Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(7): 1636-1643

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

The naive Bayes text classification algorithm based on rough set in the cloud platform

Yugang Dai and Haosheng Sun

Northwest University for Nationalities, Key Lab of China's National Linguistic Information Technology Lanzhou, China

---

ABSTRACT

This paper improves the naïve bayesian classification algorithm , combining with the rough set theory we can get a naive bayesian classifier algorithm based on the rough set. We implement this algorithm on a cloud platform using map-reduce programming mode and get a excellent result. A recall rate of 76.4 was achieved when classifying Tibetan Web pages .

Key words: Cloud platform; parallel computing; map-reduce programming mode; naive bayes algorithm based on the rough set

---

INTRODUCTION

Along with the national financially supporting on the Tibetan information infrastructure, and the rapid development of Tibetan economy, the scale of Tibetan netizen is also constantly growing. Automatic classification of large-scale Tibetan Web document can improve the disorderly situation of webpage text information, so as to reduce the query time, improve search quality.

At present there are a lot research about Chinese web document classification , and usually we use the classification , clustering algorithm and so on in the research process . This paper we mainly use the classification algorithm . Classify various types of information in the Tibetan web so as to get the useful information . Nowadays the commonly used classification algorithms are the decision tree , the k-nearest neighbor , the support vector machine(SVM) , the vector space model(VSM) , the naive Bayes algorithm and the neural network algorithm , among which we choose the naive Bayes algorithm this paper . Compared with other algorithms , the naive Bayes algorithm simple and practicable .

The algorithm work well in a small amount of data in some language such as Chinese. But the technology on Tibetan is not so perfect , When applied to large data the efficiency even worse . Nevertheless during the big data time , the ability of obtaining large-scale data processing and improving the calculation efficiency is the main issues needed to be resolved about webpage text classification technology. Many experts are all devoted to this direction . Eventually combined with the newest frontier science named cloud technology propose the method of parallel processing. Put the massive data segmentation into a plurality of small amounts of data , and in general use the map-reduce method[5] . It handles the large data sets through the cluster, and establishes a mainstream parallel data processing model in the cloud platform.

THE ANYLYSIS OF COMPUTING TASK

The naive Bayesian classifier algorithm has some defects. Here to improve this model we introduce the rough set theory, establish the model of the naive Bayesian classifier algorithm basing on rough set . In the study of the naive Bayesian classifier algorithm[1] ,[2]basing on rough set[3] theory , through introduces the knowledge reduction basing

on information entropy to improve the limits of conditional independence assumption , Put forward the algorithm of approximate attribute reduction basing on information entropy , then proposes the naive Bayesian classifier algorithm basing on rough set . The experiment proves , this algorithm is better than the naive Bias classification algorithm[18] in classification ccuracy .

The maximum independent attribute reduction algorithm [6] , deals with the null value attribute in the decision table , eliminate the redundant attributes and the rely solely on attributes , then selected partly attribute set and delete . All of this are in the premise of not affecting the classification ability . Through the above operation improves the limitation of the condition attribute independence , in line with the requirements of naive bayesian classifier[7],[8] ,[9]. So we get a naive bayesian classifier algorithm based on the rough set [4]. Its input includes the data set , the condition attribute , the decision attributes and the output is the classification results .To quickly recap , improve the naive Bayes text classification algorithm via the learning of rough set [17],[19], then apply in the cloud platform , using the map-reduce method by means of the parallel processing . In order to further study , the next is a simply analysis of the computation task .

Calculate the maximum posteriori probability C_{mpp} and get the best classification document discrimination. If there are M classifications, we are required to calculate the M posterior probability of document to be classified. Take the maximum posteriori probability C_{mpp} ,The correspond class name as the category of the text to be classified. The calculated of posterior probabilities completes by the map-reduce parallel processing model[10],[11].The calculation formula of the posterior probability is as follows：

$$C = \ln P\left(C_j\right) + \sum_{k=1}^{r} \ln P\left(\omega_{ik} \mid C_j\right) \qquad (1)$$

Completing the calculate of type (1) can do the work of the following statistics and computing using the map reduce formula.

