Text Categorization using Jaccard Coefficient for Text Messages

Ankita Jadhao[1], Dr. A. J. Agrawal[2]

[1, 2] Ramdeobaba College of Engineering and Management, Nagpur, India

Abstract: ***There is wide growth in web application and electronic documents in day to day which needs automatic text classification of documents. Proper Classification methods provide the good results of the experiment and gives proper direction to the further processing of the text. The text is e-documents, news report, blogs, messages, comments on social media, e-books, web content etc which required text mining to extract meaningful knowledge from it. Some natural language techniques and machine learning algorithm are good to get the meaning of that e-document and classify them. There are lots of techniques are there for classification of the text documents, this paper is to understand different techniques and highlight the important methodology among them and helpful to selecting the classification technique which is appropriate to the text-classification process. And detail implementation of one of this method to classify the text message in two categories according the terms found in it. The coming text message is suspicious or not. In this case the Jaccard coefficient method gives the best result to classify message according to the words found in it. Text classification processes include several steps such as feature selection, vector representation and learning algorithm.***

Keywords: Document Classification, Natural Language processing, Information retrieval, Text mining

1. Introduction

There is many advancement in the world web applications nowadays everything are available on web it generate structured and unstructured data, it is very difficult to handle the unstructured data. To store such type of documents classification is necessary we have to classify those documents according to their category. Text mining required for classification of world wide web, social media data, online forums, government documents, blog repositories, digital books, news block, chat rooms, spam filtering, opinion mining for reviews, political situations, movies, improving the search result, sentiment analysis all these fields required text classification process.[1]

For classification text documents are assigning and automatically classify by using a machine learning technique. NLP, data mining and machine learning techniques these three are work combine for finding the patterns and classify the documents. Text mining is mainly work for extracting the information from the given text file that we called information extraction and then perform the retrieval and classification process.[2] Classification can be of three types supervised, unsupervised and semi-supervised. For the proper result documents or text must be correctly classified. The main task is to classify the text according to its features. There are many challenges in classification first is to representation of the document then annotation of the document, dimensionality reduction to handle algorithms and proper feature selection. By considering all these issues we have to choose one right algorithm for the classification.

The text documents are classified in to the predefine class or categories. Each document can be classify into exactly one, multiple or no category. The task is, if there is text documents D $\{d_1,d_2,d_3,\ldots.,d_n\}$ and 'k' no of predefine category or class $(c_1,c_2,c_3,\ldots..,c_k)$ then each document assign one category c_j to document d_i which is more relevant category to that document.

The main source of text document is the web, daily huge text data are generated at website which is mostly unstructured. Near about 80% data of any organization are in the unstructured format, in the form of email, news, junk file, reports etc. it means 90% of world digital data are in the unstructured format. Manually handling such huge and unstructured data is not possible, so need to automatically retrieval the knowledge from that data and analysis the document. So it required a good text mining system which gives more correct result for this huge amount of data. Handling and categorized this huge amount of data is another problem; identify the correct technique for our problem is necessary. [4]There are different techniques are available each of the techniques is having some merits and demerits considering all of them we can know which technique is more efficient in that case. The aim of this paper is to give the general view of all the techniques and comparative study of all the techniques in case of structured and unstructured data format. Classifying the structured data is any easy job as compared to unstructured data format. In structured data format already classified dataset are available at some repository by extracting the knowledge from that we can easily classify with the help of extracted knowledge other text document. In the unstructured from extraction of knowledge is a difficult task.

The process of classification of documents involves several steps like pre-processing, feature selection, vector representation and machine learning algorithm. Each step plays an important role in the whole process of classification. The first part is to document representation, we have to transfer the full text document into the vector document in which the whole complexity of the document is reduce. The text document is represents in vector of term weights form i.e. the count of the each term in the documents. [1]The main problem is the high dimensionality of the text sometimes the potential feature of the new query is exceeds than the number of training text documents. In that case we have to reduce the dimensions, drop the redundant and irrelevant feature from the potential feature that's increase the speed and accuracy of the classification. Dimensionality reduction techniques are further classified into two techniques Feature Extraction (FE) and Feature Selection (FS). Feature Extraction is the process in which the language dependent factors are removing. The pre-processing steps are in the feature extraction process such as tokenization, stop word removal and stemming. After that document is in the clear word format first is the tokenization in which the each word from the sentence is identified and separated. The sentence contains many words like "the", "a", "was", "and"… etc. they all are remove. Which reduce the size and last is the stemming is nothing but the change word which we get in its similar conical form that means store each word into its root form e.g. "walking", "walked", "walker" into its root form "walk".

