A Vector-Based Algorithm for Chinese Text Classification[1]

Luo Chang ri
Department of Computer Science
Central China Normal University
Wuhan 430079,PR. China
Luocr2@hotmail.com

He Ting ting
Department of Computer Science
Central China Normal University
Wuhan 430079,PR. China
hett@163.net

Abstract

In this paper, vector-distance-weighted algorithm and representative-vector-distance algorithm are described and used to implement the process of automatic text classification. Two experiments have been done by means of the algorithms (experiment1 is based on vector-distance-weighted algorithm and experiment2 is based on representative-vector-distance algorithm). Characters are selected as features. The average precision of experiment1 and experiment2 is 80.36% and 69.27%, respectively. Comparing the two experiments, it can be concluded that the efficiency of text classification can be improved by means of vector-distance-weighted algorithm.

Keywords: text classification, vector-distance-weighted algorithm, Natural Language Process.

1. **Introduction**Text classification is to assign one or more appropriate class-types for a text based on its content. In western countries, the study about text classification and its relative fields begins as early as 60 or 70s last century, when Salton put forward the VSM (Vector Space Model) theory and the VSM is used successfully in application. The study on Chinese text classification by means of computer starts in the 90s. The research has achieved a lot, for example, FuDan University and Institute of Computing Technology of Chinese Academy of Sciences have tracked and studied the TREC test. Early on PeKing University and TingHua University have made a study of the technology of web classification on their search engine "Network Sky" and "The guide of Network" respectively [LIU Bin, HUANG Tie Jun, CHENG Jun, GAO Wen 2002]. The results of these fields are not as satisfactory as those in English research because of the uniqueness of Chinese language.

There are two methods for text classification. One is the rule method. This method is applied earlier. For example, the Construe [Church , K.W.Lisa.F.Rau 1995] is developed based on it. The other method is based on statistics, such as Bayes, VSM, KNN, SVM and so on. In the following section, the authors will discuss the expression of texts in VSM, computation of the feature weight, the similarity between text and classes, and the vector-distance algorithm. At the same time, the results of experiments are figured out and analyzed.

2. Vector Space Model

Vector Space Model (VSM) is widely applied in IR system for its simple conceptions and its simulation of close space to close meanings. The classification method used in texts is introduced from IR system.

2.1. Vector expression of training texts

A text is composed of characters、words and phrases which are termed as the features of text. According to "Bayes hypothesis", presuming the effects of features to the class adscription are

[1] This paper is supported by Nature Science Foundation of Hubei, China (ID:2001ABBo12)

independent, the text can be expressed as the vector of feature collection. After the handling of the training set, "Term-Documents" matrix space A_{txd} could be obtained. The row and column are expressed by the feature and text respectively. So we can get the training set vectors space showing in figure1.

	doc$_1$	doc$_2$	doc$_3$............doc $_d$
t$_1$	a$_{1,1}$		
t $_t$	a$_{1,t}$	....	...a$_{t,d}$

Figure1.

In general, A is a sparse matrix, it can be compressed by using the method described in [Zhan xue gang, Lin Hongfei, Yao Tianshun 1999].

2.2. Feature weighting

From section 2.1, the matrix A_{txd} can be inferred .If $a_{i,j}$ is used to express delegate the non-zero element, then $A=[a_{i,j}]_{txd}$. In order to show the importance of the features in texts, $a_{i,j}$ can't be expressed by the frequency of the features that occur in texts. Generally, feature weight is computed and normalized. TF-IDF model is used to compute the term weight as follows: [2]

$$a_{i,j}=\frac{\text{Local(i, j)} * \text{Global(i)}}{\sqrt{\sum_{i=1}^{t}\left(Local(i,j)*Global(i)\right)^2}} \text{.......(1)}$$

Where, $a_{i,j}$ is the term weight; $Local(i,j)=\log_2(1+tf_{i,j})$, $tf_{i,j}$ is the frequency of term i in document j. $Global(i)=\log_2((n/df_i)+1)$, n is the number of the train set, df_i is the number of documents containing term i.

A new document, formula (1) can be used to calculate the term weight. Because n=1 and df_i=1, so $Global(i)=1$. In the experiments, formula (2) is used to compute the term weight and formula (3) is used to normalize it:

$$S_{i,j}=10*\left(\frac{1+\log(tf_{i,j})}{1+\log(l_j)}\right) \text{.......(2)} \quad S'_{i,j}=S_{i,j}\Big/\sqrt{\sum_{i=1}^{t}\left(S_{i,j}\right)^2} \text{.............(3)}$$

Where, $S_{i,j}$ is the weight of term i in document j, l_j is the length of the document j, and the

[2] about other model please read the paper[Diao Qian, Wang Yongcheng, Zhang Huihui, He Ji 2000] and the chapter 15 of [Christopher D.Manning, Hinrich Schutze]

frequency of terms in document is summed up as the document length.

