ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 8, Issue, 06, pp.32421-32425, June, 2016

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

GENRE DETECTION OF DOCUMENTS USING HYBRID TECHNIQUES OF MACHINE LEARNING

Nihar Ranjan, *Kavyashree Pushpan, Shraddha Samgir, Anjali Nair and Rutuja Murhekar

Department of Engineering, Sinhgad Institute of Technology and Science,
Savitribai Phule Pune University, Pune, India

ARTICLE INFO

Article History:
Received 23rd March, 2016
Received in revised form
10th April, 2016
Accepted 05th May, 2016
Published online 15th June, 2016

Key words:

Text mining, Text classification, ID3, Neuralnetwork, NLP

General Terms:

Text classification, Machine learning.

ABSTRACT

Document classification is an example of supervised machine learning. It is used to assign one or more categories to documents for making it easier to sort. Classification falls under Text mining. Complete task of text mining is to give the user, the benefit of the textual information and user could be able to perform the task of text retrieval, classification and summarization. The problem of assigning a specific category to a particular document can be done manually but when huge amount of documents exist, this task becomes practically impossible. Also, the time required to sort small amount of documents is also large compared to an automated classification system. Therefore, to reduce the manual efforts, a system is introduced that makes use of various machine learning algorithms.

Copyright©2016, Nihar Ranjan et al. *This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Citation: Nihar Ranjan, Kavyashree Pushpan, Shraddha Samgir, Anjali Nair and Rutuja Murhekar, 2016. "Genre detection of documents using hybrid techniques of machine learning", *International Journal of Current Research*, 8, (06), 32421-32425.

INTRODUCTION

Today is a world where social networks are highly influential. Even though people are under same roof they communicate with each other through any of the social networks like Facebook, whatsapp, hike, etc. Twitter is also a platform where people share their opinions frankly. All these communication leads to a lot of textual information. Not only these social networks, but even the smallest of the things are internet based, ultimately textual based, i.e., any work that is done manually earlier are converted to online applications like sending e-mails instead of letters, bank applications, and many such other cases. Some organizations consider even online resume. Many government documents are also preserved as a soft copy. Government officers have many confidential information which is stored in document format. This overloading of data could be seriously problematic and it is a really hard job to manage these large amount of textual data. Without any searching parameter, these textual data could be unwanted. This is where text classification is important and needed by the society and it is considered to be the best solution. Even Google uses text mining to search the enormous amount of queries that people have. It is almost impossible to search the data without any indexing or label over the topic. This indexing could be done by gathering the documents to various categories which is nothing but the process of classification. Categories may have a wide range since the data is huge. The various categories may be the content or the language or the culture or many other such categories. Text classification is nothing but a process or a task carried out in supervised machine learning. Basically there is a large number of documents taken as input and these documents are then categorized to a single or different category according to the content of the document. Different machine learning algorithms are used to design such classification systems like support vector machine, neural network, ID3, k-means, k-nearest neighbours and many more. These classifiers can be used individually or even combinations of these classifiers could be used. By combining various classifiers the efficiency of the system is improved. Various pre-processing tasks are carried out on the documents before applying classification algorithms on them. These pre-processing tasks are done with the help of NLP. Pre-processing task is done to reduce the

Corresponding author: Kavyashree Pushpan,
Department of Engineering, Sinhgad Institute of Technology and Science, Savitribai Phule Pune University, Pune, India.

amount of irrelevant data thus making the job of any classifier easy. Every classifier have their advantages and disadvantages, but by combining any classifiers, the system performance would increase. Ultimately the process of document classification becomes faster. And the complication of searching a document is reduced.