After the feature extraction feature selection is the important step. We have select the feature in such a way that the original meaning retains the same and not to choose that much feature which increase the noise and reduce the accuracy of the classifier. The key technique is to remove irrelevant and redundant term only select the meaning full or important term which having high score.[6]

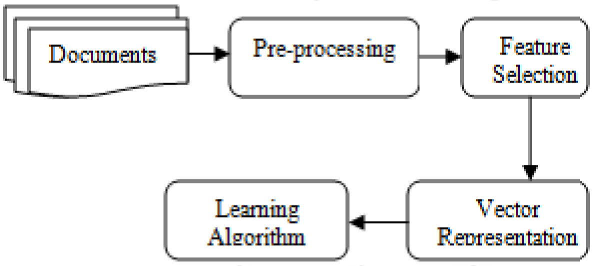

Figure 1: Document Classification Process

The structure of article is, first section gives the brief description of the text classification process, problems in it and application of text classification. Second section shows the review of text classification. Third section gives detail study of six classifier algorithm their advantages and limitation. Fourth section is about to one of the algorithm implementation. Fifth section shows the comparison of all algorithms and last section is the conclusion from study and implementation.

2. Literature Survey

[1]Vandana Korde et al (2012) discussed in her paper that the text mining studies are now important because all documents are available electronically, all web source are electronics document. Most of the documents are unstructured and semi structured, it is necessary to classify them.[2] The main goal is to extract the information from the document and perform the operation like retrieval, classification on it. The action is takes place with the help of natural language processing; machine learning and data mining it automatically classify the documents.

[2] Zakaria Elberrichi, et al (2008) proposed the new approach for text classification in the paper, with the help of the pre-define knowledge by WordNet. In which the author do first extract the 'k' best features from it and then categorized it on the basis of that 'k' best features. [3] Author shows experimental result with the 20Newsgroup dataset and Reuters21578 dataset the main problem in this method s that words having synonyms and it is not easy to find the perfect synonyms for it.

[3]William B. Cavnar et al (2010) proposed a highly effective method for classify the documents using the N-gram frequency technique. This method is pretty simple method in which only the occurrence of the terms are take into consideration instead of NLP which do parsing and lexicons etc which is more costly. Since it depends on the statistic it can overcome problem of OCR because it not only works on the particular word frequency it works on the N-gram occurrence.

[4] Anna Huang (2008) shows the experiment for text clustering according to similarity measures. Author describes some techniques and shows experimental results of it. Experiment on seven different dataset and five different clustering algorithm, [10]jaccard and Pearson coefficient gives the more coherent result. Author says that the three components affect the result more first is the document representation, distance and similarity measure and the clustering algorithm itself.

[5]Fabrizio Sebastiani et al (2010) the author says in his paper that text categorization evolved from the '80s, it is the fully blossomed field in computer and research it delivers the more effective, efficient, cheap solutions for the applications. It tackles with the wide variety of the problem and proposed different solutions.[8] Key to this success(i) The ever-increasing involvement of the machine learning community in text categorization, which has lately resulted in the use of the very latest machine learning technology within text categorization applications, and (ii) The availability of standard benchmarks (such as Reuters-21578 and OHSUMED), which has encouraged research.

3. Machine Learning Algorithm

In few years there is rapid growth of the Internet, everywhere Internet are use. So there is need to focus on this area more for more advancement. Document must be classified for efficiency purpose that's why there is rapid progress in it. There are varies method are available for categorizing of the documents. We have to choose one of the best methods among them. The documents of three types supervised, unsupervised, and semi supervised we discus here about the supervised techniques and new challenges in it. All the text documents are automatically correctly classifies into the predefine categories. Rocchio's Algorithm, K-nearest neighbor (KNN), Bayesian classifier, Decision Tree, Support Vector Machines (SVMs), Neural Networks, Latent Semantic Analysis, Fuzzy Correlation and Genetic Algorithms etc. these are the some supervised text classification techniques. Some of these techniques are discus below:

