A Review Paper On Automatic Text And Image Classification For News Paper

Komal P. Bhoyar DMIETR Wardha.

Neha S. Hiwase DMIETR Wardha.

Sneha K. Ankurkar DMIETR Wardha.

Prof. Preeti Bhagat DMIETR Wardha.

Abstract- Now a day's use of digital data is increasing rapidly. Knowledge discovery and data mining have attracted a great deal of attention with a need for turning such data into useful information. Therefore by taking advantage of data mining we put forward a website for News Paper. In this work, we are applying the various algorithm of text mining and image mining which helps us to classify and categories the news according to type of news.

This paper presents an innovative and effective technique. In this it include the process of pattern deploying and pattern evaluation to improve the effectiveness for finding the relevant information. Web newspaper provides the valuable source of information. It may be in the format of text or may be in the format of image. In order to take the more advantage from the available information, we are applying the text mining and image mining technique. We present an approach which based on the extracting the individual information from the data and mine them separately.

Index Terms- classification algorithm, clustering, OCR technology image to text conversion, text classification.

I. INTRODUCTION

Now a days there are many newspaper and magazines have provide a various websites which provide the news and other material. It provides the valuable information and maintains its quality as well. Some of the search engine based on the clustering based tool for finding the related information.

In order to improve the complete page mining, our approach is based on the extracting the individual news from the web page and mining these separately. This should considerably increase the quality of the results, because these news are short and it content the relevant and descriptive keywords, and made by the humans which gives the higher quality of the classification compare to what we get from the arithmetic calculation technique to do this task.

The removal of the non-relevant information also makes the news item extraction useful as way of data cleaning. Although news item extraction is relatively easy task for human who can do it just by visual inspection. In this paper we present a pattern based strategy that will provide the satisfactory quality as well as the being robust in the sense that it perform well for different news. To extract useful information and classify them according to their section text classification algorithms are applied. Before applying classification algorithm it is necessary to formatted them into same format because the news can be in the form of image or pdf.

II. LITERATURE REVIEW

In classification of text and image categorization lots of work is done. We observe various research articles on text and image classification. News paper contain lots of information and text mining is use to extract the valuable information.

News Item Extraction for Text Mining in Web Newspapers[1] is use to improve complete page mining. In this technique individual news is extracted separately and then mining is done on it. The work of extraction of important information is automatically performed by using pattern-detection strategies.

In Google Newspaper Search[2], the production pipeline is created which takes newspaper microfilms as an input and produce news articles as an output. Extracting information from page image and article segmentation is important in this paper.

Effective Pattern Discovery for Text Mining[3] , describes the effective pattern to reduce the low-frequency and misinterpretation problems for text mining. In this it uses two processes pattern deploying and pattern evolving to improve the discovered patterns in text documents.

As data grows rapidly day by day in Image Mining: Issues, Frameworks and Techniques [4], highlight the need of image mining. It defines the unique characteristics of image database that contain a whole new set of interesting and challenging research issues.

Image Mining: Trends and Developments[5] examine that how the image mining is different from image processing, pattern recognition, and traditional data mining. In this analysis of essential components which is needed for image mining is done and also how this component are interact with each other to describe new interesting image pattern are examine. Two main categories of image mining are defined named as a functional-driven image mining framework and an information-driven image mining framework.

After the study of various papers we saw some insufficiencies in previous methods. Such as in some methods the news paper converted into numbers of clusters and this clustered news items are not labeled. As it involve such type of manual work there is occur a difficulty in quality measurement and recall. In Pattern-based news item extraction text mining is perform for solving any task but it cannot give high accuracy.

Also we examine that many paper are perform only text mining or only image mining. In previous work it is found that they first store various types of data in database, then mining is perform on that stored data and finally that data get stored again in new way. So this process required relatively more time to perform mining technique and can cause insufficiencies.

