International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 9 2914 – 2917

Feature Selection Technique for Text Document Classification: An Alternative Approach

S.W. Mohod
Deptt. Computer Engineering, B.D. College of Engg.
Sevagram,
Wardha, India
sudhirwamanrao@gmail.com

Dr. C.A.Dhote
Prof.,Ram Meghe Institute of Technology & Research,
Badnera.
Amravati, India
vikasdhote@rediffmail.com

Abstract **-**Text classification and feature selection plays an important role for correctly identifying the documents into particular category, due to the explosive growth of the textual information from the electronic digital documents as well as world wide web. In the text mining present challenge is to select important or relevant feature from large and vast amount of features in the data set. The aim of this paper is to improve the feature selection method for text document classification in machine learning. In machine learning the training set is generated for testing the documents. This can be achieved by selecting important new term i.e. weights of term in text document to improve both classification with relevance to accuracy and performance.

Keywords-*Text classification, Feature selection.*

*****

I. INTRODUCTION

With the rapid growth of the world wide web and data in digital format, the task of automatic document classification is important for organization. The information and knowledge discovery from documents are done manually in many companies. Proper classification of electronic documents, online news, blogs, e-mails and digital libraries requires Text Mining, Machine learning and natural language processing techniques to extract required knowledge information. Text mining makes an attempt to discover interesting information and knowledge from unstructured documents. The important task is to develop the automatic classifier to maximize the accuracy and efficiency to classify the existing and incoming documents.

In reality a large portion of the available information does not appear in structured databases but rather in collections of text articles drawn from various sources. Unstructured information refers to computerized information that either does not have a data model or the one that is not easily used by a computer program. The term distinguishes such information from data stored in field form in databases or annotated in documents. However, data mining deals with structured data, whereas text presents special characteristics and is unstructured. The important task is how these documented data can be properly retrieved, presented and classified. Extraction, integration and classification of electronic documents from different sources and knowledge information discovery from these documents are important.

In data mining, Machine learning is often used for Prediction or Classification. Classification involves finding rule that partition the data into disjoint groups. The input for the classification is the training data set, whose class labels are already known. Classifications analyze the training data set and construct a model based on the class label. The goal of classification is to build a set of models that can correctly predict the class of the different objects. Machine learning is an area of artificial intelligence concerned with the development of techniques which allow computers to "learn". More specifically, machine learning is a method for creating computer programs by the analysis of data sets since machine learning study the analysis of data. The challenging task is of text classification performance, because many problems are due to high dimensionality of feature space and unordered collection of words in text documents. This paper will mainly focus on implementation of the text document feature selection.

'Bag of words' [1] is the simplest representation of textual data. Vector space model (VSM)[2] is widely used in document classification system. Thousands of term word occurs in the text document, so it is important to reduce the dimensionality of feature using feature selection process [3], to resolve this problem different techniques can be used. Researchers have used different feature selection methods such as X2 Statistics (CHI), Information Gain (IG), mutual information, term strength, document frequency. With the help of these approaches it is possible to reduce the high dimensionality of features. Proposed IDFDDF (Inverse Document Frequency Divide Document Frequency) is the most effective method to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifier. In this approach document preprocessing is also important to reduce the complexity and high dimensionality of term words occurs in the text document.

II. DOCUMENT REPRESENTATION

One of the pre-processing techniques is the document representation which is used to reduce the complexity of the documents. The documents need to be transformed from the full text version to a document vector. Text classification is again an important component in most information management tasks for which algorithms that can maintain high accuracy are desired. Dimensionality reduction is a very important step in text classification, because irrelevant and redundant features often degrade the performance of classification both in speed and classification accuracy. Dimensionality reduction technique can be classified into feature extraction (FE) [4] and feature selection (FS) approaches given below.

A. *Feature Extraction*

FE is the first step of pre-processing which is used to presents the text documents into clear word format. So removing stop words and stemming words is the pre-processing tasks [5] [6]. The documents in text classification are represented by a great amount of features and most of them could be irrelevant or noisy [7]. DR is the exclusion of a large number of keywords, base preferably on a statistical process, to create a low dimension vector [8]. Commonly the steps taken for the feature extractions (Fig.1) are:

Tokenization: A document is treated as a string, and then partitioned into a list of tokens.

Removing stop words: Stop words such as "the", "a", "and"… etc are frequently occurring, so the insignificant words need to be removed.

Stemming word: Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form, e.g. connection to connect, computing to compute etc.

