MATEC Web of Conferences **44**, 01015 (2016)
DOI: 10.1051/matecconf/20164401015

A Chinese text classification system based on Naive Bayes algorithm

Wei Cui [1]

[1]*Department of Information Technology , Luzhou Vocational&Technical College, Sichuan Luzhou, 646005*

Abstract. In this paper, aiming at the characteristics of Chinese text classification, using the ICTCLAS(Chinese lexical analysis system of Chinese academy of sciences) for document segmentation, and for data cleaning and filtering the Stop words, using the information gain and document frequency feature selection algorithm to document feature selection. Based on this, based on the Naive Bayesian algorithm implemented text classifier , and use Chinese corpus of Fudan University has carried on the experiment and analysis on the system.

Keywords: Chinese word segmentation. Text categorization; Information gain; Naive Bayes algorithm

With large network retrieval system, document management, information filtering system, such as wide application , the growing importance of text categorization have been increasingly emerging。The common method to text classification are mainly Support Vector Machine method, the K-Nearest Neighbor , Naive Bayes theorem, Neural Net, etc. These methods in the English and European languages, automatic text classification had extensive research and achieved good results. However Chinese than English in word formation is more complicated, so the Chinese text classification and English text categorization in text preprocessing phase, there is a certain difference. This article is based on Naive Bayes algorithm and Chinese lexical analysis system ICTCLAS, design a suitable Chinese text categorization system.

1. Chinese text preprocessing.

We first for Chinese text preprocessing, including structure processing, word processing, to stop words and so on. Extract on behalf of the text characteristic of metadata (characteristics), saved in a structured form, as the center of the document representation.

1.1 segmentation preprocessing

In the organization of the text, Chinese is very different to the European and American language which is Represented by English. In western languages, the word and the word was separated by Spaces, don't need word processing. In the Chinese text, words are joined together.

Article available at http://www.matec-conferences.org or http://dx.doi.org/10.1051/matecconf/20164401015

Therefore, word segmentation is difficult. Chinese word segmentation technology is facing two biggest problems, the ambiguous segmentation and the unknown word recognition[1].

In this paper, use ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System) [2] to finish the job . The system adopts the combination of, statistical method and rule method, based on N - the shortest path to unambiguous text at the beginning of Chinese text segmentation, then through multiple hidden markov model algorithm to identify the word which is not log in, very good segmentation results have been achieved. System segmentation accuracy is as high as 97% above, not on word recognition recall rate is as high as 90%.

1.2 Data cleaning

After word segmentation processing of text, not all of the features are helpful to structure vector space model and classification, on the contrary, will make the word which without help for text classification as a feature , can cause great influence to the precision of classification. In addition, take out of stop words can largely reduce the number of feature item, has great help for text dimension reduction, so in front of the vector space model , to clean up the words thoroughly which are without help for classify. This system is based on the part-of-speech tagging for data cleaning. After dealing with the word segmentation, for punctuation w, partical u, conjunctions, prepositions p c, quantifier q cleaning, etc.

1.3. Take out the Stop words

Stop words mainly refers to the utterances, adverbs, conjunctions, pronouns, prepositions, interjection, quantifier, numerals etc, in the text . For example: the auxiliary "di, di, de, ba "; The adverb "very, very, very, all".

To get rid of the Stop words in technology implementation is not complicated, just build a stop words dictionary, will stop at each word after word segmentation and matching words in dictionary entry, if the match is successful, will remove the word.

2. the text feature selection

Feature selection is to improve the efficiency of text classification, reduce the computational complexity . Text feature selection is usually through the judgment of key . The commonly used methods are: document frequency, information gain cross entropy, mutual information, statistics and expectations, and so on . This paper uses the information gain and document frequency method of key judgment.

2.1 the information gain

Information gain is an evaluation method based on entropy，is defined as a feature in text information entropy difference before and after.

For the characteristics of w and c document category，the IG considered the appear frequency of w in document c to measure the information gain of w for c, the greater W for c information gain, the most important characteristics for category c items w, the greater the contribution. Information gain evaluation function definitions as shown in formula 1[3].

Among them, m said category number , $P(c_i)$ refers the probability of c_i such documents in the corpus , $P(w)$ said corpus contained in the feature item w document frequency, $P(c_i \mid w)$ said document contains feature when w belongs to the conditional probability of class c_i, $P(\overline{w})$ said the item does not contain features from the corpus of w document frequency, $P(c_i \mid \overline{w})$ said document does not contain feature when w belongs to the conditional probability of class c_i.

