ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

Feature Extraction for Document Classification

S.Vidhya [1], D.Asir Antony Gnana Singh [2], E.Jebamalar Leavline[3]

P.G. Student, Department of CSE, University College of Engineering, Anna University, BIT Campus, Tiruchirappalli, Tamil Nadu, India[1]

Teaching Fellow, Department of CSE, University College of Engineering, Anna University, BIT Campus, Tiruchirappalli, Tamil Nadu India[2]

Assistant Professor, Department of ECE, University College of Engineering, Anna University, BIT Campus, Tiruchirappalli, Tamil Nadu India[3]

ABSTRACT: Document classification is a significant and well studied area of pattern recognition, with a variety of modern applications. The purpose of document classification is to allocate the contents of a text or document for one or more categories. It is employed in document association and management, information retrieval, and certain machine learning algorithms. Feature extraction acquires an important subset of features from a dataset for improving the document classification task. Correctly identifying the related features in a text is of vital importance for the task of document classification. The document categorization problem is more challengeable when the data are in high-dimensional. In text mining, feature extraction and document classification are important techniques. The main aim of feature extraction is to reduce the dimensionality and eliminate irrelevant features so that efficiency and performance of the classification algorithms is improved. In this paper a term frequency (TF) with stemmer-based feature extraction algorithm is proposed and the performance of the algorithm is tested using various classifiers and it is observed that the proposed method outperforms other methods.

KEY WORDS*:* Document classification, Text mining, Feature extraction, High-dimensionality.

I.INTRODUCTION

Data Mining is the extraction of interesting or potentially useful patterns for knowledge discovery from huge amount of data. Knowledge helps to make prediction or decision making that can be used for industrial, medical, and scientific purposes. Text mining is one of the types of data mining. Text mining extracts only text data from huge volumes of data. If the dataset is of high-dimension, the information retrieval time is increased and the accuracy of the mining algorithm can be reduced. To overcome this problem, a technique known as feature extraction is introduced. Feature Extraction is the process of eliminating the irrelevant and redundancy features from the dataset. Using feature extraction technique accuracy is improved while assigning text into one or more categories. A feature extraction technique is proposed for document classification to improve the accuracy, to reduce the dimensionality, and to reduce the processing time.

Due to the consistent and rapid growth of unstructured textual data that is available, text categorization, the machine learning task of automatically assigning a predefined category label to a previously unlabelled document, is essential for handling and organizing this data. So feature extraction is applied to text categorization in order to improve its scalability, efficiency, and accuracy. Since each document in the collection can belong to multiple categories, the classification problem is usually split into multiple binary classification problems with respect to each category.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

II.RELATED WORK

The Anukriti and Bansal proposed a novel learning-based framework to extract articles from newspaper images using a Fixed- Point Model [1]. The input to the system comprises blocks of text and graphics, obtained using standard image processing techniques. The predetermined point form uses contextual information and features of each block to learn the layout of newspaper images and attains a reduction mapping to assign a unique label to all block. The hierarchical model works in two stages. In the first stage, a semantic label is assigned to each segmented block. The labels are then used as input to the next stage to group the related blocks into news articles [2].

The Karel Fuka and Rudolf Hanka developed a growing text classification application wherein important-term selection is a critical task for the classifier performance [3] [4]. They discussed some of the theoretical problems in text mining, which are identified and described. Pattern recognition techniques are found to be useful for text classification tasks.

Nawei Chen discussed an automatic document classification for organizing and mining the documents. Information in the documents is often conveyed using both text and images that complement each other [5]. Typically, only the text content forms the basic features that are used in document classification. On the other hand, the features formed by the visual words and the typical bag-of-words representation are commonly used to build the classifiers for document classification using the classification algorithm such as Naïve Bayes. They reported in their results that it performs better for classifying biomedical documents compared to methods which were previously used in the TREC Genomics track and employing the image-based representation [6].

In the literature [7], an overview of supervised and unsupervised text classification and clustering machine learning techniques is presented. The techniques described are those most widely used for text classification tasks. A number of issues particularly to the text classification task of the news source material, from its collection and organization to particular problems related to the evaluation of method correctness and categorization efficiency on Croatian news documents.

The Gnana Vardhini and H,Anju Abraham discussed an structural extraction [8], In this method the tree-mining algorithm is used for textual extraction, and also the algorithm is developed using fuzzy c-means clustering algorithm. Once the classification is carried out the supervised classification algorithm is used to combine both structural and textual feature vectors to build the classifier model [9].

