Document Categorization Using Reinforcement Learning

Sukhdeep Kaur

Abstract— Automatic categorization of documents has become an important research issue since the explosion of digital and online text information. The text information is stored in the form of files and documents and to extract this particular information, a well-defined process is followed. The organization of documents is not so easy task. So there is need to present these documents in a categorized form. A particular text document is categorized, on the basis of high frequency words. So, the main idea is to convert the unstructured documents into structured form. The final goal is to achieve high accuracy in the formation of categories by following particular steps.

Index Terms— Document Categorization, Data Mining, Pre-processing.

I. INTRODUCTION

Document categorization is a process to categorize the document based on its content. In modern arena Document categorization is most critical problem because most of the organizations perform the process of categorization manually. Everyday Hundreds of files and e-mail messages are saved by the employees of organizations**.** So there is a need of automatic classification or categorization of documents [1]. Automatic Document categorization reduces a lot of burden of company or organization .Document classification, also known as document categorization; is a process of assigning documents to categories based on their content matching. A particular document may fit into two or more different categories. A non-categorized document may increase the searching time. So, there is a need of such system which will automatically generate categories and consequently burden over the employee's reduces.

It also reduces the search time that spent to search the document. Document categorization process is also known as text classification. Following diagram gives the inner view of document categorization, which shows how the no of documents are assigned categorizes. This is like a matrix where the documents are arranged row wise and columns represent the particular categories. In this diagram, D1, D2.......Dn represents documents and C1, C2 ....Cn represents Categories.

	D1	**D2**	**D3**	**D4**	**Dn**
C1					
C2					
C3					
C4					
Cn					

Fig.1 Diagram showing the brief view of Document

Document categorization is the basic need of all organizations because not only it reduces the search time but it also allows the users to easily access information from any type of text document. The main goal of document categorization is to allow a user to access the information that is present in the document .This feature can be used in the future to perform many useful tasks [3]. Document categorization (DC) process is very useful to various fields. The fields like business field and technological field. The process of document categorization in past was very time consuming process. It required several resources to do the process manually. This manual process increases the time to categorize the documents into the different classes. But now through automatic classification of document into the different categories a lot of effort and time is saved.

II. PREVIOUS WORK

The literature review starts with the classical methods of document categorization. Document categorization is not a new idea. It has been developed in 1950.The main purpose of document categorization is convert unstructured document into structured form .so various approaches can develop to solve this problem

Y. h. and a. k. Jain (1998) presented the classification of text documents. In this paper author has discuss about problems of classification. This paper has discussed the various approaches for the classification of documents such as naive bayes, decision tree. These approaches apply to different types of seven categories. The approaches Navies bayes and subspace method are compared after implementing them on seven categories.

Debnath Bhattacharya (2000) has discussed the history of document categorization. In this paper author has showed different approaches to solve problem of document categorization.

Manuscript published on 30 June 2012.
* Correspondence Author (s)
Ms.Sukhdeep kaur*, student of M-Tech(CSE) from Lovely Professional University (India). She completed her B-Tech(CSE) in L.P.U Itself. Lovely Professional University Punjab, India.

This paper describe different algorithm to convert unstructured document to structured document. The approaches used for same are lexical chain Method, linearization method and neural network approaches. It explains how it can work and also explain the limitations of methods.

Arzucan augur (2002) has represented supervised and unsupervised learning approaches to text categorization. This paper various approaches of supervised and unsupervised can be discussed and used. Automatic categorization of documents has become an important research issues since the explosion of digital and online text information. There are mainly two machine learning approaches to enhance this task: supervised approach, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labeled documents; and unsupervised approach, where there is no need for human intervention or labeled documents at any point in the whole process.

Hang lie, kanji Yamanishi (2002) has discussed the framework of Document classification using finite mixture model. It is a new method to document to classify the document into categories. In this it describes finite mixture model based on soft clustering of words. Then conducting the statistical hypothesis. The main problem is to classifying the document into no of categories. Each categories also determine already containing the newly documents also determine which category contain newly document to be assigned. To address this issue, the method is also called the hard clustering. But this method is degrading classification of results. But this problem can be solved by soft clustering, soft clustering uses the finite mixture model .This model classifying the document based on the soft clustering of words.

