Research on the Classification and Selection of Archive Texts with the Improved C4.5 Algorithm

Xianbin Lv

***Abstract*—**In the age of information explosion, how to get the information we need from mass information has always been a problem for us. Hence, many data mining techniques have been developed. In this paper, the scope of data mining was further narrowed based on a data model constructed from text categorization. Logarithmic calculation was converted to a simpler arithmetic hybrid operation, which reduced the time overhead of generating decision trees by the algorithm through eliminating the procedure to call the library function, thus reducing the time overhead of the entire text classification process, with the Fayyad and Irani Boundary Theorems as well as the decision tree C4.5 algorithm introduced. The experimental results showed that the improved algorithm in this paper had a classification time of 2 minutes and 44 seconds, which was shorter than the original C4.5 algorithm and its average classification accuracy of 92.91% was close to that of the original C4.5 algorithm. Therefore, the improved C4.5 algorithm could be well applied to the calculation of the file classification and had good results.

***Keywords*—**C4.5 algorithm, decision tree, improvement, text classification

I. Introduction

With the development of communication technology, the number of information is increasing rapidly every day. How to effectively classify and manage information so that users can rapidly and accurately acquire information they needed is one of the important issues in the modern society. Text is the most common form of information expression; hence the amount of text information is huge. Text classification is an indispensable part of data mining research and is effective in efficiently managing and utilizing information; therefore it has been an important direction in data mining research. Many scholars in China and abroad have studied text classification. Jiang et al. [1] proposed a self-organizing algorithm based on fuzzy similarity for feature clustering. Experimental results showed that the proposed method could speed up the extraction of text classification features. Shi et al. [2] proposed a new semi-supervised classification algorithm based on tolerance rough set and set learning. Two common text corpora, namely Reuters 21578 collection and WebKB collection were experimentally evaluated to verify the effect of the method in the improvement of text classification. Jiang et al. [3] put forward a graphics based text classification method which expressed document collection as graph collection, extracted frequent subgraph using the weighed graph mining algorithm, and further processed them to generate the feature vectors of classification. Zhou et al. [4] designed a C-LSTM neural network by combining the advantages of convolutional neural network and recurrent neural network and proved its excellent performance through experiments. Shi et al. [5] put forward a rough set and integrated learning based semi-supervised algorithm for text classification and proved its effectiveness through experiments. Javed et al. [6] adopted two-stage FS algorithm to improve the rate of feature extraction in text categorization by using FR metrics such as BNS or IG in the first phase and FSS algorithm such as Markov Blanket Filter (MBF) in the second phase. Uysal et al. [7] proposed a potential semantic feature based on genetic algorithm to eliminate vectors with large singular values to get a better representation of documents in text categorization. A new filter based probability characteristics selection method proposed by Uysal et al. [8] performs well in classification preciseness and processing time. Wan et al. [9] combined k-nearest neighbor classification and support vector machine training algorithm together, which was highly efficient in processing texts. Based on the characteristics of archival text, this study proposed an improved C4.5 algorithm to improve the efficiency of archival text classification.

X. B. Lv is with Jining No.1 People’s Hospital, No.6 Jiankang Road, Rencheng District, Jining, Shandong, 272000, China (lxianbjn@126.com).

II. Text classification

A. Overview of text classification

In the era of big data, the data circulated in computers include not only the resources needed in people's work and study, but also a lot of harmful information. Therefore rational and effective management of the information has become a quite important problem. Text data is the most common part of mass data, and text classification as the key technology for processing text data is of great significances to efficient management and use of information.

Text categorization refers to sorting texts into predetermined text categories according to content. It is a supervised learning process. It acquires the model of relation between text characteristics and text class from the labeled text sets and then determines the categories of new texts using the relation model.

In the perspective of mathematics, text classification can be regarded as a mapping process, i.e. mapping unknown categories of texts to predefined categories.
Because of the particularity of text, there are many differences between text classification and other pattern classification. For example, text set has a high-dimensional feature space and distributes sparsely. The relationships between characters and between words are close. The relationship between them should be taken into full consideration before classification. The flexible and changeable meanings of texts and the existence of various polysemants and synonyms bring great difficulties to text processing by computer.