III. PROPOSED METHODOLOGY

A. *Document Classification*

In real world classification of document is quite difficult to set it into proper category. In many field the classification of document is required for example newspaper, spam filtering Sort journals and abstracts by subject categories and language identification. Hence to reduce human efforts, text classification is implemented with the help of programming language. The text classification is the task to classify the document into predefined classes. There are two approaches that we are classifying the documents, Rule based approach and machine learning based approach. In rule based approach it define the set of rules to classify document and in machine based approach use a set of sample documents that are classified into the classes (training data), automatically create classifiers based on the training data.

Documents can be classifying according to subject or other attribute like document type, author, or date. But most of the articles are based on subject classification. In newspaper, news are classify according to their classes. There are many categories in newspaper like sports, science, astrology, crime, daily news etc. It is very difficult to classify the news manually hence the documents classification is very useful to this section. Following is the architecture of document classification. Documents can be in the form of text, pdf or images, hence first of all the inputted data must be formatted into same pattern like store the data in the form of text to the database because it is necessary when some queries and mining algorithms are applied to recognize strategic data from database. Following are the steps for document classification.

B. *Image Processing*

If the documents are present in the form of images which contain text, hence to extract the information in the form of pixel OCR technology is used. OCR technology is used for character separation, it segments the document into individual connected component then each connected component is separated into individual characters. After separating individual character, apply preprocessing to collect the property of character such as character size, typical distance between characters, calculate pixel of single character etc. to recognize these properties JAI(Java Advance Imaging) API(Application Program Interface) tool is used. It is a user defined modules of java allows to develop an algorithm often using API developed software itself.

To implement classification or filtering algorithm for images it needs the properties of images hence these facilities are provided by Java Advance Imaging, although the API not part of full feature image processing software, their existing functions and extensions are ease of implementation with low cost and makes an attractive option for image processing algorithm development.

Optical Character Recognition is the electronically conversion of scanned images or printed text into computer readable text. It is widely use to collect data from original papers like passport, bank statements, business card or any printed records. This technology is also known as dynamic character recognition, real-time character recognition and intelligent character recognition.

To convert printed image or scanned image into text format three steps are required in OCR technology that are pre-processing, character reorganization and post-processing. Pre-processing is used to improve the chances of successful reorganization. It divides the words and character and establish baseline to separate the word. Then segmentation of image is done to separate the artifacts these artifacts are the broken piece of character or words. Segmentation is done simply by aligning the image to a uniform grid based where the vertical grid lines are intersecting black pixilated area. The white space indicates the space between two words hence it helps to make individual words. Character recognition calculates width and height of rectangle of black pixel, using JAI technology. After finding the properties of each character it finds which characters are nearly matches to the character store in a library or training sets this process called as post-processing. Matched characters are collected in a string array, which gives a full text document.

C. *Converting documents into no of words*

A document is a sequence of words; such documents can be represented as an array of words. As we provide the training set of words to classify the document we need the no of words present in a document. To identify the stop words, important words in a document, it split into word.

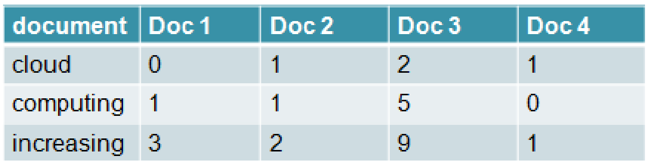
D. *Stop word removal process*

Words in a document that are frequently occurring but meaningless in term of Information retrieval called as stop words. In past decade when mining technique is used to find the important document by searching the documents it matches the word which is enter by user to words in various documents and it gives less accurate answer because most sentence contain similar percentage of stop words and these are most frequent words in English grammar like “I”, “has”, “the”, “who”, “it”, “or” etc.
Removing stop word helps to decrease the size of index and at the time of calculating frequency of words stop words are not present hence the output get more correctly. There are set of stop words used to eliminate the stop words present in documents. This technique is very useful for mining process.