The basic statistical work : The statistics of each sample term frequency ; Count the document frequency of feature word (DF) ; Count the total number of feature word vocabulary(VC) ; Count the label number of each class(LC);Calculation the normalized term frequency of each feature words, The calculation method is $NTF = \dfrac{\ln\left(1+D_k\right)}{\sqrt{\sum_k D_k^2}}$;Calculate the inverse document frequency feature words, $IDF = \ln\left(LC/DF\right)$;The calculation of weight normalized term frequency on feature words in the class, $T_{ik} = WNTF\_IDF = NTF \times IDF$.;Calculate the sum of normalized weight by all the feature words in a class, $FS = \sum_k T_{ik}$.

Through the above computing can get a naïve Bayesian classification model based on rough set, which describes by a sparse matrix and two vectors. The each row of sparse matrix Q corresponds to a feature words, and each column of it corresponds to a class, the matrix element is the normalized weight values of the corresponding feature words. Vector V express the sum of normalized weights by one feature word in all the class; and vector N express the sum of normalized weight by all the feature words in a class . From the above analysis , put the text data quantification , to lay the foundation of the following calculation .

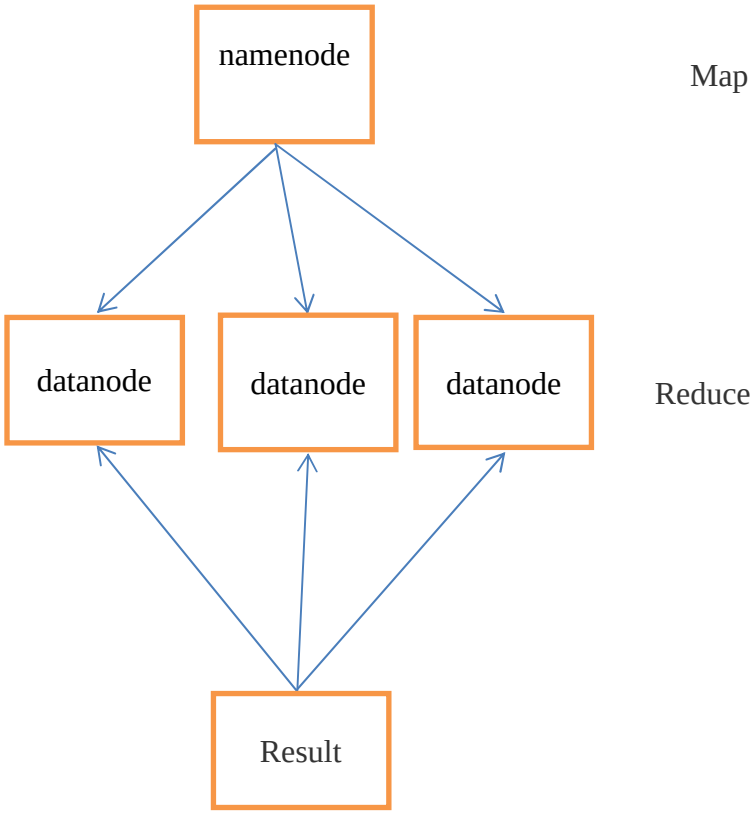

Figure 1 Distributed algorithm process

THE MAP-REDUCE PARALLELIZATION OF CLASSIFICATION ALGORITHMIC

We have a certain understanding of the theory from the above analysis , then we analyze the calculating process concretely , mainly include the implementation of text preprocessing , the generate of naïve Bayesian classification model based on rough set , the map-reduce parallel method of classification (test) stage and so on .

A The implementation of text preprocessing

The object of text preprocessing includes two categories, One is the corpus webpage texts which are collected and artificial categories annotated in advance, using in training and testing stages; the other kind is the webpage to be classified which is crawls down from the wap site.The target of the text preprocessing: Put the document into a text input format of the naïve Bayesian model based on rough set. The features is each line of the text file representing a sample text, that is a sample text emerging in the file in the form of row. The organizational structure is "Label+(\Table)+Term_1+(Blank-Space)+…+(Blank-Space)+Term_R". Input: The key is the path to the document name, the value is webpage, The content of training set or test set.