2.3. Feature selecting

There are multi-selection for features, such as characters, words, phrases or their compounds. It is universally accepted that using words as features is superior to other selections. Words can be extracted directly from texts. Because automatic extracting of words is not satisfactory, segment training set is then the first choice. After segmentation, mutual information method can be used to filter the feature. In literature [LIU Bin, HUANG Tie Jun, CHENG Jun, GAO Wen 2002], the authors employ characters and characters plus words as the features respectively, and compare the results. The results improved little as shown, after adding more than 200 000 words as features. So, employing the characters as the features for text classification study is meaningful.

In experiments, characters are selected as the features, but not all the 6763 Chinese characters defined in GB2312-80. The characters of training set are selected as features. The number is less than GB2312-80. There are 5468 characters in the training set.

3. Vector-distance algorithm

There are many algorithms which are based on vector space, such as Support Vector Machine, Nerval Network, KNN, Bayse, Vector-distance and so on. In this paper, Vector-distance algorithm is employed. The simple vector-distance algorithm is used in the vector space model. Simply speaking, this algorithm is used to compute the vector distance between the document to be classified and the classes of training set. Two methods are used in the experiments.

Method I : vector-distance weighted algorithm.

This is an algorithm to compute the weighted similarity, that is, handling every text of training set to get training set matrix, handling the text to be classified to get vector, computing the similarity between the texts in training set and the text to be classified, and then weighting the similarity, if the training texts have the same class. The main steps are as follows:

Step1: Formula (1) is used to handle the training set text vector to get the training set matrix space;

Step2: Formula (2) is used to handle the new document vector, while formula (3) is used to normalize the vector;

Step3: Formula (4) is used to figure out the similarities;

$$Cos(d_i, d_j) = \sum_{k=1}^{n} \left(a_{i,k} \times s'_{j,k}\right) \Big/ \sqrt{\left(\sum_{k=1}^{n} a_{i,k}^{2}\right) \times \left(\sum_{k=1}^{n} s'^{2}_{j,k}\right)} \text{.......... ....}(4)$$

Where, d_i is the feature vector of the new text, d_j is the vector of jth document in training set, n is feature number;

Step4: The same class training set texts are judged in order to calculate the weighted similarities, using formula (5):

$$SumSim\left(d_i, C_j\right) = \sum_{t=1}^{n} Cos\left(d_i, d_t\right) * T\left(d_t, C_j\right) \text{......... ......}(5)$$

where, d_i is the new document, $Cos(d_i, d_t)$ is as the same as the formula (4), n is documents' number of the collection , $T(d_t, C_j)$ is the class determine function. Suppose d_t belongs to class C_j, then $T(d_t, C_j)=1$, otherwise $T(d_t, C_j)=0$.

Step5: The classes are sorted in descending according to the similarities computed in step4, and are outputted; the first one is the class the new document will belong to.

Method II: representative-vector-distance algorithm

In this method, the representative vector vc_j is formed by the mergence of every same class training texts. When there is a new document, construct a vector d_i for it. Then calculate the cosine-distance (similarity) of d_i and vc_j, then sort the similarity and output the result, the main steps are as follows:

Step1: Handle every kind of texts of the training set to get all the kinds of representative vectors. Furthermore, get the representative vector matrix of the whole collection, to get the normalization matrix, using formula (1);

Step2: The vector of new document is gotten by using formula (2) and (3);

Step3: The similarities are calculated by using formula (4);

Step4: At last, get the class that the new document belongs to, according to the size of the similarities.

4. Evaluation

The effectiveness of text classification is measured as Recall and Precision calculated by the following equations:

$$\text{Precision} = \frac{\begin{array}{c}\textit{the number of classification that are}\\ \textit{correctly assigned to documents}\end{array}}{\textit{the number of training set}},$$

$$\text{Recall} = \frac{\begin{array}{c}\textit{the number of classification that are}\\ \textit{correctly assigned to documents}\end{array}}{\begin{array}{c}\textit{the number of classification that are correctly}\\ \textit{assigned to documents in whole training set}\end{array}}.$$

In order to synthetically consider the effectiveness of text classification, F1 test value is used as follows:

$$\text{F1 test value} = \frac{\Pr ecision \times \operatorname{Re} call \times 2}{\Pr ecision + \operatorname{Re} call}$$