E. *N-gram algorithm*

N-gram is a sequential list of n words, used to encode that the phrase will appear in the future. N-gram is used for approximate matching of words by converting the sequence of words into set of N-gram. These set of N-gram gives the frequency of words in document. This technique is use to provide the information about frequency of words for mining process various types of mining algorithms are used to mine the document this N-gram technique is very effective for inputting to the mining process. Following is one of the example of N-gram algorithm which finds the frequency of words.

EX: Cloud computing is increasing day by day.



document	Doc 1	Doc 2	Doc 3	Doc 4
cloud	0	1	2	1
computing	1	1	5	0
increasing	3	2	9	1

Fig 1: Frequency of words using N-gram algorithm.

There are various types of n- gram process like 1-gram, 2-gram, 3-gram and so on. 1-gram is consist of all individual words in a one document, 2-gram gram contains 2 two phrase words in a document and so on following is one of the example for n-gram algorithm.

Example: A Major League Baseball game was held in Salt Lake City 40 years ago.
1-grams are: {a, Major, League, Baseball, game, was, held, in, Salt, Lake, City, 40, years, ago}
2-grams are: {a Major, Major League, League Baseball, Baseball game, game was, was held, held in, ... }
3-grams are: {a Major League, Major League Baseball, League Baseball game, Baseball game was,...}
Similarly we can count 4-grams, 5-grams and so on.

The count of n-gram also use for auto complete keywords or predict people’s next search keyword.

F. *Naive Bayes Classifier*

The Naive Bayes classifier is a simple classifier which is based on Bayes theorem. It is one of the most basic text classification techniques but performs well in many complex and real world problems with various applications in email personal email sorting, spam detection, document categorization etc. Naive Bayes classifier is very efficient since it is less computationally intensive in both CPU and memory and it requires a small amount of training data. The training time with Naive Bayes is significantly smaller as compare to alternative methods. This is largely used because of independent assumptions which allows all parameters or attributes of documents for each to be learn separately.

In the case of document classification every features of documents is proportional to the size of training data set, the number can be increases of training data set in many cases so naive Bayes can be very efficient for document or text classification. The probability model for classifier is conditional model $p(C|F_1,.......,F_n)$.

Where, C is the class variable which classify the document, and F_1 to F_n are several features. The problem is that if the number of features n is large or when a feature can take on a large number of values, base on such a model on probability tables is infeasible, therefore reformulate the model to make it more tractable. Using Bayes theorem this can be written as,

$$p(C|F_1,\dots,F_n) = \frac{p(C)\ p(F_1,\dots,F_n|C)}{p(F_1,\dots,F_n)}.$$

The naive Bayes classifier combines this model with the decision rule. This is known as *MAP* (maximum a posterior) decision rule. The corresponding classifier is the *classify* function defined as follows:

$$\mathrm{classify}(f_1,\dots,f_n) = \underset{c}{\mathrm{argmax}}\ p(C=c)\prod_{i=1}^{n} p(F_i=f_i|C=c).$$

This naive Bayes classifier is use to categories the news item according to their section. Following example shows how documents are classify. It consider a one news item which contain the sport information, some heavy or important words are recognize from that document after removing stop words and calculating n-gram. This heavy words like “ball”, “game”, “play”, “dance” etc. are match with training sets provide to every classes, this extracted heavy words features are match with the features of “sports” class hence it predict the their category and store in the database.