Literature survey

Now a day's, text classification has become a topic of interest to every individual. In the following we examine some basic and advance review papers related to text classification. Paper (Archna and Elangovan, 2014) is discussing about the various classification techniques used in text mining and a study on each of them. The machine learning algorithms specified in this paper are naïve bayes, rule induction, decision trees, nearest neighbours. As per the paper, all decision Tree's algorithms have less error rate and it is easier algorithm as compared to KNN and Bayesian. The disadvantage of decision tree algorithm are that they require certain knowledge and statistical experience to complete the process accurately. (Bhumika *et al*., 2013) Many text classification process and classifiers were discussed and elaborated. They also explained the document classification process which ended with performance evaluation. In this last step of text classification the efficiency of classifier is calculated. As described in the paper, for finding estimates of precision and recall relative to the whole category set, two different methods may be used like Micro-averaging and Macro-averaging, some other measures are also used as Break–even point, F-measure, Interpolation. Many classifiers are explained in this paper which are Rocchio's Algorithm, K-Nearest Neighbours, Naïve Bayes, decision tree, decision rule, neural network, LLSF, voting, Associative classifier, Centroid based classifier. All these classifiers were explained thoroughly. All algorithms are good for classification; even hybrid classifier can also be used. A comparative study is taken place in the paper. (Efstathios Stamatatos *et al*., 2001) An approach to text categorization in terms of stylistically homogeneous categories, either text genres or authors were mentioned. The paper (Thorsten Joachims, 1998) have discussed about support vector machine for text categorization. As per the paper SVM provides both theoretical and empirical evidence which are well suited for text categorization. The experimental results show that SVMs consistently achieve good performance on categorization task. As per the paper (Soundarya and Balakrishnan, 2014) a combination of decision tree and Bayesian network generally have different operational profiles, when one is very accurate the other is not, viceversa. According to this paper decision tree provides accurate results with Bayesian networks. Using four classifiers, namely, naïve bayes, nearest neighbour, decision tree, subspace (Li and Jain, 1998) document categorization was done. Categorization was done by individual classifiers and also with a combination of 2, 3 or 4 classifiers. It showed that combination of classifiers is always better. As per the paper (Soundarya and Balakrishnan, 2014) a combination of decision tree and Bayesian network generally have different operational profiles, when one is very accurate the other is not, vice versa. According to this paper decision tree provides accurate results with Bayesian networks. Using four classifiers, namely, naïve bayes, nearest neighbour, decision tree, subspace (Li and Jain, 1998) document categorization was done. Categorization was done by individual classifiers and also with a combination of 2, 3 or 4 classifiers. It showed that combination of classifiers is always better. An observation was made that (Ikonomakis *et al*., 2005) the performance of classifier is someway relevant to its own training corpus, if the training corpus is of high quality then the performance of the classifier is also good.

Implementation

Classification is the important task to categorize the document to various classes. Several classification algorithms are used in this phase which determines the classes of documents by various methods. For every text classification task, processing the textual data is must. In the beginning, the database has to be created with words in every category which are previously defined. These words have to be added manually so as to make the dataset stronger. This is known as training dataset. Basic architecture of classification is as shown in Figure 1.

The main task of document classification is i) document pre-processing, ii) feature selection, iii) training the classifier.

Pre-processing

This phase reduces the size of the input text document. Pre-processing a data means actually cleaning a data for the further processing of data. It mainly involves tokenization, stemming and stop words removal. This pre-processing task is done with the help of NLP. Natural Language Processing deals with the interaction between machine language and natural language

Tokenization

Tokenizer chops words from the entire textual data into tokens. Each word is separated in a sentence to form a token. For example, "I love milk", this sentence is tokenized as "I', "love", "milk". For this system tokenizer function is used for this process of tokenizing.

Stop words removal

Stop words are worthless for classification of documents. Stop words are just used in the sentence to make that sentence meaningful, but for classification process these stop words must be removed since the stop words doesn't fall in any category. Stop words are "a, an, the, of, etc."

Stemming

Stemmer removes the derivational endings to reduce word forms to common stem. Porter's Stemmer is one of the best algorithm used for stemming. For example, with the help of the Porter's Stemmer, "Specializations" can be reduced or stemmed to "Specialization" → "Specialize" → "Special"..

Feature selection

Feature selection identifies the meaningful words from the document and keeps only those words and remaining words are discarded. This is done with the help of Term Frequency.

Term Frequency

Term frequency is the number of times the term appears in a document. Basic formula for term frequency is the ratio of frequency of a term by maximum frequency of any of the term in the document.

$$tf(f_i, d_j) = \frac{freq_{ij}}{\max freq_{ij}}$$

Where tf (f_i,d_j) is the term frequency of the document d_j of training set.