B. Text preprocessing

In archival texts, not all words have an important role in the classification of texts. Therefore, the texts need to be preprocessed before classification, including the following items:
(1)Text segmentation [10]: according to the grammatical rules and text features of the Chinese and English texts, the text string is divided into words or phrases.
(2)Remove the stop words [11]: stop words are generally function words and the words that are not strongly colored that have no special effect on text categorization, or those whose deletion can reduce the dimensions of the feature words. Before removing the stop words, a stop word table should be set up, according to which the words or phrases in step (1) are removed, by which the calculation amount of text classification is reduced.
(3)Merge similar words: combine words that are similar in meaning and represent them in one word.
(4)Word frequency statistics [12]: the statistical software is used to count the number of appearance of the words processed by the above operations in the text.
(5)Text transformation and representation [13]: a series of strings obtained by the words processed in the above steps should be converted to institutionalized data which can be identified by the computer.

C. Feature selection

In this design, information gain algorithm [14] is applied for feature selection, which is a widely used method in the machining learning field. The method determines the category of the document by calculating the difference between the information entropies of a certain feature word appearing or not in a document, with its calculation formula as follows:

$$G(v) = -\sum_{j=1}^{m} P_u(A_j) \log P_u + P_u(v) \sum_{j=1}^{m} P_u(A_j|v) \log P_u(A_j|v) + P_u(v) \sum_{j=1}^{m} P_u\left(A_j \middle| \bar{v}\right) \log P_u\left(c_j \middle| \bar{v}\right) \tag{1}$$

where v refers to feature words for text classification, A_j refers to document classification, with a total number of m; $P_u(A_j)$ refers to the probability of a document coming from A_j, $P_u(v)$ refers to the probability that feature v appears in a document, p_u refers to the various types of probabilities associated with feature v in each case. It can be learnt that the use of feature v can help to classify the texts accurately with information gain express. Based on the calculation of information gain, some excellent classification features can be screened out.

III. Text classification algorithm

A. C4.5 algorithm

Suppose the sample data set to be B, with its category assembling to be A, $A = \{A_1, A_2, ..., A_m\}$, $|A_j|$ refers to the record number of A_j. Based on category A, sample set B is divided into m data subsets B_j,(1 ≤j ≤ m). Suppose the attribute set of B is C_n, $C_n = \{C_1, C_2, ..., C_n\}$, where C_j has h values $\{c_{1i}, c_{2i}, ..., c_{hi}\}$. The data set C is divided into h different subsets B_j^C, where the absolute value of B_j^C represents the sample number of subset B_j^C, A_j^C represents the number of category C_j in B_j^C. The formula for calculating the information gain rate is as follows:

$$entropy(B) = -\sum_{j=1}^{m} p_j \log_2(p_j) \tag{2}$$

where $p_j = \frac{|A_j|}{|B|}$, $A_j = B_j^C$.
Therefore,

$$entropyC(B) = -\sum_{J=1}^{m} \frac{|B_j^C|}{|B|} \times entropy(B_j^C) \tag{3}$$

where $entropy(B)$ refers to the information entropy of the sample data set, p_j refers to the probability that any sample belongs to category A, $entropyC(B)$ refers to the information entropy that divides sample data sets according to attribute C_j. Suppose the information gain of attribute C_j to be $Gain(C_j)$, $Splitl(C)$ represents split information amount, then:

$$Gain(C_j) = entropy(B) - entropyC(B) \tag{4}$$

$$Splitl(C_j) = -\sum_{j=1}^{h} \frac{|B_j^C|}{|B|} \log_2 \frac{|B_j^C|}{|B|} \tag{5}$$

Then, the information gain rate is:

$$Gain\_Ratio(C_j) = \frac{Gain(C_j)}{Splitl(C_j)} \tag{6}$$

The attributes with the maximum gain rates are divided into subsets with different nodes, based on which the branches and leaves of the decision tree are built. Recursion uses each attribute to establish the branches of the decision tree until the class of the nodes is the same, and the decision tree is constructed.

B. Improvement of the C4.5 algorithm

In the actual operation of C4.5 algorithm, a large number of attribute values are often required as candidate partition points. The calculation of a large amount of information entropy and split information required for classification means that a large number of logarithm operations must be performed, which will slow down the overall operation efficiency. Moreover, in the process of constructing the decision tree, if the included data has many continuous attributes, it will increase the complexity of the decision tree model, easily lead to over-fitting phenomenon and greatly affect the data classification and selection operation.

a. Improvement of the formula

The C4.5 algorithm is optimized by Fayyad and Irani boundary theorem [15], with the theoretical idea as follows: In any case, the best demarcation point for data appears at the data space boundary. According to this idea, the discretization process of continuous data can be greatly simplified, and a certain amount of calculation time can be reduced.