III METHODOLOGY

In the proposed system the documents are categorized using the reinforcement learning technique that provides efficient solution than other existing techniques. For implementing this research idea various resources are needed. The methodology for document categorization involves document collection, pre-processing of data, term weighting, and dimension reduction

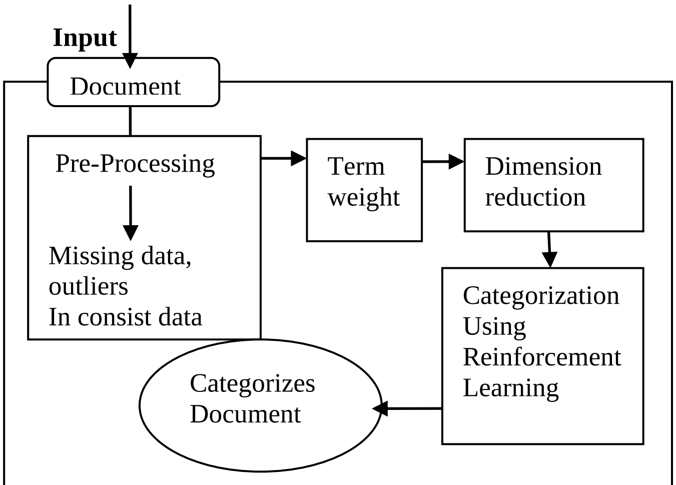


Fig2. Architecture of document categorization

Step 1: Collect the documents from different resources that can be required for process of document categorization.

Step 2: Next step consists of pre-processing on documents. This step is removing certain things from the documents like:-

- Missing data
- Smoothing and filtering data by removing stop words
- Stemming

Pre-processing of data is important because the documents that are collected from different sources contains missing data, and inconsistent data, stop words so all these things are needed to get removed from the document and hence improved results of categorization are obtained. Pre-processing of data is needed to improve the quality of data.

Step 3: In this step term weighting is performed. This process can be done to identify features (most frequently used words), which further means to check the feature component of original document and in the next step term weighting will be performed. This method, the weight of a term in a document is equal to the number of times the term appears in the document, i.e. to the frequency of the term in the document. To estimate the usefulness of weights by adding up of each terms that present in sentence. This can be done by using the term frequency (TF). This can identify the presence of terms in documents.

$$F=W_i(O_j) \qquad (1)$$

The relevance of a word to the document is proportional to the number of times it appears in the document. The discriminating power of a word between documents is less, if it appears inmost of the documents in the document collection. This method, the weight of a term in a document is equal to the number of times the term appears in the document, i.e. to the raw frequency of the term in the document

Step 4: Next step is dimensionality reduction that can be done for purpose of to reduce the dimension of document. In purpose of use the dimension reduction to discard the irrelevant and redundant features that leads to reduce the better results of document categorization.

The dimension reduction is important because the document have multiple or thousand of features so that to reduce the no features to select some best terms from document .suppose to choose the n best term and assign weights to terms on basis of to find frequency of these selected best terms from the document. so helps to reduce the features of document Dimension reduction that can be done by using of TermStrength that can find the closely related terms in document. Terms strength is calculated as

$$TS= P\,(T \in Di/T \in Dj) \qquad (2)$$

Step 5: To apply the categorization technique to documents for purpose of categorizing the document. To use the matrix in which all documents can be represents in row wise or direction and all terms that present in the document that can be show as column. To implement the categorization of documents by this technique.

Consider matrix Q in which documents can be denoted as d1, d2..........dn and terms can be represented as t1, t2...........tn.
Q [da*tb] = da (tb)
If tb belongs da
Then

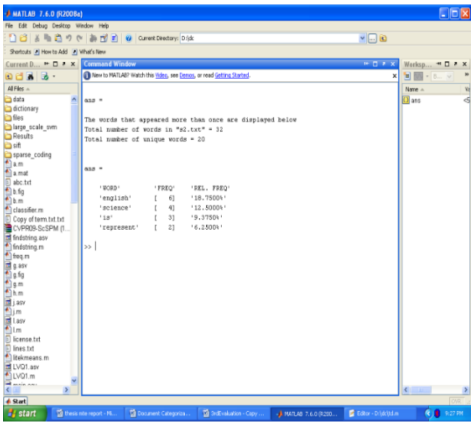
Return 1

In reinforcement learning, supervisor gives rewards to agents on the basis of observation. In proposed document categorization system, rewards are given in the form of 0 or 1 in matrix form. If a particular document consists of the category then reward 1 is given, otherwise reward 0 is given. Now on the basis of this categorization is performed. Hence, proposed idea is using reinforcement learning.