The map job: The document treatment on the training set and the test set need to parse the pathname, and put the directory name as the class name; For all the values to be deal with call the Tibetan word segmentation tool; the output of intermediate result is<class name or URL,{ Term_1+ (Blank Space)+…+ (Blank Space)+ Term_R }>;

The reduce job: Put the intermediate results with the same key into a file ,when dealing with the training set and test set document. Such a category corresponding to a file; Ordered by URL with the webpage document to be classified, and the output is a large file, then block storage according to the file size.

B The generate of naïve Bayesian classification model based on rough set

1)Statistical information job

The task is to complete the four values of the statistics: The frequency of feature words in each class, the document frequency of feature words appear in a class, the feature types and the number of documents in a text. The input is the text input format of the naïve Bayesian model based on rough set in the preceding stage: put the class name label

as the key, following by feature words set which has already segmented as the value.

The map job: Count the frequency of each feature words occurs in each line of text D_k ;Calculate the normalization denominator of term frequency $D_k : LN = \sqrt{\sum_k D_k^2}$;Calculate the normalized frequency of feature words in a document; Output the intermediate results of normalized frequency: its key is constituted by Label and Token, which is emit:<(Label, Token), NTF>;A feature word appears in a document of class, then output the intermediate results<(Label, Token), 1>,the key is constituted by Label and Token; Output the intermediate results when there appears a feature word<Token，1>,the key is token; The statistics of document numbers belong to a class, output the intermediate results,<Label, 1>,the Key is Label;

The reduce job: As to the <(Label, Token), NTF> set, if (Label, Token) is the same, then sum by the NTF, The results are the sum of normalized frequency of each feature word in all documents by the given class ;As to the <(Label, Token), 1> set, if (Label, Token)is the same, then sum, and get the number of feature words in a document; As to the <Token, 1> set, if Token is the same, then output<Token, 1>,so the next map-reduce job can count all the Token numbers; As to the <Label, 1> set, if Label is the same, then sum, and you can get the numbers of document containing in a class.

2)The job of calculating normalized weight
The primary mission of the Job is to compute normalized weights value of each feature word in each class. Its input is the four output files of the above statistical information job.

The map job: Obtain the total number of documents containing in one class; Obtain the times of feature words appearing in the document; Calculate the standard IDF value of feature words, and output the intermediate results<(Label, Token), IDF>; In the <Token,1> set, if Token is the same, then output the intermediate results <Vocab_Count,1>.

The reduce job: As to the intermediate results <(Label, Token),IDF> and the <(Label, Token),NTF> output of the above job. If the key value (Label, Token) is the same, then calculate $WNTF\_IDF = NTF \times IDF$, and get the normalized weights value of a feature word in a class, then output $<(Lable, Token), WNTF\_IDF>$; As to the intermediate results<Vocab_Coun t,1>, the key value Vocab_Count are all the same, directly sum the value, and we can obtain the total number of feature words vocabulary in the training.

3)Weight summary job
The weight summary job is to subtotal the normalized weights of feature words in a certain class.

The map job: Its main task is to output the input key by classify, so as to summarize on the reduce stage, for each input key value pairs $<(Lable, Token), WNTF\_IDF>$, outputs the intermediate results respectively $<Token, WNTF_IDF>, <$Label$, WNTF_IDF>$, $<Anyone, WNTF\_IDF>$.

The reduce job: Sum the key value of 'Label', 'Token', 'Anyone' respectively, we get the normalized weights sum of a feature word in all categories, output <Token, FS>;and the normalized weights sum of all feature words in a class feature word feature word, output <Label, LS>;and the normalized weights sum of all the feature words in all the class, output <Total_Sum, TS>.

C The map-reduce parallel method of classification (test) stage
Preprocess the document to be classified, organize them into a input text format of the naïve Bayesian model based on rough set. Each document to be classified in the text appears in the form of row. The execution mechanism of map-reduce is the input divides into a plurality of data blocks according to the size of the input file, and parallel computes the corresponding tasks through multiple mapper and reducer.