According to the size of the attribute value, a data set is ranked. The distance between points is calculated to obtain the boundary point. The obtained number of boundary points equals to predictive set data-1. In this way, the number of classification calculations can be reduced so as to improve the calculation efficiency. Besides, Cini indicator [16] is introduced to replace the information entropy to obtain the best boundary point.

Gini indicator is calculated as follows:

$$Gini(\mathrm{B_t}) = 1 - \sum_{i=j}^{\mathrm{k}\text{-}1} (P_j)^2 \tag{7}$$

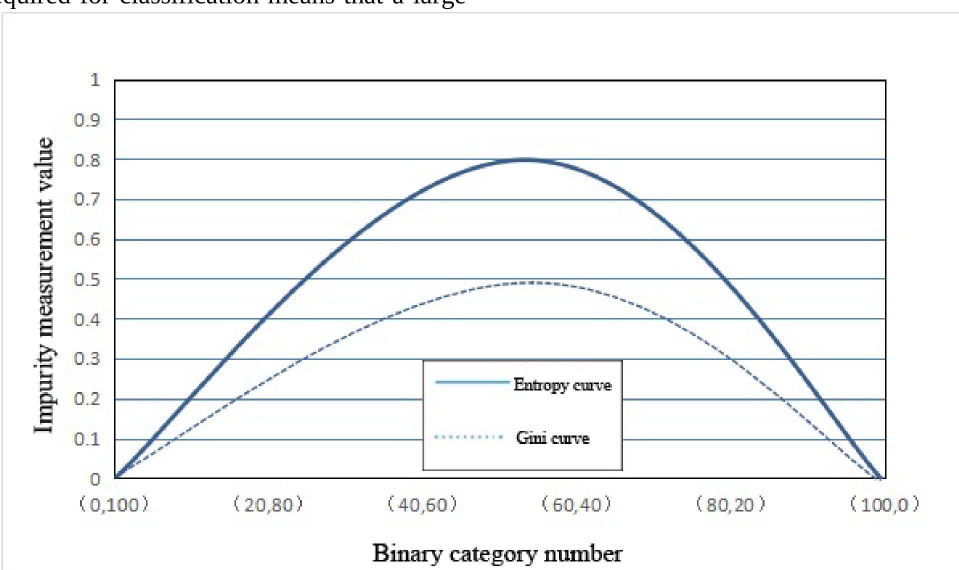

Fig. 1 Experimental results of impurity measurement of information entropy and Gini indicator

The entropy curve shows that the information entropy is used as an experimental result of the impurity measurement [17]. The Gini curve indicates that the Gini index is used as an experimental result of the impurity measurement. The impurity measurement value of information entropy is between 0-0.8 while that of the Gini indicator is between 0-0.5. The impurity measurement values of the two curves at different class points are different while their tendency and best boundary point are the same. However, compared to the logarithmic calculation of information entropy indicator, the operation of the Gini indicator is simpler, with higher calculation efficiency.

b. Over-fitting improvements

According to Occam's razor law [18], the algorithm is optimized: If not necessary, do not add entities. For text categorization, a more concise and effective computational model is the more desirable one.

Suppose there are T records and s classes in sample data set B, the set of the decision tree leave nodes of B is $\{L_1, L_2, ..., L_q\}$, with q nodes; suppose the class of the first leave to be $\{L_{k1}, L_{k2}, ..., L_{ls}\}$, then the total number of classes per leaf is $|N_l|$, the number of each class in node 1 is $|L_{li}|(1 \le i \le s), \max(|L_{li}|), (1 \le i \le s)$ represents the maximum value. The generalization error [19] is expressed as:

$$\mathrm{EFDT} = \frac{\sum_{l=1}^{q}\sum_{j=1}^{s}\left(|N_l| - \max\left(|L_{lj}|\right)\right)}{T} \tag{8}$$