3.1 Rocchio's Algorithm

Rocchio's Algorithm is used to classify the documents using vector space method it incorporates relevance feedback for it. This algorithm finds the optimal query vector i.e. the one vector which maximizes the similarity to the relevant documents while minimizing the similarity to the non-relevant documents. Using the training set build the prototype vector for each class and calculate the similarity between test document and of prototype vectors, and classify the documents which having maximum similarity. Formula is:

$$C_i = \alpha * \text{centroid } c_i - \beta * \text{centroid } c_i$$

The text document classify into that category is having positive weight, and vector of all remaining document are given as negative weight. This algorithm is very easy to implement, fast and having relevance feedback mechanism but accuracy is low. [7]

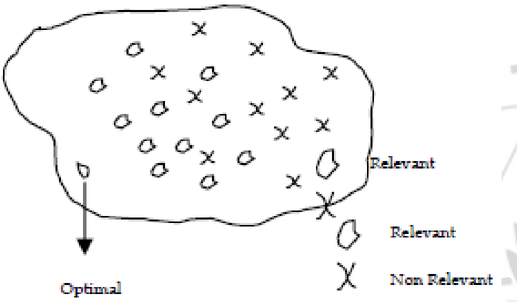

Figure 2: Rocchio Optimal query for separating relevant and non relevant document

This algorithm works well if the documents are having unique vector similarity. But there is one limitation that it fails to classify the multimodal classes and their relationship.

3.2 K-nearest neighbor (k-NN)

K-nearest neighbor methods works on the distance function. It is the simple algorithm in which the all training is used to first classifies on the similarity measures and then using this knowledge calculate for test cases. This is a instant based learning method which categorized the documents on its closet feature space in calculated training set. Mostly the Euclidean Distance formula is used for calculation of the distance between the vectors. The category is selected on the basis of the nearest point which is assign to that category. [7]

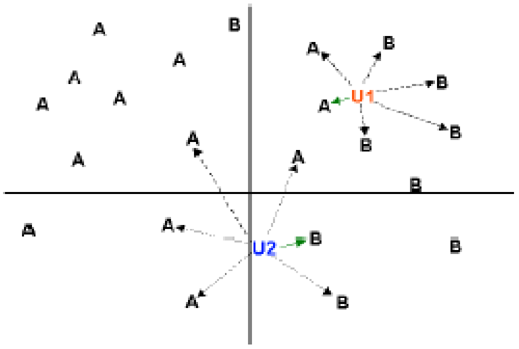

Figure 3: k-Nearest Neighbor

This method is more effective than the Rocchio algorithm in this method we consider all the characteristics in it. In this method the time spam is long to find the optimal solution but it classify more correctly. It is most famous method because of its simplicity and accuracy. It is well in categorizing the multi-categorized documents. But the one drawback of this method is that as increase the tanning sample its feature increases and generate the noise so its accuracy may be degraded.

3.3 Naïve Bayes Algorithm

Naïve Bayes classifier is works on the simple probabilistic model in which the bayes theorem used with strong (naïve) independent assumption between the features. The assumption makes Bayesian classifier more efficient because while classification one feature cannot interrupt the other feature. It is more efficient in case of limited training data because we considered the independent variable for each class and not the full covariance matrix of the model. [7]Naïve bayes classifier work efficiently in some real world classification applications with some specific condition. There is one disadvantage is that it having relatively low performance to other algorithms such as SVM etc. therefore so many researcher try to enhance the method to get more efficient results.

$$P(ci|D) = \frac{P(ci)P(D|ci)}{P(D)}$$

$$P(D|ci) = \prod_{j=1}^{n} P(dj|ci)$$

$Where\ P(Ci) = \mathrm{P(C{=}Ci)} = \frac{Ni}{N}$

$and\ P(dj|ci) = P(d|ci) = \frac{1 + Np}{M + \sum_{n=1}^{M} Nm}$

Naïve Bayes classifier are mostly used for the spam categorization, web content, email etc. this model works well with numeric and textual data. Naïve bayes doesn't show good result in case of when features do not consider the count of words and the features are mostly correlated with each other.

3.4 Support Vector Machine

Support vector machine methods are the discriminative classification method which is famous for more accurate result. SVM works on the decision plane which separate negative and positive values. SVM required positive as well as negative training set for classification which is uncommon. The decision takes place on the basis of closest of the decision surface.[7] It works to minimize the risk of inaccurate result the hypothesis is that it having lowest true error.