5. Results and Discussions

In experiments, the authors employ the Modern Chinese Corpus of State Language Commission as training set and use vector-distance algorithm to implement the process of automatic text classification. The corpus is a balanced corpus and has been manually classified. It includes three parts, namely: Human& Society Science, Natural Science and Integration. Politics, history, society, economy, arts, literature, military affairs& gym, and life, are the 8 classes included in the first part; mathematics & physics, biology & chemistry, astronomy & geography, medicine& sanitation, agriculture &forests, ocean & weather are included in Natural Science, and Integration part contains application documents and others two classes. The size of every text is about 3-4kB, and the content of the texts is selected from newspapers, books and general magazines.

In the experiments, 11 classes are randomly selected, including literature-colloquialism, politics-law, society-education, mathematics & physics, biology & chemistry, military affairs& gym, astronomy & geography, medicine& sanitation, arts, agriculture &forests, ocean & weather, from each of which 100 passages are selected. The total amount is 1100. Half of them are selected as training set, and the rest for test. Two experiments have been done with them separately. Experiment 1 is done with method I, and experiment 2 with method II. The results of experiment 1 and experiment 2 are revealed in table 1 and table 2 respectively [Note: Literature is for Literature-colloquialism, politics is for politics-law, society is for society-education, mathematics is for mathematics & Physics...and so on].

Table1. The data of experiment 1

	Right	Recall%	Precision%
Literature	39	78	98
Politics	49	98	87.46
Arts	16	32	93.82
Medicine	18	36	94
Astronomy	38	76	96.55
Mathematics	40	80	97.82
Biology	23	46	94.91
Society	44	88	86.18
Agriculture	20	40	94.55
Military	4	8	91.64
Ocean	50	100	89.09

Table2. The data of experiment 2

	Right	Recall%	Precision%
Literature	25	50	95.46
Politics	50	100	83.09
Arts	11	22	92.09
Medicine	12	24	93.09
Astronomy	33	66	95.46
Mathematics	50	100	78.18
Biology	5	10	91.82
Society	32	64	89.64
Agriculture	14	28	93.46
Military	7	14	92.18
Ocean	34	68	95.27

Table 3. The comparison of experiment 1 &experiment 2

	Correct pages	Precision	Recall	F1 value
Experiment 1	341	62%	62%	62%
Experiment 2	273	49.64%	49.64%	49.64%

It is obviously shown in the above figures that:

1. The effectiveness of experiment1 is better than that of experiment2. That is to say, the merging of the same class training texts to get representative vector for classification effectiveness does not work well.
2. In experiment1, the classification effects of literature-colloquialism, politics-law, society-education, Mathematics & physics, astronomy & geography, ocean & weather are better and more stable than others, especially politics-law, society-education, mathematics & physics, ocean & weather. The Recall of them achieved as 80% or even more. The main reason is that the features of these classes are distinct, thus little influence from other overlapping classes is found.
3. The Recall of biology & chemistry, military affairs& gym, medicine& sanitation, arts and agriculture &forests is low, especially that of military affairs& gym, only about 10%. After analyzing the corpus and the outcome of the experiment on these classes, it is found that:

   ① The boundaries of classes are seriously overlapped. The discrimination becomes lower, when using the characters as the features.

   ② Some samples in corpus are excerpts of one or several passages. Then, it is likely that the title of the sample is about medicine and is classified into medicine& sanitation, however, the content of the text is about chemical components and chemical reactions of the medicine, so the text is judged as biology & chemistry class by machine.

   ③ The corpus of military affairs& gym contains military affairs and gym. After the corpus is analyzed, it is found that the majority of military documents contain historical and military facts, and they overlap with politics-law. The samples of gym overlap with those of medicine& sanitation and education in content. The samples which overlap with other classes' samples are classified correctly, but they are classified wrongly by machine. The same reason is found in other classes whose Recall is low.

For above problems, the effectiveness of case ① can be improved by means of changing the feature selection, such as selecting word as the feature. As for the case ②, it is difficult to improve the effectiveness by statistical method only. Semantic comprehension should be added to help us improve the classification effectiveness.