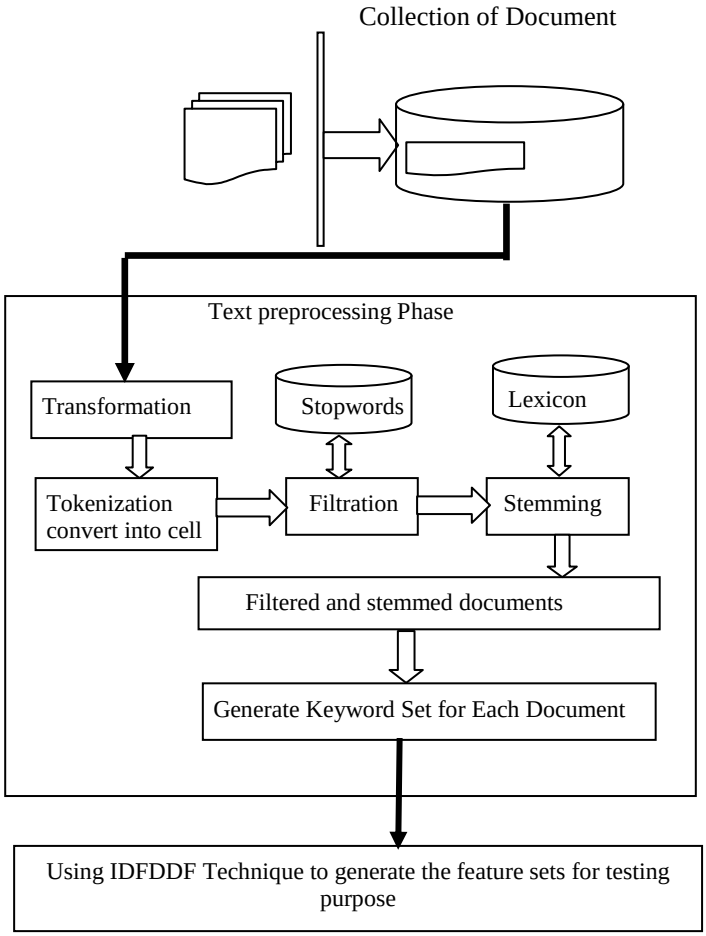


Figure 1 Feature selection Process

B. *Feature Selection*

After feature extraction the important step in preprocessing of text classification, is feature selection to construct vector space, which improve the scalability, efficiency and accuracy of a text classifier. In general, a good feature selection method should consider domain and algorithm characteristics [9]. The main idea of FS is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word [7]. The selected feature retains the original physical meaning to provide a better understanding for the data and learning process [4]. For text classification a major problem is the high dimensionality of the feature space. Almost every text domain has much number of features, most of these features are not relevant and beneficial for text classification task, and even some noise features may sharply reduce the classification accuracy [10]. Hence FS is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers.

In our approach TF, IDF, TFIDF, TFDF and Proposed method compare with data set R8 of Reuters 21578.

III. FEATURE SELECTION APPROACHES

Feature selection helps in the problem of text classification to improve efficiency and accuracy. In our approach we are examining different feature selection methods and then will find wheather our proposed method is effective to other studied method.

A. *TF (Term Frequency)*

Term frequency in the given document is simply the number of times a given term appears in that document. TF used to measure the importance of item in a document, the number of occurrences of each term in the document. Every document is described as a vector consisting of words such as

D = <Term1, Term2, Term3,……., Termn >

Where D means the Document and Term means the word on that document and n represents the number of words in the document.

Importance of the term't' within the particular document with 'ni' being the number of occurrences of the considered term and the denominator is the number of occurrences of all terms.

$$TF = \frac{ni}{\sum_K n_k}$$

B. *DF (Document Frequency)*

One way of calculating the document frequency (DF) is to determine how many documents contain the term't' divide by the total number of documents in the collection. |D| is the total no of documents in the document set D, and | {di tj ∈ di ∈ D}| is the number of documents containing term tj.

C. *IDF (Inverse Document Frequency)*

The inverse document frequency is a measure of the general importance of the term in the corpus. It assigns smaller value to the words occurring in the most of the documents and higher values to those occurring in fewer documents. It is the logarithm of the number of all documents divided by the number of documents containing the term.

$$IDF = \log \frac{|D|}{|(di \supset ti)|}$$

or

$$IDF = \log \frac{|D|}{|\{di\ tj \in di \in D\}|}$$

Where |D| is total no of documents in the corpus & $|(di \supset ti)|$ is number of documents where the term 'ti' appears.