When the generalization error is the same, in order to ensure the accuracy and efficiency, it is more desirable to simplify the model. When the training error is less than one value, the branch of the decision tree is stopped. The value should be obtained through multiple calculations according to the requirements of classification.

c. Algorithm improvement description

The improved C4.5 algorithm code is as follows:

```
Input: sample data set B (a total of 1 attributes, the first
one is the label number of B)
Output: Decision tree classifier
1.  Calculate the node's information gain rate
For (j = 1; j <= l; j ++) {
// If the attribute is not continuous, calculate the information gain rate.
If((attribute[j])=discontinuous){
Gain_Ratio=gain_Ratio(j);
} else{
//find the Gini indicator according to ranking of the sequential attributes of the data set to find the best demarcation point;
BB1=classificationBs(attribute[j]);
Point  p[]= search For Point(B1);
Gini[]  gn=Gini(p);
gini Min = search For Min Gini(gn) ;
Point pm = search Point By Gini(gn[min]) ;
Calculate the information gain rate of the best demarcation point;
Gain Ratio = gain Ratio( pm) ;
}
2.  Build the decision tree
// Node Number represents the number of leaf nodes in the decision tree. Each new node split is used as a new leaf node.
while(node.type=new){
// EFDT represents the decision tree generalization error)
EFDT=Error For Decision Tree();
//If the generalization error is greater than a certain value, the decision tree branch growth continues, otherwise, the tree growth ends;
If(EFDT>δ){
Tree Growth() ;
Node.type = old ;
}else{
Stop Tree Growth() ;
}
}
```

IV. Experiment Verification

A. Classification structure

Based on the analysis of the text file data of Jining No.1 People's Hospital, the file number of the archive text largely reflects the subject information of some documents and the subject matter of most texts while the text title can display the theme. Hence, in this experiment, the classification is performed by classifying the text titles and themes into three levels according to their importance values, as shown in Fig. 2.

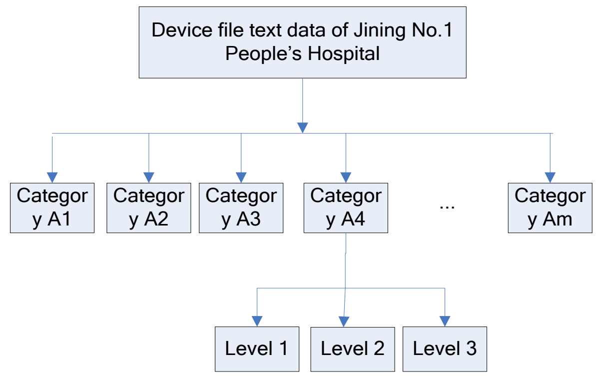

Fig. 2 File text classification structure

B. Text preprocessing

a. Design of the stop word table

To remove the stop words, a stop word table should be established. Firstly, the numbers from 1 to 10 are of no value to classification and should be included to the stop word category. Then, the words such as notification, method, stipulation, of and record as well as some symbols and numbers in the file titles should be included, as shown in Fig. 3. By using specifically designed stop word table, the amount of word data in the archive text can be reduced with the classification accuracy unaffected.

	A
1	Stop Words
2	notification
3	method
4	stipulation
5	summary
6	suggestion
7	on
8	report
9	opinion
10	work
11	hospital
12	of
13	plan
14	Jining City
15	Health Bureau
16	situation
17	in

Fig. 3 Stop word table

b. Text segmentation processing

Generally, word segmentation software is used to divide the sentences into words or phrases according to set grammar rules.

Then, stop words are removed from these words and phrases according to the above mentioned stop word table. Finally, the remained words and phrases are input to the text document, as shown in Fig. 4.

"Logistic equipment maintenance" "Disinfection equipment installation and use" "Office equipment specifications
Issued "" management equipment "" use of water and electricity equipment "" care instrument maintenance "" use
Department records "" Operation and treatment equipment "" Purchase contract "" Instructions for use Optical Instruments "
"Optical Microscope Maintenance" "Radiological Inspection Operation" "Nuclide Device Operation" "Technology
Specifications "" Equipment Qualification "" Surgical Instruments "" Refrigeration Equipment Management "" manual
Installation "" Maintenance Person "" "" Purchase Report "" Maintenance Acceptance "" Owner "" Retired Equipment "" Equipment Disposal "
"Equipment Purchase Origin" "Purchase Invoice" "Boiler Equipment Use Management" "Measuring Instruments" "" ""
"Diagnostic Device Purchase" "Department Diagnostic Use" "Instrument Assignment Planning" "Therapeutic Instruments
Maintenance "" Jining City Management "" Jining City Health Bureau "" Medical Equipment Maintenance "",
Other Equipment Maintenance "" Equipment Installer "" Traffic Equipment Use "" Traffic Equipment "
"" "" "" "" "