The map job: Load the naïve Bayesian classification model basing on rough set; As to each feature word in the input Key value pairs <Label, Token>,take out its relevant parameters to calculate the class conditional probability in the naïve Bayesian classification model basing on rough set, and calculate the posterior probabilities of the given document on all the class; Use the category corresponding to the maximum posteriori probability as the input document categories. If proceeding the test task, then output<(the original mark before calculate, the class name to

have identified),1.0>; If proceeding the classification task, then output<(the original mark before calculate, the class name to have identified), >.

The reduce job: As to the documents to be classified, directly output the intermediate results; for the test set document, to the intermediate results <(the original mark before calculate, the class name to have identified),1.0>. If its key is the same, then sum the value, the output is easy to generate confusion matrix, and conduct the classification evaluation.

Above is the calculation of the naive Bayes text classification algorithm based on rough set in the cloud platform . The key is the interface part between the map stage and the reduce stage . Timely feedback and verification to get the best classification results . Figure 2 is the text classification process , just shows the main calculation steps and give an intuitive understanding of the algorithm .

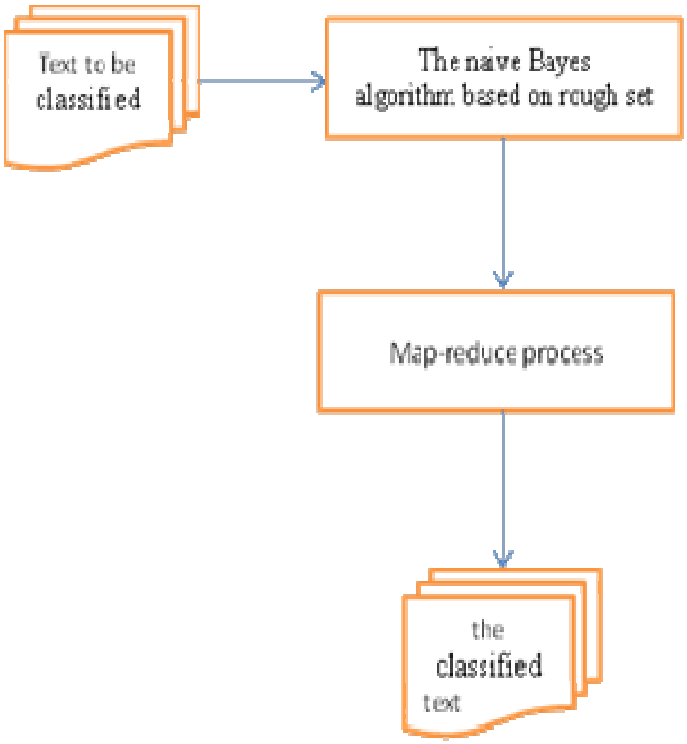

Figure 2 Text classification process

THE EXPERIMENT SITUATION

A Comparative experiments of one machine

Comparison the performance of processing the same size data on the same hardware configuration environment between a computing node in the hadoop[12],[13] cluster and the serial software of naïve Bayesian classification algorithm basing on rough set. In the experiment we uses the weka system[14] and choose 20 newsgroups data sets, which collects 328 Tibetan sample documents. Then put them in 20 folders separately. And the total file size is 26.8 MB. Both of the comparative experiments, the scale of data changes from small gradually increase to big, making the cross-validation experiments.

The experimental results show : When the input data is small the processing efficiency of hadoop computing nodes are below the non-parallel computing method . This is because completing the naïve Bayesian classification algorithm basing on rough set on hadoop cluster needs to complete multiple map-reduce operations, and the start or interaction of each operation requires a certain degree of resource consumption.

B Experiments on the cloud platform

The configuration structure of cloud platform: A machine as the service host node of namenode and job tracker. The other six machines as services slave node of datenode and task tracker. The hardware configuration of each node are as follows: Intel Xeon X3330 CPU,8 GB RAM, 2 TB SATA hard disk and the onboard Intel dual Gigabit network

controller. According to the method of hadoop project on the official website configurates clusters, basing on the Hadoop0.20.2 version.