III.PROPOSED SYSTEM

This section details the proposed system and its specification.

3.1 Preliminaries:
3.1.1 Null Stemmer: A dummy stemmer that performs no stemming at all.
3.1.2 IDFT: Sets whether if the word frequencies in a document should be transformed into $f_{ij} \times \log$ (no of documents with word i) where f_{ij} is the frequency of word i in document (instance) j.
3.1.3 TFT: Sets whether if the word frequencies in a document should be transformed into $\log(1+f_{ij})$ where f_{ij} is the frequency of word i in document (instance) j.
3.1.4 J48: J48 is an extension of Iterative Dichotomiser3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the algorithm.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

Table 1. Terminologies

Terminology	Description
NS	Null Stemmer
IDFT	Inverse Document Frequency Transform
TFT	Term Frequency Transform
J48	Decision Tree Algorithm

3.2 Algorithm:

Input: Dataset D={D_1,D_2,..D_n}
Where D->Data set
Di->Document

Output: F={F_1,F_2,….F_n}
Where F->A set of selected features from D
F_i->is the selected features

Begin
Load D
For(i=1;i≤N)// Where N is no of documents
{
Perform tokenizing [D_i]
Append the tokens to T_{list}
}
WC=find length (T_{list}) //wc is total no of tokens of the document
Sc=find (SW_{list})
For(i=0;i<wc)
For(j=0;j≤Sc)
If(SW_{list} (j)==T_{list}(i))
S_{list}={s_1,s_2,.s_n}
Eliminate T_{list}
Else
Append R_{list}//where R_{list} is the stop words removed text
X=length of NS
For(i=0;i<x)
{
Perform stemming R_{list}(i) and store V_{list}
}
Calculate TFT, IDFT{
Append the terms to TF_{list}
}
tf − idf is tf – idf(t, d) = tf(t, d) × idf(t,)
idf(t) = log2 |D|{ d:t∈d }

Preparation of feature vector an append FV_{list}
For(i=0;i<v)//where V is the stemming word list
{
Perform feature vector creation (FV_{list})
}

3.2.1 Algorithm Description:

The algorithm has eight phases; first the dataset is loaded, followed by tokenizing. In this step every word in the document is spilt into tokens, and then stored it as T_{list}. After that, prepare the stop words list and store it SW_{List}. Apply

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

the for loop condition to check the stop words. If the stored word is present in the document, then delete the word from the document. After performing the stop word elimination, do stemming, in stemming process apply null stemmer algorithm to convert the words into their root format. Then store it as R_{list}. Then calculate both TFT, IDFT using tf − idf is tf – idf(t, d) = tf(t, d) × idf(t,), idf(t) = log2 |D|{ d:t∈d }. Based on the term frequency, form the feature vector of the document. Then store it as Fv_{list}.

3.3 Feature extraction process: Feature extraction is a special form of dimensionality reduction. Extract features from the document through the evaluation of their weights in different related domains. The feature extraction is a preprocessing stage of the knowledge discovery. This preprocessing step aims at converting the free-text review sentences into a set of words and, at the same time, enriching their semantic meaning. Three subtasks involved are part-of-speech tagging, stemming, and meaningful word selection.
Steps for Document classification,
i .Read Document
ii.Tokenize
iii.Stop words removal
iv.Stemming
v.Feature vector representation
vi.Feature extraction
vii.Learining algorithm

3.3.1Term Frequency calculation:
Term frequency (TF) denotes the relative importance of a term to a document, which means that the more times a word appears in a document, the more important it is to that document, while tf-idf weight is a numerical statistic which reflects how important a word is to a document in a certain type (class) of collection or corpus. The mathematical representation of tf − idf is tf – idf(t, d) = tf(t, d) × idf(t,)in which idf(t) = log2 |D|{ d:t∈d }, |D| is the total number of document in the corpus, and |{d: t ∈ d}|is the number of document where term t occurs.