D. *Our Approach : IDFDDF (Inverse Document Frequency Divide by Document Frequency)*

Using the inverse document frequency and document frequency it is the division of IDF and DF, it is denoted as IDFDDF.

$$\text{IDFDDF} = \frac{\log \frac{|D|}{|\{di\, tj \in di \in D\}|}}{|\{di\, tj \in di \in D\}|}$$

TFIDF is commonly used to represent term weight numerically using multiplication of term frequency and inverse document frequency [11]. IDFDDF is commonly used to represent inverse document frequency divided by document frequency. Here we can also get the numerical value and assign to the related term. Using these we can select the relevant (important) term from the total number of features in the corpus. The proposed method of feature selection is described is as follows.

Step1: Collect the different Data sets.
Step2: Text tokenization:
1. Its conversion is very important for separating the sentences into words.
2. Produce the sequence of term features of document.
3. Remove all non-alphanumeric characters then convert into word string.
Step3: Filtration:
First these words are converted into lower case
Remove all stop words using already well defined Blockade list.
Step4: Stemming:
In this process system removes the word's prefixes and suffixes. Applying stemming algorithm or own method.
Step5: We get the highly relevant words from the document.
Step6: All the terms and frequencies are collected from each document. Evaluate and retain values of TF.
Step7: Repeat Step2 to Step6 for all the documents of corpus.
Step8: Evaluate and retain all values of DF, IDF, and IDFDDF.
Step9: Obtain the word feature set of corpus.

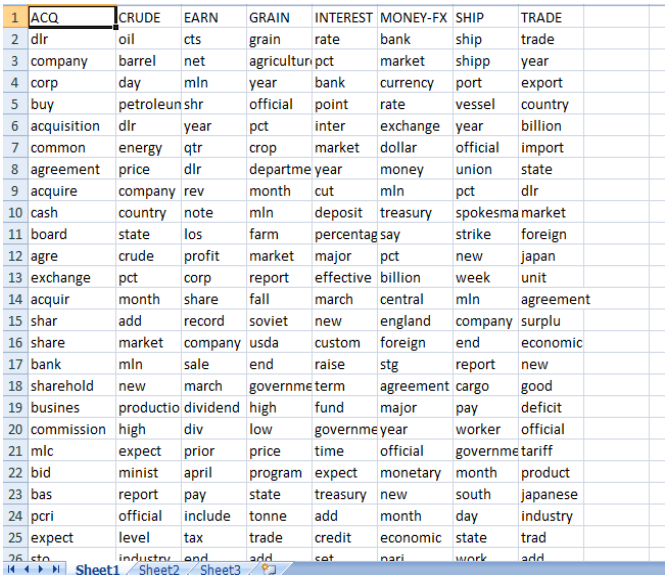
IV. EXPERIMENTAL RESULTS

To evaluate the performance of our proposed method we have perform experiment on data set R8 of Reuters 21578. The experiment have been performed on Pentium® Dual-Core CPU, 3GB RAM, Windows Vista 32-bit Operating System, and MATLABR2008a. Table 1 shows the detailed information of the data set. Performance of our proposed method using above mention dataset is shown in Figure 2.

To reduce the dimensionality of original text documents using preprocessing step, the documents are converted into the vectors after which removes the non-alphanumeric characters and then insignificant words (keywords) called as filtration process hence removes the noisy elements from the term vector. After removal of noisy elements perform stemming process which is important for the text document feature selection. In this process remove the lexicons i.e. s, es, ing, ed, est etc. This process can be executed using our own method for lexicon removal instead of any existing algorithms. Stemming process generates the different word in single form in which we get the original term features of the corpus. Using traditional document frequency and inverse document frequency algorithm calculates the numeric values for the term feature of the corpus. With the help of these values proposed IDFDDF method is used to generate new numerical values for the corresponding term. These term values are ordered by their IDFDDF values in descending order. For creating the training set for testing the corpus documents, conduct feature selection by picking up top few terms. It has been observed that using top most minimum terms related to corpus generated using proposed IDFDDF method are relevant with the class of the data set.

