categories, including city, international, documentary, education and sports, economy, culture and society, and other categories. 1490 papers were randomly selected as training samples, including 239 papers in cities, 301 papers in the world, 105 papers in documentaries, 195 papers in education and sports, 148 papers in economics, 115 papers in tourism, 216 papers in culture and society, 171 papers in other fields, and the remaining 921 papers were used as testing samples. The results showed that the training samples were consisted of 239 papers in cities, 301 papers in the world, 105 papers in documentaries, 195 papers in education and sports, 148 papers in economics, 115 papers in tourism, 216 papers in culture and society, 171 papers in other fields. News text is stored in a text document as a list, and data reading is read as a list. Based on the dictionary segmentation method, this lab segments the words in training and test sets.

2.2 Structured Representation-Constructing Word Vector Spaces

Structured representation of text categorization is mainly to count the frequency of words in text, and the method is a vector space model. Vector space model denotes the text as a vector, in which each feature of the vector is represented as the words appearing in the text[3]. Typically, each different string that appears in the training set is treated as a dimension, including common words, proprietary words, phrases, and other types of pattern strings.

Because the text is stored as a vector space, the dimension is high. To save storage space and improve search efficiency, some words or phrases are automatically filtered out before text categorization. These words or phrases are called stop words. Such words are usually vague words, and there are some modal auxiliaries, usually they can not play the meaning of the text classification features. In this experiment, we use TFIDF to calculate the weight of each word, transform it into a word frequency matrix, save it into a dictionary file, and output it to build a stop-word list.

2.3 Weight Policy

TF-IDF indicates that the word frequency reverses the frequency of the document. It is assumed that a word or phrase appears in a high frequency in an article, and rarely appears in other articles, and it is considered suitable for classification[4][5].

2.4 The Classifier

KNN algorithm calculates that most of the k nearest neighbors in a feature space belong to a certain category, and the sample also belongs to this category. The algorithm involves several main factors: distance measurement, k-value selection and so on[6].

First, an experiment of K value selection is performed, and the optimal k value is selected from the seven news text data by a simple cross-validation method. The test set is validated by using the model obtained from the training set to interfere with the sample selection of the training set and the test set.The results show that the effect is the best when K=4.

Secondly, distance measurements are used to measure the distance between individuals in a space. Euclidean distance is the most common distance metric used to measure the absolute distance between points in a multi-dimensional space.

$$\mathrm{Dist}(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Minkowski distance is a common measure of the distance between numerical points, but it is not a distance, but a set of definitions of distance: the Minkowski distance between two n-dimensional variables A= (x11, x12,..., x1n) and B= (x21, x22,..., x2n) is defined as:

$$D_{12} = \sqrt[p]{\sum_{k=1}^{n} | x_{1k} - x_{2k} |^p} \quad (2)$$

The experimental results are as follows (k=4):

Table.1 distance measurement experimental result

News Text Categorization Error Rate	KNN
Minkowski distance	32.2%
Euclidean distance	34.6%

Each coordinate contributes equally to the Euclidean distance. In the seven classes of KNN, when each component is a quantity of different properties, the magnitude of "distance" is related to the unit of the index. It equates the difference between different attributes of news classification text, which can not meet the requirements of accurate classification. The effect of population variation on distance is not taken into account. Minkowski distance is a generalization of Euclidean distance and a generalization of several distance measurement formulas, which can effectively solve the above problems.

2.5 KNN Optimization

When using KNN to classify Lao news texts alone, we need to do data normalization and data dimensionality reduction to solve the above problems to a certain extent, because of the large number of features, resulting in larger dimensions, and the huge computational complexity in distance measurement.

Normalization is a form of normalization that maps data into the interval [0, 1]. Construct the normalized object of mean variance, normalize it with the training set, recording the parameters needed and standardization the test set with the attributes of the training set. Describe the range of data distribution, including: variance, standard deviation, etc. After the mean variance normalized data are written, the accuracy is calculated using the KNN classifier.

In this experiment, Text classification data is normalized by zero-mean.Z-Score can be applied to numerical data and is not affected by the amount of data because its function is to eliminate the inconvenience caused by the amount of data. This is done by subtracting the average of the data by attribute and dividing it by its variance. The variance of each attribute is aggregated around 0 with a variance of 1.

$$x^* = \frac{x - \bar{x}}{\sigma} \quad (3)$$

Where $\bar{x}$ is the average of the original data and σ is the standard deviation of the original data.

Table.2 Comparison of experimental results of normalized treatment

	Not normalized	Normalized treatment
classification accuracy	67.8%	69.2%

Data dimensionality reduction removes noise and unimportant features, and retains some of the main features of high-dimensional data, thereby achieving the goal of increasing data processing speed. Currently, Singular Value Decomposition, Principal Component Analysis, Factor Analysis and Independent Component Analysis are often used for dimensionality reduction.