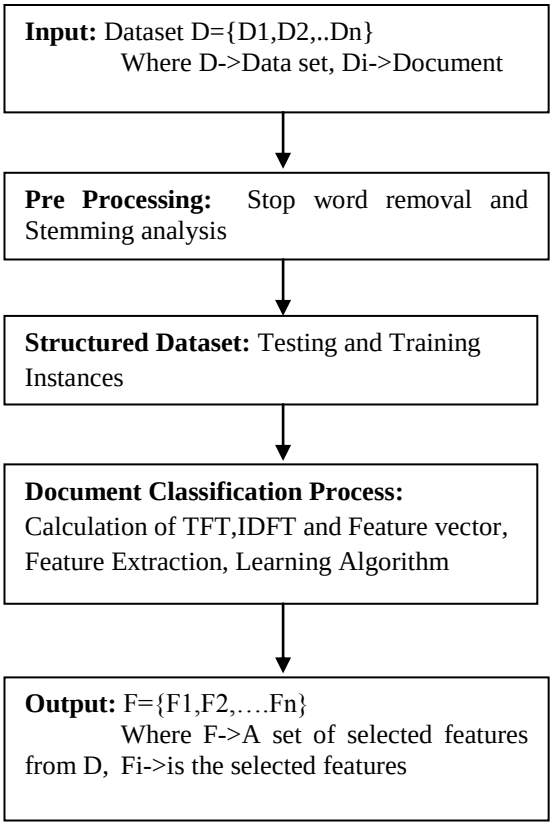
3.3.2 Learning algorithm:
Document classification or document categorization is a problem in library knowledge, information discipline and computer discipline. The rational classification of papers has generally been the prefecture of library knowledge. The troubles are overlapping, nevertheless, and there is for that reason interdisciplinary examine on document classification. Documents may be confidential according to their subjects or according to other attributes. There are two main philosophies of issue classification of documents: The comfortable based come near and the request based come near. In this element can classify the papers with previously qualified documents as news group datasets. Testing datasets are uploaded in user side and to perform preprocessing steps to remove noises and stemming progression and to conclude supply the grouping of the text documents. Using tree based algorithm one can improve the accuracy.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015



3.3.3 Evaluation criteria:
This module evaluates the performance of the proposed algorithm using attributes as follows. True positive (TP) are the positive tuples that were suitably labeled by the classifier. If the result from a prediction is 'p' and the actual value is also 'p', then it is called a true positive (TP). True Negative (TN) are the negative tuples that were acceptably labeled by the classifier. False Positive (FP) are the negative tuples that were imperfectly labeled as positive. However if the real value is n then it is measured to be a false positive (FP). False Negative (FN) are the positive tuples that were mislabeled as negative. Accuracy is calculated as (TP+TN)/(P+N) where, P=TP+FN and N=FP+TN. Or TP+TN/(TOTAL). According to experimental results, correctly classified occurrences for dataset are 1183 Accuracy of algorithm is 98.5 which are high. This new algorithm is a capable technique for this type of dataset.

3.4 Dataset for the conduction of the experiment: Reuters dataset is taken for extracting the feature, after extraction process the objects are split into two subsets: one for training (or development) and the other one for testing (or for performance evaluation).

Table 2.Details of Dataset

S.No	Dataset	Document	Classes
1	ReutersCorn	1554	2

3.5 Experimental Procedure: Following steps are performed for conduction of the experiment .
Step 1: Feed the newsgroup dataset to the Matlab.
Step 2: Remove the Stop words from the dataset (i.e. is, the, on..etc)

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

Step 3: Do stemming word analysis for removing the commoner morphological and in flexional endings from the words.
Step 4: Calculate the term frequency (TF) and inverse term frequency (ITF)
Step 5: Extract features based on the TF and ITF
Step 6: Calculate the accuracy for the extracted features using Stemmer algorithm.
3.6 System Specification: In order to justify, the effectiveness of the proposed algorithm, the experiment was conducted in the software environment MATLAB R2013a and personal computer with the configuration of Windows 7 Operating system and i5 processor.

IV. RESULT AND DISCUSSION

In order to justify the performance of the proposed method the experiment is conducted and the results are tabulated in Table 3.