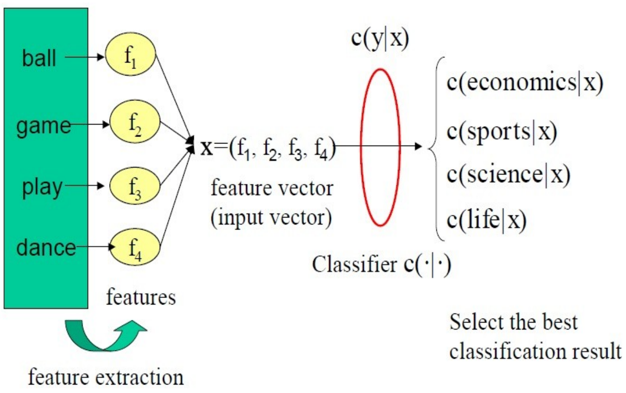


Fig 2: classification of documents using Naive Bayes algorithm

G. *Clustering using K-means algorithm and LWC algorithm*

Clustering is similar to classification in which data are in group form, grouping is accomplished by finding similarities between data according to characteristics found in the actual data these groups are called as clusters. Like tuples or records in a database are grouped together. There are many algorithms are used to make a clustered data K-mean is one of the algorithm use for clustering.

K-mean is an algorithm classifies or groups your object into k number of group based on attributes/features. K is positive integer number; the grouping is done by minimizing the sum of squares of distance between data and the corresponding cluster centroid. It takes any random object as an initial centroid or first k object can also serve as initial centroid. Then K-means algorithm will do the three steps until convergence first it determine the centroid coordinator then determine the distance of each object to the centroid and last group the object based on minimum distance that means find the closest centroid.

This algorithm is used to clustered classified data form naive bayes algorithm into database with its special attributes for retrieve the news item to user view and use the clustered data and their special attributes at the time of searching.

Similarly LWC algorithm is also used to cluster data but also sequencing the data. The LWC algorithm work similar to K-mean algorithm but it used complex analyses of these high-speed data streams such as clustering and outlier detection, classification, frequent item sets and counting frequent items. The LWC means Light Weight Clustering is included in K-mean algorithm for getting high threshold high accuracy in lower running time.

When new classified data element is arrived, the algorithm searches for nearest instance already in the main memory to a pre-specified distance threshold. The threshold means similarity measure acceptable by the algorithm to consider two or more elements as one element according to the element attributes values. This algorithm is used to searches nearest data matching document which having highest similar words.

IV. CONCLUSION

In text classification we have to perform the clustering of the news items which are from same category. There are two aspects which present in extraction of news items is the removal of the non-relevant information and also makes the news item extraction useful as a way of data cleaning and also space require for storing the news items in the database is smaller than that of the method used in previous processes.

In this we will use the various technique and algorithm which is easier than the previous one. By studying the previous technique the quality of news item extraction we conform the quality and the robustness of our strategies.

V. REFERENCES

[1] News Item Extraction for Text Mining in Web Newspapers Kjetil Nørv°ag∗ and Randi Øyri .Department of Computer and Information Science Norwegian University of Science and Technology ,7491 Trondheim, Norway.

[2] Google Newspaper Search – Image Processing and Analysis Pipeline Krishnendu Chaudhury, Ankur Jain, Sriram Thirthala, Vivek Sahasranaman, Shobhit Saxena,Selvam Mahalingam.

[3] Effective Pattern Discovery for Text Mining Ning Zhong, Yuefeng Li, and Sheng-Tang Wu

[4] Image Mining: Issues, Frameworks and Techniques Ji Zhang Wynne Hsu Mong Li Lee, Department of Computer Science, School of Computing National University of Singapore . Singapore, 117543.

[5] Image Mining: Trends and Developments Ji Zhang Wynne Hsu Mong Li Lee School of Computing National University of Singapore . Singapore 117543.

[6] http://www.aboutdm.com/2012/12/introduction-to-n-grams.html

[7] http://en.wikipedia.org/wiki/N-gram

1. Prof. Preeti Bhagat
Lecturer of DMIETR wardha.
2. Komal Bhoyar
Student of DMIETR Wardha, BE CSE 8th sem.
3. Neha hiwase
Student of DMIETR Wardha, BE CSE 8th sem.
4. Sneha Ankurkar
Student of DMIETR Wardha, BE CSE 8th sem.