2.2 document frequency

Document frequency is the number of text feature item。The idea is: the DF value below a certain threshold of entry is low frequency words, they contain less or no category information. And get rid of the low-frequency

words from the feature space, reducing space dimension to improve the accuracy of classification.

3 text classification

Text classification use a tagged category of text data sets to train a classifier , this text data set is called the training set , then use the trained classifier of unmarked categories of text classification. Text classification algorithms, which are frequently used KNN algorithm, the SVM algorithm and Bayesian algorithm, etc.

On many occasions, Naïve Bayes classification algorithm is comparable to obtained with the decision tree classification algorithm and neural network, the algorithm can be applied to large databases, and the method is simple, high classification accuracy and speed. Theoretically, compared with all other classification algorithms, Bayesian classification has the smallest error rate, in its class under the premise of conditional independence assumption was set up，it is the best classification algorithm. In many cases, however, its class conditions independent simple assumption is not established, but the practice has proved that even so, it in many areas are still able to obtain better classification results.

Naive Bayes method [3] will break down training document into eigenvector and decision class variables. Assumed the characteristic vector of each weight relative to the decision variables is relatively independent, that is to say, each component independently on the decision variables. Naive Bayesian conditional independence assumption between attributes greatly simplify the calculation of joint probability, formula as shown in (2).

To construct a classifier module is the key to the training process module, using Bayes classification algorithm, structure specific classifier. Training process is generally more time-consuming, the system will training all text once, the characteristics of related information in the file. In training once，testing directly from the configuration file read relevant information without the need for training again, to save time.

Assume that the training corpus containing N text $D=\{D_1,D_2,\cdots D_n\}$, these texts are now owned by M text category $C=\{C_1,C_2,\cdots C_m\}$, training corpus set consists of L key text $W=\{W_1,W_2,\cdots W_L\}$.

When text D_i belongs to the category C_j, there is $P(C_j \mid D_i)=1$, Otherwise $P(C_j \mid D_i)=0$. If a given category variable text, the text category of prior probability is estimated according to the formula (3).

If use $F(W_k,D_i)$ expressing the number of occurrences of key word W_k in the text D_i, the key word W_k in the category C_j's probability estimation according to the formula (4).

Total number of key + frequently number of the eigenvalues in class C_j

Any text can be seen as a collection of a series of orderly arrangement of key, in the Naive Bayes model assumes that feature vector relative to the decision variables between each component is relatively independent, the category C_j generated in the probability of text D_i according to the formula (5).

According to the test text belongs to the text characteristic data computing the probability of each category, and in accordance with the largest probability to test text classification. Test text D_i belongs to the category C_j of probability according to the formula (6).

4. Experiments and analysis

The corpus for the experiments is Fudan university's Chinese corpus (part), divided into two parts, training corpus and test corpus. Training set is composed of a set of completed classification of text , used to summing up the characteristics of each category to construct classifiers . The classification of the test set is used to test the classifier effect is a collection of documents. The number of the training set and testing set text basic is a 2:1 ratio, such as table 1.

There are a total of 170 training document, system characteristic value of a total of more than 6000 after training. Feature extraction using information gain (IG), the dimensions of the feature extraction for 3000 d, the classes of recall, check, and the comparison of F1 value as shown in table 2.

As you can see, "traffic" and "sports" two categories of recall, precision and F1 value is 100%. "Economic" recall only at 77.8%, is relatively low. "Education" only 71.4% of precision, also is the lowest in all categories. This result is due to the two categories of training and test text similarity is relatively low.

Test documentation to extract 1000 d, 2000 d, 3000 d when the experimental comparison results as shown in table 3.

Visible, the classification effect of classifier with the increase of dimension increased after decreased first. Document the eigenvalue extraction dimension at 3000 d classifier classification effect is best.

Feature extraction using information gain (IG) and document frequency (DF) extract 3000 d experimental results as shown in table 4..

Visible, the classification of the document frequency (DF) effect to lag behind the information gain (IG) classification effect. In order to verify the effects of the feature selection algorithm for classification will be effected according to the characteristics of DF and IG choose after 30 words out before analysis.