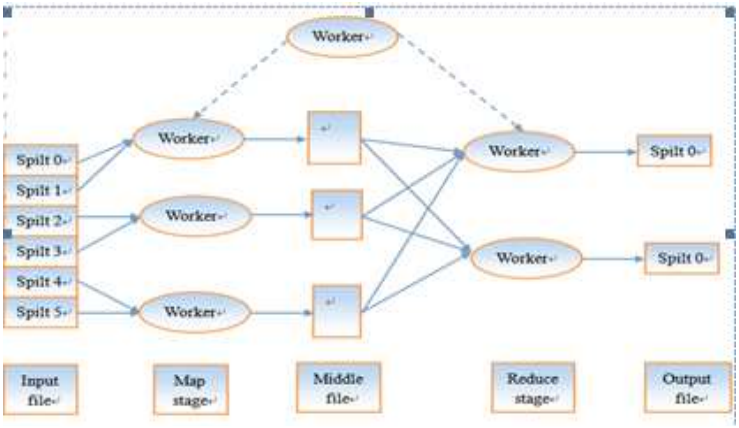

Figure3 Map-reduce execution framework

Figure 3 is the map-reduce execution framework. For this input file is too large , we decide the input file into five files just as shown in the figure 3 , then give them to three workers in the map stage , through calculate we obtain the intermediate results , and output it to the middle file , and next we give them two workers in the reduce stage , after calculate we get the final result , and output it to the output file . This is a simple process of map-reduce , and some actual experiment is much more trivial than this . In this paper we just use the above execution and obtain much more satisfied l results .

Experimental data : Search corpus in Tibetan website , and making the training and testing . for the time being, we take sports , tourism , the Tibetan Medicine (TM) , education, the Tibetan Buddhism (TB) , recruitment , culture and so on , all these seven classes as an example . Each category has 600 documents , and has a total of 80 MB in size after decompression .

Table 1: experimental data

Subject category	sports	tourism	T M	Education	TB	Recruitment	culture
The number of text	100	50	100	50	50	100	150

Shown in the table 1 we collect 50 tourism messages , 50 education news , 50 Tibetan Buddhism information , 100 sports news , 100 Tibetan medicine messages , 100 recruitment information and 150 culture messages , through analysis these records we can get a lot useful information .

Table 2 The overall experimental results

algorithm	The naive Bayes algorithm based on rough set	The naive Bayes algorithm
The recall rate	76.1%	73%
The average accuracy	77.6%	74.9%

The classification results : The confusion matrix of test output shows in table 2 . In the naive Bayes text classification algorithm based on rough set model the recall rate is 76.1% and the total classification recognition rate is 77.6% , and in the naive Bayes text classification algorithm model the recall rate is 73% and the total classification recognition rate is 74.9% . We can see through improving the algorithm both the recall rate and the efficiency are significantly improved .

Below we analysis the seven items partly by the recall rate and the efficiency , so that we can get the detailed information on specific ,with which the Tibetan people can take effective measures to the wanted news .

Table 3 Comparison of two algorithms for the recall rate

Subject category	sport	tourism	T M	Education	TB	Recruitment	culture
The original algorithm	76%	72.5%	72.3%	71.6%	69%	68.1%	68.6%
The improved algorithm	79%	75.1%	74.7%	73.8%	72%	72.5%	71.2%

Table 4 Comparison of two algorithms for the accuracy

Subject category	sport	tourism	T M	Education	TB	Recruitment	culture
The original algorithm	76.1%	73.2%	72.5%	70.3%	71.6%	73.1%	72.8%
The improved algorithm	79.4%	76.4%	74.2%	73.2%	74.3%	75.4%	76.1%

Shown in the table 3 and table 4 are the recall rate and the efficiency by the seven items . For the recall rate , the culture class is the lowest , and the most are mistaken for education . For the classification accuracy we see , the classification accuracy of education is the lowest , and most of them are taken as employment . The factors influencing the classification recognition rate are as follows:

Have relationship with the formula selection of key parameters $P\left(\omega_{ik} \mid C_j\right)$ in the naïve Bayesian classification model basing on rough set; Affect by the Tibetan word segmentation tool performance, there is no special feature words selection work in this study, so the stop words provided by segmentation tool has great influence on the results; Associating with sample quality in the corpus, there should be a greater discrimination between each class.