The main idea of Principal Component Analysis is to map n-dimensional features to k-dimensional features. This k-dimensional feature is a new orthogonal feature, also known as the main component, which is reconstructed from the original n-dimensional features[7].

In this lab, the data is normalized in advance, So when PCA is used to reduce dimension, Feature centralization is not required, After feature centralization, the calculated direction can better describe the original data, The concrete manifestations are as follows, reducing the seven dimensions of seven types of news texts in Lao to three dimensions, in which the eigenvalues and eigenvectors are calculated by using the covariance matrix, and sort the eigenvalues in descending order, and then the sample points are projected onto the selected feature vectors. Thus, the 7-dimensional features of the original Lao news text are transformed into 3-dimensional features, which is the projection of the original features on the 3-dimensional features.

Table.3 Comparison of experimental results of PCA data dimensionality reduction

	Unprocessed	PCA dimensionality reduction
classification accuracy	69.2%	71.4%

3 Experiment and Analysis

The Lao news text classification data used in this paper is unbalanced, so it will affect the classification effect of the classifier. Loss is also uneven in its response to categories, so accuracy is not a measure of accuracy in unbalanced classification tasks. In this experiment, over-sampling is used to solve the problem of unbalanced classification data[8]. The experiment shows that over-sampling can improve the experimental results.

There are three common methods to judge the classification model: Confusion Matrix, Receiver Operating Characteristic curve and Area Under Curve[9]. This experiment uses the confusion matrix to evaluate the KNN classifier. The number of KNN classifier is counted in the confusion matrix, which is difficult to measure the quality of the model. Therefore, the basic statistical results of the confusion matrix extend the following three indicators: Accuracy, Precision, Recall. With the above four metrics, you can convert the number of results in the confusion matrix into a ratio of 0-1, Facilitate standardized measurement, The confusion matrix is used to calculate the number of observations in the wrong class and the right class of the classification model respectively, and then display the results in a table[10].

The indicator is calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

TP: True Positive
TN: True Negative
FP: False Positive
FN: False Negative

Table.4 Mixing Matrix

Confusion matrix		True value	
		Positive	Negative
Forecast Value	Positive	TP	FP (Type II)
	Negative	FN (Type I)	TN

From the confusion matrix, we can draw the following conclusions:

Accuracy: In the test set of 921 news texts, we predicted a total of 658 samples, so the accuracy (Accuracy) = 658/921 = 71.4%.

Precision: The categorization shows that 161 of the 921 news texts are international news, but only 121 of them were correctly predicted. So Precision = 121 /161 = 75.2%

Recall: Take international news as an example. Of the 135 international news texts, only 121 are classified as international news texts. So, Recall = 121/135 = 89.6%

F1-Score: Calculated from the formula, for international news texts, F1-Score= (2 * 0.752* 0.896)/(0.752+ 0.896) = 81.77%

Experimental results show that the method used in this paper is more effective than the data normalization processing and data dimension reduction before.

4 Conclusion

A KNN-based Chinese and Lao text classification method is proposed, which uses data normalization and data dimensionality reduction to classify Lao news texts. Experimental results show that the method has achieved good results.

5 Acknowledgements

This paper is supported by National Nature Science Foundation No.61662040, 2016FB101

References

1 CHEN Weidi,QIN Yuping. Review of text classification methods based on machine learning[J]. Journal of Bohai University(Natural Science Edition), 2010, 31(2): 201-205.

2 Fan Cunjia, Wang Yousheng, Bian Hang. An Improved KNN Text Classification Algorithm[J]. Foreign Electronic Measurement Technology, 2015(12): 39-43.

3 HU Xue-Gang, DONG Xue-Chun, XIE Fei. Chinese Text Classification Method Based on Word Vector Space Model[J]. Journal of Hefei University of Technology(Natural Science), 2007, 30(10): 1261-1264.

4 Zhao Shuang. Research on Text Feature Selection Method Based on Domain Ontology[J]. Journal of Fujian Computer, 2016, 32(7): 41-41.
5 https://baike.baidu.com/item/%E9%80%86%E6%96%87%E6%A1%A3%E9%A2%91%E7%8E%87/11018305
6 Anonymous. Overview of KNN Algorithms [J]. Communication World, 2018, 341(10): 279-280.
7 Lin Haiming, Du Zifang. Problems that should be paid attention to in the comprehensive evaluation of principal component analysis[J]. Statistical Research, 2016, 30(08): 25-31.
8 Anonymous. Research on text classification method on unbalanced data sets[J]. Computer Engineering and Applications, 2013, 49(20): 118-121.
9 Research on Chinese text classification feature selection method [D]. Southwest University, 2010.
10 https://en.wikipedia.org/wiki/Confusion_matrix