Feature selection algorithm adopts DF, the organization form of word for < word: document frequency>, the result is:

"China:10 / the national:6 / protection:5 / environment:4 / Beijing:4 / American:4 / Britain:3 / Iraq:3 / Baghdad:3 / East Asia:3 / railway:3 / text:3 / word:3 / computer:3 / education:3 / japan:2 / continuous:2 / correct:2 / Shanghai:2 / involve:2 / correct:2 / transportation:2 / military:2 / Guangdong province,:2 / data:2 / World Cup:2 / bus:2 / Brazil:2 / sport:2 / Finance and economics:2"。

Feature selection algorithm USES the IG, the organization form of word is > < word: information gain, the result is:

"The national :0.0227 / protection:0.0210 / China:0.0183 / Baghdad :0.0174 / Iraq :0.0174 / Railway:0.0170 / East Asia:0.0165 / environment:0.0159 / text:0.0151 / education :0.0118 / aviator:0.01145 / navy:0.0114 / battlefield:0.0114 / military :0.0114 / world:0.0114 / ritain:0.0113 / teaching :0.0112 / international:0.0112 / traffic:0.0112 / Water transport:0.0112 / The World Cup :0.0109 / sports :0.0109 / Brazil :0.0109 / prosperity :0.0109 / International Monetary Fund:0.0109 / football team:0.010906 / Guangdong:0.01090 / operation:0.0107 / bleeding:0.0107 / solid :0.0107 / calculator :0.0106".

By comparison with the former 30 words, it can be seen that DF based text categorization effect to lag behind the IG based on the text classification effect. The lack of DF in that on the whole, the rare words may not rare in one type of text, and contains important information.

References

[1] Chen ping .svm-based Chinese text classification research and implementation of relevant algorithm [D] .at northwestern university in 2008

[2] The ICTCLAS official website http://ictclas.org/ 2010 6

[3]Yu fang .a web text classification system based on Naive Bayes method: web CAT computer engineering and application, 2004

[4] Wenhua, Dai .text classification and clustering based on genetic algorithm research of Beijing: science press, 2008

$$IG(w) = -\sum_{i=1}^{m} P(c_i) \lg P(c_i) + P(w)\sum_{i=1}^{m} P(c_i \mid w) \lg P(c_i \mid w) + P(\overline{w})\sum_{i=1}^{m} P(c_i \mid \overline{w}) \lg P(c_i \mid \overline{w}) \quad (1)$$

$$P((w_1, w_2, \cdots, w_n) \mid C_j) = \prod_{i=1}^{n} P(w_i \mid c_j) \tag{2}$$

$$P(C_j) = \frac{\sum_{i=1}^{N} P(C_j \mid D_i)}{N} \tag{3}$$

$$P(W_k \mid C_j) = \frac{1 + \sum_{i=1}^{N} F(W_k, D_i) P(C_j \mid D_i)}{L + \sum_{s=1}^{L} \sum_{i=1}^{N} F(W_s, D_i) P(C_j \mid D_i)} \tag{4}$$

$$P(D_i \mid C_j) = P((w_1, w_2, \cdots, w_n) \mid C_j) = \prod_{k=1}^{n} P(W_k \mid C_j) \tag{5}$$

$$P(C_j \mid D_i) = \frac{P(C_j) P(D_i \mid C_j)}{\sum_{j=1}^{M} P(C_j) P(D_i \mid C_j)} \tag{6}$$

Table 1. Experiment training set and test set

Classes	Enviro-nment	Comp-uter	Trans-portation	Edu-cation	Eco-nomic	Mil-itary	Sports	Med-ical
Training set	22	26	22	20	21	20	22	22
test set	11	12	11	10	9	9	11	10

Table 2. The results of classification algorithm

Evaluation	Category							
	environment	computer	traffic	education	economic	military	sports	medicine
recall ratio	81.8%	92.3%	100%	100%	77.8%	100%	100%	81.8%
precision ratio	100%	100%	100%	71.4%	100%	75%	100%	100%
F1value	90%	96%	100%	83.3%	87.5%	85.7%	100%	90%

Table 3. The extraction of different dimensional results more classification algorithm.

Dimension.	micro recall rate	Micro accuracy	Macro recall rate	Macro accuracy	Macro F1 value
1000	85.37%	85.37%	84.43%	87.39%	84.27%
2000	87.8%	87.8%	87.32%	89.60%	87.31%
3000	91.46%	91.46%	91.72%	93.3%	91.57%
4000	90.24%	90.24%	90.58%	92.17%	90.44%

Table 4 different performance comparison method to extract features

proposed method	micro recall rate	Micro accuracy	Macro recall rate	Macro accuracy	Macro F1 value
Information gain	91.46%	91.46%	91.72%	93.3%	91.57%
Document frequency	81.71%	81.71%	79.84%	84.75%	78.33%