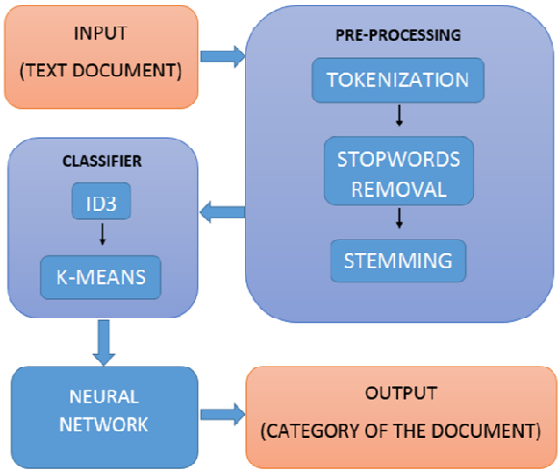

Figure 1. Basic Architecture

ID3

ID3 is a decision tree algorithm. ID3 stands for "Iterative Dichotomiser 3". Decision tree reconstruct categorization of training document in the form of tree structure. In tree structure, root node represents question and leaf node represents category. Decision tree works as shown in figure 2. Once the document is scanned, after pre-processing task, a hash value is generated for each document. This hash value is generated with the help of MD5 algorithm. This hash key is unique for every document, thus the hash key and the category of the document is stored in the database. If a new document is scanned and obtains the same hash key as a previous one then k-means and neural network is skipped and the time is saved, that is, if the hash key generated for the new document is matched with the hash key generated by the old document, then remaining part is skipped and directly the category of the document is shown.

The advantages of ID3 are that it implicitly perform feature selection or variable screening and for data preparation, it require comparatively less effort from users. They are easy to interpret.

The disadvantages of ID3 are that it is very time consuming and if the data is classified incorrectly, then it cannot be updated, instead a new tree has to be generated.

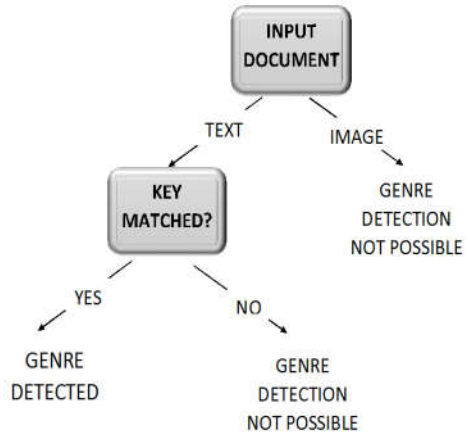

Figure 2. Decision Tree

K-means

It is a clustering algorithm. This algorithm is used for grouping of words that fall under same category. After pre-processing, a bag of meaningful words are obtained. The words with the count greater than the threshold value are grouped. These words are then matched with the keywords and category of each group is detected. The words that fall under same category has to be grouped again. For this k-means is used. Now the category that has the maximum count of words is the ultimate category of the document. Thus, this system uses k-means for grouping.

The advantage of k-means is that it forms tighter clustering than hierarchical clustering and it has high computational speed.

The main disadvantage of k-mean is that it is difficult to predicate the k-value

Neural Network

Neural network, is a mathematical model inspired by biological concept i.e. neural networks. A neural network is analgorithm where input set is a number of terms while output set contains the genre or category. A neural network consists of an interconnected groups. There are two types of learning algorithm that is supervised learning and unsupervised learning algorithm. These algorithm can be used for training a neural network. For classification of a documents (text document), weight are assigned to the input set, which is propagated forward along network. After that output set are decided which gives the conclusion of the text document means genre of documents. Perceptron is used to map the input weight to the output weight; it basically maps the input to the network that leads to the specific output.

Two types of perceptron 1. Single-layer perceptron 2. Multi-layer perceptron. But multi-layer perceptronis mostly used.

The main advantage of neural network is that it is simple to implement and easy to use. It is very powerful and it can also

model complex functions. Neural network can be executed in any application and no problem is faced.

The main disadvantage of neural network is that it has high complexity and it cannot be retrained, if you add data later, this is almost very hard to add to an existing network. It also requires high processing time. Handling time series data is a tough job in neural networks and sometimes it is impossible.

RESULTS

A number of documents are given as an input to the system. System scans each document. Each document is pre-processed and given to the classifier. The output of the system shows the category of the documents which were uploaded as an input. The category of a specific document must be relevant to the content of that particular document. A document may contain various category's content in them, then the classifier must decide that which category's content are more in that document and accordingly should display the result. Thus, expected output of the system is that it must show the correct category of the document uploaded, according to the content of the document.

Precision and Recall

Evaluation factors for this project are precision, recall and f-measure. Precision could be defined as the amount of retrieved documents that are query relevant whereas recall could be defined as the amount of relevant documents that are retrieved effectively.

$$Precision = \frac{|\{relevant\ doc\} \cap \{retrieved\ doc\}|}{|\{retrieved\ doc\}|}$$

$$Recall = \frac{|\{relevant\ doc\} \cap \{retrieved\ doc\}|}{|\{relevant\ doc\}|}$$

F-measure is the harmonic mean of precision and recall.

$$F = 2 \times \frac{precision\ \times recall}{precision + recall}$$

The evaluation factors for this project is shown in Table 1.