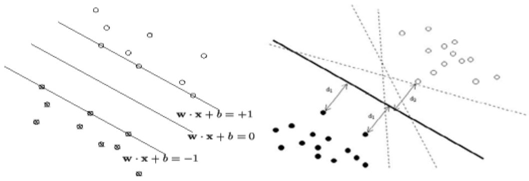

Figure 4: Illustration of optimal separating hyper plane, hyper planes and support vectors

SVM can handle more dimensional input space and drawback is it is relatively more complex, required more space and time for training and categorization of documents. SVM is the best technique for document classification.

3.5 Neural Network

Artificial neural Networks are neural structure of the human brain. It having three stages first is input layers middle one is for calculation and the last stage is output stage. It process one records one time and learn by comparing of the records with the predefine classification of the records. And in case error occurs in it the fed back into the network for further iterations.

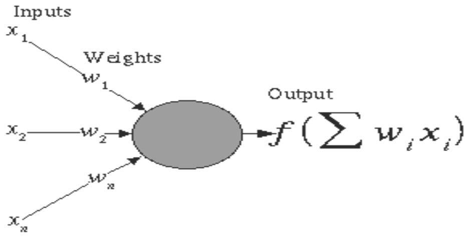

Figure 5: Artificial Neural Network

The main advantage of this method is it can handle high dimensional features and noisy data also. Linear speed up is possible in matching process by increasing computational element. [7]The main drawback of this method is their computation cost and CPU usages and the memory usages. The back-propagation neural network model used to improves the efficiency. The output of the neuron is:

$$O_{pj}(net_j) = \frac{1}{1+e^{-\lambda net_j}}$$

and

$$net_j = bias * W_{bias} + \sum_k O_{pk} W_{jk}$$

Neural network methods give the good result in complex data and it is also use for both discrete and continuous data. And neural network try to minimize the error in training processes.

3.6 Jaccard Coefficient

It is also known as Tanimoto coefficient, it measure the similarity of the text as intersection of terms divide by the union of terms. In union it negates the shared terms in both the document. Otherwise in the union part shared terms are considered twice.[10] The similarity measures in jaccard coefficient ranges from 0 to 1. And its corresponding difference is $D_J = 1 – SIM_J$. If the measure is 1 it means two are completely similar and if the measure is 0 it means two documents are completely different.

4. Implementation

Text messages coming from user sometimes it related to crime by detecting the crime related message before time we can easily stop the crime activity. For this we need to identify the crime related message first and detect under which category of crime that message comes. Simply means check the message is suspicious or non suspicious. Categorize the text message in the two classes from the terms present in that text message. Incoming message and documents contains some common terms on the basis of that incoming message is classify into the more similarity document.

The documents are one is collection of suspicious message text file and another is non suspicious text file. Check incoming text message terms in both the documents. Calculate the jaccard coefficient for both the document which one having more value that means it more similar to that if the incoming message having the jaccard coefficient value is higher with suspicious text file then classifies that text messages into the suspicious message and vice a versa.

These two document file each contains more than the 100 messages in it. When the incoming messages arrived then first check jaccard coefficient with the suspicious message text file and store that value with the formula: [10]

$$SIM_J(\vec{t_a},\vec{t_b}) = \frac{\vec{t_a}\cdot\vec{t_b}}{|\vec{t_a}|^2+|\vec{t_b}|^2-\vec{t_a}\cdot\vec{t_b}}.$$

Let "t_a" is a term in incoming message and "t_b" is the terms in the text file. Calculate similarity score for each document and classify incoming message into the most similar category.

Example: Plan to attend marriage of New Delhi in this month.

First pre-processing is done on the text message. After the stemming, stop words removing only gets the words like [plan, attend, marriage, New Delhi, month]. Calculate the Jaccard coefficient for these terms with suspicious text file and also for the non suspicious text file. By comparing both the score categorized that message into which one has highest score.

5. Comparison

Classification of the text data is the challenging field it needs three techniques text mining, natural language processing and the machine learning algorithm to extract the knowledge from the text data. But first issue is representation of data most of the literature gives the statistical of syntactic solution. While extracting the feature there is lots of redundant feature in it some are irrelevant feature which degrades the result as well as increase the cost of computation. Above describe algorithm have its own advantage and disadvantage.