Fig. 4 Text after segmentation

C. Test results

In this paper, 2000 device file text data is extracted from Jining No.1 People's Hospital and preserved to the three storage directories of different levels based on the value of the text. Then, the C4.5 algorithm and C4.5 improved algorithm were applied to the classification of the text file data, with the results shown in Table 1.

Table 1 Text file classification results of C4.5 algorithm and C4.5 improved algorithm

	C4.5 algorithm	C4.5 improved algorithm
Total number of file text samples	20000	20000
Correct classification number	18627	18582
Incorrect classification number	1373	1418
Average classification accuracy (%)	93.135	92.91
Decision tree construction time	3 minutes and 7 seconds	2 minutes and 28 seconds
Time of classification	3 minutes and 47 seconds	2 minutes and 51 seconds

As shown in table 1, the average classification accuracy of the improved C4.5 algorithm and the C4.5 algorithm were relatively similar, which were 92.91% and 93.125% respectively. The decision tree construction time of the improved algorithm was 20.05% faster than that of the original algorithm due to the Occam's razor improvement. The time of classification of the improved algorithm was 2 minutes and 51 seconds, which was 24.67% shorter than that of the original algorithm due to the decrease of the time building the decision tree. On the premise of ensuring the accuracy of classification, the improved C4.5 algorithm spent less time in classifying texts compared to the conventional C4.5 algorithm, which greatly improved the classification efficiency of the archive texts.

V. Conclusion

In the era of information explosion, data mining has become one of the key points of research. Text classification, as an important part of data mining, can manage and utilize text information well, and has a high research value in practice. At present, the most common text classification methods include k-nearest neighbor algorithm, neural network algorithm, support vector machine and decision tree algorithm [20]. Decision tree algorithms which mainly include ID3 algorithm, CHID algorithm, CART algorithm and C4.5 algorithm [21] is simple and convenient in text classification.

Before text classification, it is necessary to preprocess text first. In this paper, preprocessing such as segmenting texts, eliminating stop words and merging similar words were performed, and then the information gain method was selected as the feature selection method.

In the C4.5 algorithm, logarithmic operation needs to be done repeatedly, which can increase computation load. In order to reduce the time overhead, the C4.5 algorithm was optimized by using the Fayyad and Irani boundary theorem. The logarithmic calculation in the calculation process was converted to a simpler four hybrid operation, which saved time. The final test results showed that the improved algorithm finished archival text classification in 2 minutes 44 seconds, shorter than the C4.5 algorithm; the average classification accuracy of the file text was 92.91%, which was close to that of the conventional C4.5 algorithm. All the findings suggested that the improved C4.5 algorithm could greatly shorten classification time and moreover achieve a high accuracy.

Text classification should not only have a full understanding of data mining, but also take the characteristics of text into account. With the development of Internet technology, information on the network has increased the difficulty of text classification. The structure and relationship between texts are more complex, which put forward higher requirements on text classification. Different text classification methods should be developed according to the characteristics of data from different fields. In this study, the C4.5 algorithm was improved according to the characteristics of archival texts, and it had a better performance in text classification. The algorithm can be promoted to other fields.

Since there are many words or phrases that have little effect on the result of text categorization, this paper first preprocessed the archive text and then made a feature selection. The C4.5 algorithm was improved to make classification of the archive text by taking the text data of the device file in Jining No.1People's Hospital as the research objects. The results showed that the file text time of the improved algorithm was much shorter than that of the original C4.5 algorithm, with a classification accuracy rate of 92.91%. Therefore, the C4.5 improved algorithm proposed in this paper could be effectively used in the classification of archive texts.

The algorithm in this paper was proved effective in classifying texts. But the algorithm remains to be optimized to shorten time and improve accuracy before being applied in texts with larger scale.