Table 1. Evaluation Factor

No of Files	Precision	Recall	F-measure
5	100	97	98.47
10	90	85	87.42
15	86	80	82.89
20	90	87	88.47

Graphical representation of the evaluation factor is shown in Figure 3.

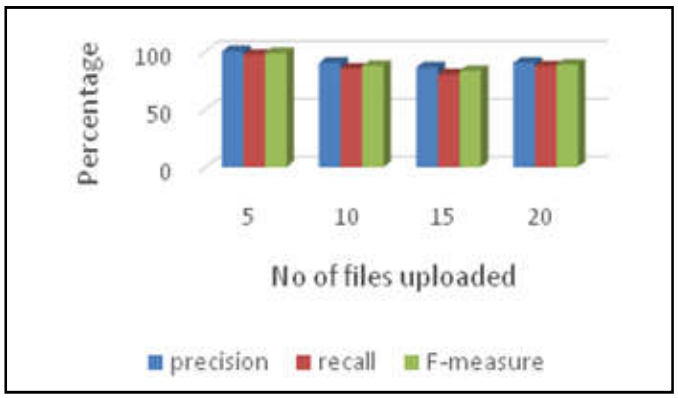

Figure 3. Graphical Representation of Evaluation Factors

Conclusion

The system was introduced to label or categorize the unknown or unlabelled documents. This system has 5 predefined categories. The system is trained for categorizing the document using some specific keywords. Each category has their own keywords. The keywords would be specific to the particular category. The input to this system would be an untitled document. The document is scanned by the system and forwarded to the pre-processing unit. The pre-processing unit performs the tasks such as tokenization, stemming and stop word removal. With the help of the given training set of the predefined categories, the document is labelled under one of the categories defined. After deciding the label of the unknown document, the system also learns the newly added document and trains itself to pick some more keywords from the document. Due to this dynamic nature of system, efficiency of system is improved. The system is designed to be user-friendly and cost effective. Thus, this system is designed for those people who are worried about their unlabelled documents.

REFERENCES

Archna, S., Dr. K .Elangovan, 2014. "Survey of Classification Techniques in Data Mining", *International Journal of computer science and mobile application (IJAIEM),* Volume 2, Issue 2.

Bhumika, Prof Sukhjit Singh Sehra and Prof Anand Nayyar, "A review paper on algorithms used for text classification", *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, Volume 2, Issue 3, March 2013.

Bing Liu, Wee Sun Lee, Philip S. Yu, Xiaoli Li, "Partially Supervised Classi cation of Text Documents".

Efstathios Stamatatos, Prof Nikos Fakotakis and Prof George Kokkinakis, 2001. "Automatic Text Categorization in terms of Genere and Author".

Ikonomakis M., S. Kotsiantis and V. Tampakas, 2005. "Text Classification Using Machine Learning Techniques", *Wseas Transactions on Computers*, Issue 8, Volume 4, pp. 966-974.

Li, Y. H. and A. K. Jain, 1998. "Classi cation of Text Documents", *The Computer Journal*, Vol. 41, No. 8.

Meenakshi, Swati Singla, 2015. "Review Paper on Text Categorization Techniques", SSRG International Journal of

Computer Science and Engineering (SSRG-IJCSE) – EFES.

Soundarya, M. and R. Balakrishnan, 2014. "Survey of Classification Techniques In Data Mining", *International Journal of Advance Research in Computer and Communication Engineering,* Volume 3, Issue 7.

Thamarai Selvi R., E. George Dharma Prakash Raj, 2014. "An Approach to Improve Precision and Recall for Ad-hoc Information Retrieval using SBIR Algorithm", World Congress on Computing and Communication Technologies.

Thorsten Joachims, 1998. "Text categorization with support vector Machines: Learning with many relevant Features",

International Journal of computer science and mobile application (IJAIEM), Volume 2, Issue 19.

Timothy P. Jurka, Loren Collingwood, Amber E. Boydstun, Emiliano Grossman, and Wouter van Atteveldt, "RTextTools: A Supervised Learning Package for Text Classi cation", *The R Journal*, Vol. 5/1, June ISSN 2073-4859.

Vandana Korde and C Namrata Mahender, 2012. "Text classification and classifiers: a survey", *International Journal of Artificial Intelligence & Applications (IJAIA)*, Vol.3, No.2.

****