Table 3.Featuer extraction method against No. of features, accuracy and Runtime

Methods	Extracted No. of Features	Accuracy of J48 algorithm	Time taken by J48 algorithm
TFT and IDFT with stemmer	1183	98.5%	3.29sec
Null stemmer	2023	93%	4.217sec

4.1 Improved Accuracy:

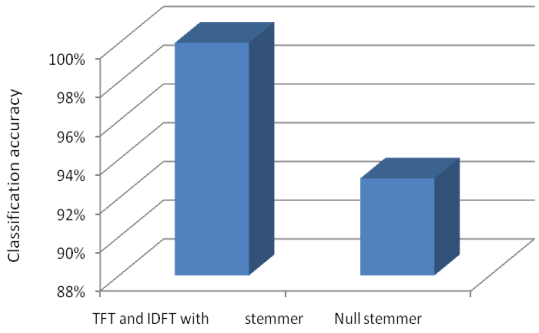


Figure 2. Classification accuracy for the Reuters dataset

4.2 Reduced Time:

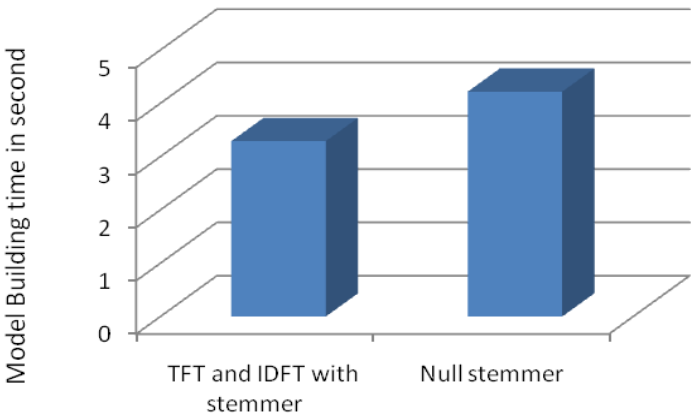


Figure 3. Time taken for model building for the Reuters dataset

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

4.3 Reduced Dimensionality:

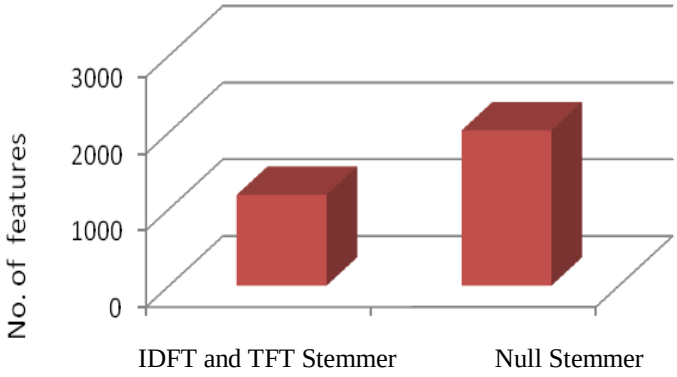


Figure 4. No. of Features Extracted for the Reuters dataset

V. CONCLUSION

In this paper, a feature extraction method is proposed for document classification. The classification accuracy was calculated using J48 classification algorithm. The effectiveness of proposed method was investigated and compared against well known other feature extraction techniques. The results of a thorough experimental analysis clearly indicate that proposed algorithm provides a considerably better performance in terms of accuracy, dimension reduction rate and processing time. For the future work, it is expected to combine the feature selection classification algorithm to improve the performance in document classification.

REFERENCES

[1] Anukriti Bansal,Santanu chaudhury,Sumantra Dutta Roy,J.B.Srivastava . Newspaper Article Extraction Using Hierarchical Fixed Point Model.Journal of Learning Research,2008.
[2] J. Bi, K. P. Bennett,M. J. Embrechts, C. M. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. Journal of Machine Learning Research, 3:1229–1243, 2003.
[3]Karel Fuka, Rudolf Hanka , Feature Set Reduction for Document Classification Problems,Journal of research papers,
[4] N. Cesa-Bianchi, S. Shalev-Shwartz, and O. Shamir. Efficient learning with partially observed attributes. Journal of Machine Learning Research, pages 2857–2878, 2011.
[5] Nawei Chen, Hagit Shatkay and Dorothea Blostein, Exploring a New Space of Features for Document Classification: Figure Clustering,In School of Computing,2010
[6] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. J. Mach. Learn. Res. (JMLR), 7:551–585, 2006.
[7] Boris Debi´c, Rock Harbor, Feature Extraction and Clustering of Croatian News Sources,In ACM,pages 59-69,2008
[8] Gnana Vardhini. H,Anju, Classification of XML Document by Extracting Structural and Textual Features,IJIR in Computer& Communication Engineering,2014.
[9] M. Dash and V. Gopalkrishnan. Distance based feature selection for clustering microarray data. In DASFAA, pages 512–519, 2008.