The one technique hybrid or single technique proposed recently which shows the good result, for more efficiency we have to explore that technique. Among all the technique the SVM, NB and k-NN shows most appropriate result in existing literature survey. However the SVM is considered as the most effective classifier its work on the Structural Risk Minimization (SRM) which minimizes error rate. And it has some problem in kernel selection and parameter tuning. The NB approach is good for the email categorization, spam filtering, it required small amount of training data. Naïve bayes approach is easy to implements but if the features are highly correlated and does not considering the word count in that case it show very poor result.

In case of k-NN if the pre-processing is proper then this method shows the good results. It can suitable for n number of documents. But difficult to find the value of k in this method it takes to long time. Jaccard coefficient shows good result for categorizing the message into two categories. These methods first identify the terms and then only check for those terms in two documents with simple formula which gives result in minimum time, efficiency is high.

6. Conclusion

Some more research is required for performance improvement and more accuracy of the classification process. Classification and clustering of semi-structured documents is challenging. By reducing the training and testing time and improves the accuracy of the classification precision and recall. By studying all the techniques learn that every technique has some drawbacks and advantages with some specified condition. Among all of the some shows the good results. While choosing the technique we have to consider the advantage as well as limitation of each technique.

References

[1] Chauhan Shrihari R, Amish Desai, “A Review on Knowledge Discovery using Text Classification Techniques in Text Mining”, International Journal of Computer Applications (0975 – 8887) Volume 111 – No 6, February 2015

[2] Vandana Korde,”Text classification and classifiers:” International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012”.

[3] Zakaria Elberrichi, Karima Abidi, “Arabic text categorization: a comparative study of different representation modes.” Int. Arab J. Inf. Technol. 9(5): 465-470 (2012)

[4] S. Subbaiah "Extracting Knowledge using Probabilistic Classifier for Text Mining" Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22, IEEE-2013

[5] M. Janaki Meena , K. R. Chandran "Naive Bayes Text Classification with Positive Features Selected by Statistical Method" ©2009 IEEE vaishali Bhujade, N.J.Janwe "knowledge discovery in text mining techniques using association rule extraction" International Conference on Computational Intelligence and Communication Systems, IEEE-2011

[6] Jiawei Han and Micheline Kamber “Data Mining Concepts And Techniques” ,Morgan kaufman publishers, San Francisco, Elsevier, 2011, pp. 285-351

[7] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee,” A Review of Machine Learning Algorithms for Text-Documents Classification”, Journal Of Advances In Information Technology, Vol. 1, No. 1, February 2010

[8] Fabrizio Sebastiani, “Machine Learning in Automated Text Categorization”, ACM Computing Surveys, Vol. 34, No. 1, March 2010.

[9] Lena Tenenboim, Bracha Shapira, Peretz Shoval “Ontology- Based Classification Of News In An Electronic Newspaper” International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria,June-July2008.

[10] Anna Huang, “Similarity Measures for Text Document Clustering”, NZCSRSC 2008, April 2008, Christchurch, New Zealand.

[11] Dasgupta, “Feature selection methods for text classification.”,In Proceedings of the 13th ACMSIGKDD international conference on Knowledge discovery and data mining,pp.230-239,2007.

Author Profile

Ankita R. Jadhao she is pursuning M.tech in Dept. of Computer Science and Engineering, Ramdeobaba College of Engineering and Managment **,** Nagpur**,** India. She passed B.E from Sipna College of Engineering Technology, Amravati, India. She had actively participated in various international conference and workshop. Her field of interest is Natural Language Processing.

Avinash J. Agrawal received Bachelor of Engineering Degree in Computer Technology from Nagpur University, India and Master of Technology degree in Computer Technology from National Institute of Technology, Raipur, India in 1998 and 2005 respectively. He has Ph.D. in Computer Science and Engineering from Visvesvaraya National Institute of Technology, Nagpur in 2013. His research area is Natural Language Processing and Artificial Intelligence. He is having 18 years of teaching experience. Presently he is Associate Professor in Shri Ramdeobaba College of Engineering and Management, Nagpur. He is the author of more than 50 research papers in International Journal and Conferences.