By comparing the two algorithms above , we can know the improved algorithm are greatly improved , either in the running time or the efficiency . Of course, the improved algorithm has his shortcomings. We can further research this to get a better conclusion .

C The Improve of algorithm

Although the naive Bayes text classification algorithm based on rough set in the cloud platform already have been improved , it also have some problems , and it can still be further improved and perfected .

The rough set theory we use in this paper is very simple . After a thorough study can further expand the application range and precision of the model . Of course, we can also combine the naive Bayes text classification algorithm with other theories . Moreover the amount of data we select in this paper is not enough , and we should use of more and larger amount of data , which can improve the accuracy much more . The last but not the least , the use of cloud technology is not so well . We should further study the distributed parallel processing technology of map-reduce method [15].

CONCLUSION

Through comparing the above two algorithms we know The naive Bayes text classification algorithm based on rough set in the cloud platform greatly enhance the calculation efficiency , for the sample size is not big enough , the difference of calculation time is not obvious , but still improved . The Tibetan netizen also obtain the desired information much more quickly and effectively through the above classification algorithm .

In a word , this is paper through study the map-reduce parallel method of the naïve Bayesian classification algorithm basing on rough set on the hadoop[16] cloud platform, improves mass data processing and the calculation efficiency in the naïve Bayesian classification algorithm basing on rough set. The experimental results show: the naïve Bayesian classification algorithm basing on rough set, running on the hadoop clusters after the map-reduce parallelization. It has a better speedup ratio, and gains a high recognition rate on the Tibetan web page classification .

Acknowledgments
This work is supported by Science and technology major special project in Gansu province (1203FK DA033) , The central college funds under Grant (NO.ycx13014) .

REFERENCES

[1] ZHANG Yaping，CHEN Debao，HOU Junqin， *Computer Engineering and Applications*，**2011**，47（15）：134-137.
[2] Wang S C，Yuan S M. *Journal of Software*，**2004**，15：1024-1048.
[3] Wei W，Ying T.A generic neural network approach for filling missing data in dataming[C] *// Proceedings of the IEEE International Conference on Systems，Man and Cybernetics*，**2003**，1：862-867.
[4] Yanqin Zhang , Xibei Yang . *Journal Of Software*, **2012**.3:551-563.
[5] Xuhui. *Data base* **2013**.9:98-99.
[6]Verbeek J J，Vlassis N，Nunnink R J. *Neural Computation*，**2003**，15（2）：469-485.

[7]Tao Dong, Wenqian Shan. *Journal Of Software*.**2011**.9:1837-1843.
[8]Qian Zhu, Yingying Zhang. *Journal Of Computers*,**2013**.51200-1206.
[9]Yang Yang, Xiang Long, Bo Jiang. *Journal Of Computers*,**2013**,10:2648-2655.
[10] Tong Yang, Ben-Chang Shia. *Journal Of Software*,**2012**.10:2189-2195.
[11]Lizhe Wang, Jie Tao, Gregor von Laszewski, Holger Marten. *Journal Of Computers*, **2010**.6:958-964.
[12]Dongbo Liu, Peng Xiao. *journal of software*,**2013**.4:761-767.
[13]Guoyuan Lin, Yuyu Bie, Min Lei. *Journal Of Computers*, **2013**.5:1357-1365.
[14]University of Waikato． Weka 3: data mining software in Java [EB/OL]． **2011**.3
http: //www.cs. waikato. ac. nz /ml /weka/.
[15]DEAN J， GHEMAWAT S.Map-Reduce: simplified data processing on large clusters [J]. **2008**,51(1) :107 -113.*//Communications of the ACM: 50th anniversary issue*，
[16] Apache Hadoop . Hadoop [EB/OL],**2011**.3. *http://hadoop.apache. org.*
[17]Wang Z Z, *Journal of Chongqing University of Science and Technology*.**2009**.8:166-168
[18]Yingxia Liu, Faliang Chang. *Journal Of Computers*, **2011**.5:849-856.
[19]Zhi-hang Tang, Wen-bin Tian. *Journal Of Software*, **2014**.2:451-457.