4. The authors select characters as features and don't filter the feature set, this probably leads to noise [Zhan xue gang, Lin Hongfei, Yao Tianshun 1999, Schutze H, Hull D, Pedersen J. 1996, ZHOU Shui-Geng, GUAN Ji-Hong, HU Yun-Fa, ZHOU Ao-Ying 2001], for example, the influence from number, and make the classification effectiveness decline. In the experiments, after the handling of the texts to Unicode, then every number code becomes a feature, so the function of number is larger than before.
5. The data of table 1、2、3 are the numbers of successfully classified documents, and each of them has the largest similarity with training set class, and into which they are classified. In the analysis of the corpus, in each class it is found that the categories of some texts judged by

machine are different from their original categories which are classified manually. The results judged by machine are regarded as correct, by means of artificial discrimination. Therefore, the results of classification are correct if they are of this case. According to this principle, the number behind "+" indicates the increasing correct result.

Table4. The adjusted data of experiment 1 (based on table 1)

	Correct	Wrong
Literature	39+1	10
Politics	49+1	0
Arts	16+11	23
Medicine	18+3	29
Astronomy	38+1	11
Mathematics	40+1	9
Biology	23+2	25
Society	44+4	2
Agriculture	20+1	29
Military	4+23	23
Ocean	50	0
Total	341+48	161

Table5. The adjusted data of experiment 2 (based on table 2)

	Correct	Wrong
Literature	25+2	23
Politics	50	0
Arts	11+12	27
Medicine	12+5	33
Astronomy	33+7	10
Mathematics	50	0
Biology	5+16	29
Society	32+9	9
Agriculture	14+7	29
Military	7+19	24
Ocean	34+7	9
Total	273+84	193

Table6. The comparison of experiment 1 &experiment 2 after the first adjust

	Correct pages	Precision	Recall	F1 value
Experiment 1	389	70.73%	70.73%	70.73%
Experiment 2	357	64.91%	64.91%	64.91%

6. According to the analysis of the experiments, the result of classification in military affairs & gym, medicine & sanitation, arts, biology& chemistry, agriculture& forests is unsatisfactory. It is partly due to the overlapping of the feature set or class boundary. According to the results of experiments and analysis of the corpus, the authors admit a document can be classified into more than one class. So, in class similarities sort list, the first-three classes of the new document are observed. If any one of them among the three is the same as the original class, it means that the classification of the new document is successful. (This adjustment is only for military affairs & gym, medicine & sanitation, arts, biology& chemistry, and agriculture&

forests.) The reasons to do so are: first, the documents in corpus may have several classes; second, the cosine distance discrepancy is very small among the first-three when compared with others; third, using this method, setting the threshold can be avoided. Table7、8 are related data.

Table7. The data of experiment 1 before and after the second adjustment (based on table1)

	Arts	Biology	Military	Medicine	Agriculture	Correct	Precision	Recall
Before adjust	16	23	4	18	20	81	62%	62%
After adjust	48	37	33	30	34	182	80.36%	80.36%

Table8. The data of experiment 2 before and after the second adjustment (based on table2)

	Arts	Biology	Military	Medicine	Agriculture	Correct	Precision	Recall
Before adjust	11	5	7	12	14	49	49.64%	49.64%
After adjust	38	34	32	24	30	158	69.27%	69.27%

The results in table7、8 is better, including that of literature-colloquialism, politics-law, society-education, mathematics & physics, astronomy & geography and ocean & weather, of which the data are not adjusted yet. The average precision of experiment1 and experiment2 is 80.36% and 69.27%, respectively.

Comparing the two experiments, it can be concluded that the efficiency of text classification can be improved by means of vector-distance weighted algorithm.

6. References

[LIU Bin, HUANG Tie Jun, CHENG Jun, GAO Wen 2002] A New Statistical-based Method in Automatic Text Classification, Journal of Chinese Information Processing Vol.16 No.6.

[Church , K.W.Lisa.F.Rau 1995] Commercial Applications of Natural Language Processing , Communications of ACM,Vol.38.No.11

[Zhan xue gang, Lin Hongfei, Yao Tianshun 1999] Hierarchical Method for Chinese Document Classification. Journal of Chinese Information Processing Vol.13 No.6 1999.

[Diao Qian, Wang Yongcheng, Zhang Huihui, He Ji 2000] Term Weighting and Classification Algorithms. Journal of Chinese Information Processing, Vol.14 No.3 2000.

[Christopher D.Manning, Hinrich Schutze] Foundations of Statistical Natural Language Processing. The MIT Press Cambridge,Massachusetts London, England.

[Schutze H, Hull D, Pedersen J.1996] A Comparison of Selective Bayesian Network Classifiers.ICML-96, 1996.

[ZHOU Shui-Geng, GUAN Ji-Hong, HU Yun-Fa, and ZHOU Ao-Ying 2001] A Chinese Doucument Categorization System without Dictionary Support and Segmentation Processing, Journal of Computer Research & Development Vol.38 No.7 2001.