TABLE1. DATA DESCRIPTION

DataSet R8 of Reuters 21578 (all-terms,3.20Mb)								
Class	acq	crude	earn	grain	interest	money -fx	ship	trade
Train Docs	1596	253	2840	41	190	206	108	251



	ACQ	CRUDE	EARN	GRAIN	INTEREST	MONEY-FX	SHIP	TRADE
2	dlr	oil	cts	grain	rate	bank	ship	trade
3	company	barrel	net	agricultur	pct	market	shipp	year
4	corp	day	mln	year	bank	currency	port	export
5	buy	petroleun	shr	official	point	rate	vessel	country
6	acquisition	dlr	year	pct	inter	exchange	year	billion
7	common	energy	qtr	crop	market	dollar	official	import
8	agreement	price	dlr	departme	year	money	union	state
9	acquire	company	rev	month	cut	mln	pct	dlr
10	cash	country	note	mln	deposit	treasury	spokesma	market
11	board	state	los	farm	percentag	say	strike	foreign
12	agre	crude	profit	market	major	pct	new	japan
13	exchange	pct	corp	report	effective	billion	week	unit
14	acquir	month	share	fall	march	central	mln	agreement
15	shar	add	record	soviet	new	england	company	surplu
16	share	market	company	usda	custom	foreign	end	economic
17	bank	mln	sale	end	raise	stg	report	new
18	sharehold	new	march	governme	term	agreement	cargo	good
19	busines	productio	dividend	high	fund	major	pay	deficit
20	commission	high	div	low	governme	year	worker	official
21	mlc	expect	prior	price	time	official	governme	tariff
22	bid	minist	april	program	expect	monetary	month	product
23	bas	report	pay	state	treasury	new	south	japanese
24	pcri	official	include	tonne	add	month	day	industry
25	expect	level	tax	trade	credit	economic	state	trad
26	sto	industry	end	add	set	pari	work	add

Sheet1 Sheet2 Sheet3

Figure 2 Generated Feature Set Using IDFDDF

V. CONCLUSIONS

The proposed method is an another approach for feature selection for text classification. R8 of Reuters 21578 data set where used for experimentation. The proposed method, per forms well for feature selection. Hence the accuracy and performance in feature selection will certainly enhance by adopting the proposed methodology.

REFERENCES

[1] M. Yetisgen-Yildiz and W. Pratt, "The effect of feature representation on MEDLINE document classification," AMIA Annual Symposium Proceedings, pp. 849-853,2005.

[2] G. Salton, A. Wong, and C. S. Yang, A vector model for automatic indexing, Communication of the ACM, vol.18, no.11, pp.613-620, 1975.

[3] F. Beil, M. Ester, X. Xu, Frequent term-based text clustering, Proc. of Int'l Conf. on knowledge Discovery and Data Mining, KDD' 02, pp. 436–442, 2002.

[4] Liu, H. and Motoda., "Feature Extraction, constraction and selection: A Data Mining Perpective.", Boston, Massachusetts(MA): Kluwer Academic Publishers.

[5] Wang, Y., and Wang X.J., " A New Approach to feature selection in Text Classification", Proceedings of 4th International Conference on Machine Learning and Cybernetics, IEEE- 2005, Vol.6, pp. 3814-3819, 2005.

[6] Lee, L.W., and Chen, S.M., "New Methods for Text Categorization Based on a New Feature Selection Method a and New Similarity Measure Between Documents", IEA/AEI,France 2006.

[7] Montanes,E., Ferandez, J., Diaz, I., Combarro, E.F and Ranilla, J., " Measures of Rule Quality for Feature Selection in Text Categorization", 5th international Symposium on Intelligent data analysis , Germeny-2003, Springer-Verlag 2003, Vol2810, pp.589-598, 2003.

[8] Manomaisupat, P., and Abmad k., " Feature Selection for text Categorization Using Self Orgnizing Map", 2nd International Conference on Neural Network and Brain, 2005,IEEE press Vol 3, pp.1875-1880, 2005.

[9] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, Xin Li "An Optimal Svm-Based Text Classification Algorithm" Fifth International Conference on Machine Learning and Cybernetics, Dalian,pp. 13-16 , 2006.

[10] Jingnian Chen a,b,, Houkuan Huang a, Shengfeng Tian a, Youli Qua Feature selection for text classification with Naïve Bayes" Expert Systems with Applications 36, pp. 5432–5435, 2009.

[11] Ying Liu, Han Tong Loh, Kamal Youcef-Toumi, and Shu Beng Tor, "Handling of Imbalanced Data in Text Classification: Category-Based Term Weights," in Natural language processing and text mining, pp. 172-194.