REFERENCES

[1] J. Y. Jiang, R. J. Liou and S. J. Lee, "A fuzzy self-constructing feature clustering algorithm for text classification," *IEEE Trans. Knowled. Data Engin.*, vol. 23, no. 3, pp. 335-349, 2011.

[2] L. Shi, X. Ma, L. Xi, Q. Duan and J. Zhao, "Rough set and ensemble learning based semi-supervised algorithm for text classification," Expert Syst. Appl., vol. 38, no. 5, 6300-6306, 2011.

[3] C Jiang, F Coenen, R Sanderson, M. Zito, "Text Classification using Graph Mining-based Feature Extraction," Knowledge-Based Syst., vol.23, no.4, pp. 302-308, 2010.

[4] C. Zhou, C. Sun, Z. Liu, F. C. M. Lau, "A C-LSTM Neural Network for Text Classification," Comp. Sci., vol. 1, no. 4, pp. 39-44, 2015.

[5] L. Shi, X. Ma, L. Xi, Q. Duan, J. Zhao, "Rough set and ensemble learning based semi-supervised algorithm for text classification," Expert Syst. Appl., vol. 38, no. 5, pp. 6300-6306, 2011.

[6] K. Javed, S. Maruf and H. A. Babri, "A two-stage Markov blanket based feature selection algorithm for text classification," Neurocomputing, vol. 157, pp. 91-104, 2015.

[7] A. K. Uysal and S. Gunal, "Text classification using genetic algorithm oriented latent semantic features," Expert Syst. Appl., vol. 41, no. 13, pp. 5938-5947, 2014.

[8] A. K. Uysal, S. Gunal, "A novel probabilistic feature selection method for text classification," Knowledge-Based Syst., vol. 36, no. 6, pp. 226-235, 2012.

[9] C. H. Wan, L. H. Lee, R. Rajkumar, D. Isa, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine," Expert Syst. Appl., vol. 39, no. 15, pp. 11880-11888, 2012.

[10] A. Bal and R. Saha, "An Improved Method for Text Segmentation and Skew Normalization of Handwriting Image," Geoderma, vol. 75, no. 1-2, pp. 117-133, 2018.

[11] I. A. Elkhair, "Effects of Stop Words Elimination for Arabic Information Retrieval: A Comparative Study," Khair, vol. 4, pp. 119-133, 2017.

[12] W. T. Hu, Y. Yang, H. F. Yin, Z. Jia and L. Liu, "Organization name recognition based on word frequency statistics," Appl. Res. Comput., vol. 30, no. 7, pp. 2014-2016, 2013.

[13] H. Ku, "The effect of computer-based multimedia instruction with Chinese character recognition," EMI, vol. 48, no. 1, pp. 27-41, 2011.

[14] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," Knowl-Based Syst., vol. 24, no. 7, pp. 1024-1032, pp. 2011.

[15] H. Yasin, M. Mohammad Yasin and F. Mohammad Yasin, "Automated Multiple Related Documents Summarization via Jaccards Coefficient," Int. J. Comp. App., 2011, 12(3):12-15.

[16] L. Ceriani and P. Verme, "The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini," J. Econ. Inequal., vol. 10, no. 3, pp. 421-443, 2012.

[17] J. Ledolter, "Binary Classification, Probabilities, and Evaluating Classification Performance," John Wiley & Sons, Inc, 108-114, 2013.

[18] J. Zahálka and F. elezný, "An experimental test of Occam's razor in classification," Mach. Learn., vol. 82, no. 3, pp. 475-481, 2011.

[19] F. Z. Zeng and X. L. Ma (2013, May).The generalization error bound for the multiclass analytical center classifier. Scient. World J. [Online]. 11, 5747-48. Available: http://xueshu.baidu.com/s?wd=paperuri%3A%2819e41038c93bf1da41c6aaa096fd6d85%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fonlinelibrary.wiley.com%2Fdoi%2F10.1002%2F9781118596289.ch8%2Fsummary&ie=utf-8&sc_us=13465608970966245477

[20] C. C. Aggarwal, C X Zhai, "A Survey of Text Classification Algorithms," vol. 45, no. 3, pp. 429-455, 2012.

[21] B. R. Patel, K. K. Rana, "A Survey on Decision Tree Algorithm For Classification," Int. J. Engin. Developm. Res., vol. 2, no. 1, pp. 1-5, 2014.