IV. RESULTS & DISCUSSIONS

After implementing the algorithm and methodology, results can obtained by using tool matlab. Matlab for document categorization is used to categorize the documents. In matlab use predefined function to solve the problem of categorization

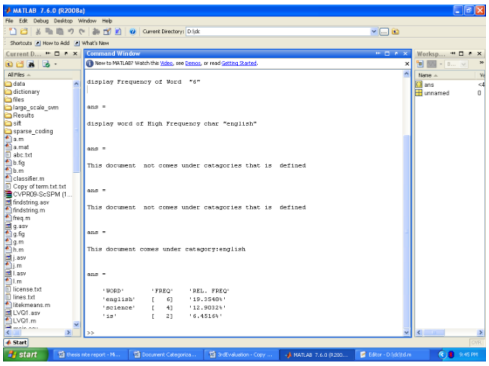
A. Next step is to calculate the frequency of words, the words that can occur more frequently in text file. This snapshot of system showing words according to the frequency. The word has high frequency it can display top as comparison to other words that have low frequency.
It can display output in matlab as following:



```
ans =

The words that appeared more than once are displayed below
Total number of words in "s2.txt" = 32
Total number of unique words = 20


ans =

    'WORD'          'FREQ'     'REL. FREQ'
    'english'       [   6]     '18.7500%'
    'science'       [   4]     '12.5000%'
    'is'            [   3]     '9.3750%'
    'represent'     [   2]     '6.2500%'

>>
```

Fig4.1 Figure showing Terms arranged according to frequency.

B. In this if frequency word belongs to document then show output, document comes from this predefined category otherwise. It can't belong to that pre-defined category. The Pre-defined categories are English, science and sports. The task of document categorization is performing by using of these different categories. In this system document is categorizes into category on basis of most frequently Occurring words in document.



```
display Frequency of Word  "6"


ans =

display word of High Frequency char "english"


ans =

This document  not comes under catagories that is  defined


ans =

This document  not comes under catagories that is  defined


ans =

This document comes under catagory:english


ans =

    'WORD'        'FREQ'     'REL. FREQ'
    'english'     [   6]     '19.3548%'
    'science'     [   4]     '12.9032%'
    'is'          [   2]     '6.4516%'

>>
```

Fig4.2 Figure showing output of document belong to category

V. CONCLUSION & FUTURE SCOPE

The proposed methodology for document categorization isusing categorizationalgorithm which has provided very interesting results .After the implementation of proposed methodologyand algorithm, it is very muchclear that, how to use the data mining approaches to categorize the documents into different categories based on the content. Traditional software used various approaches to categories the documents but their respective algorithms were not accurate enough .Traditionally the categorization process was done manually by the organization's staff. Firstly there was need to read the documents properly and then there was the task of assigning number of categories according to the predefined categories. So, the prime objective of doing this research was to reduce this burden. The system or software for document categorization can be implemented in applications like topic spotting, email routing language, guessing ad, spam filtering etc. In future this research idea can be extended to create such a system which can provide more detailed information about the categories such as: - If the existing system is telling us the category of a document such as science, then in the future it should provide us more specific category such as:-Chemistry, Biology, Physics etc. So the entire focus will be on getting more specific and particular information from the mixed set of documents. Hence more refined results will come into picture.

REFERNCES

[1] Arzucan augur (2002), "supervised and unsupervised Learning approaches to text Categorization".

[2] Debnath Bhattacharya "unstructured document Categorization" , international journal of Information Science.-30697 Korea.

[3] Dina Goren-Bar and Tsvi Kuflik (2005)" User subjective categorization using learning vector Organization and self organizinmaps,56(4):345–355, 2005.

[4] Eui-Hong (Sam) Han and George Caryopsis "Centroid based technique for document Classification". [5] G.S Thakur (2008), "Framework for document Categorization", International Journal of Information Technology.

[6] Hang lie, kanji Yamanishi (2006**),** "Document Classification using finite mixture model".

[7] Heide Britiicher, grehard knolmayer, Marc-Andre Mittermayer (2008) “document classification methods for organizing the explicit knowledge”.

[8] Ken Williams and Rafael A. Calve in 2006, “New Framework for document categorization” , New York, Yorktown Heights, NY 10598, USA.

[9] Manhood soltani, Mohammad taher (2009) “Classification of textual document using learning Vector organization”,978-1-4244-4538-7.

[10] Muhammad Fahad Umer and M.Sikander Hayat (2007) “classification of textual documents by using learning vector organization”, information Technology journal 6 (1):154-159.

[11] Marcel Rusinol and josep Lead’s (2009) “logo Spotting by Bag-of-words approach for document Categorization”.

[12] Philip S. Yu, Bing Liu Wee Sun Lee, “partially Supervised technique for the Categorization of text Documents”.

[13] Quire Zhang and jinghua (2008), “Research on medical document categorization”,International Journal of Information Technology.

[14] Riel.A, boonyasopon (2009)”Knowledge mining approach to document classification” *AIJSTPME (2009) 2(3): 1-10.*

[15] S.R.Suresh, T.Karthikeyan, D.B.Shanmugam, J.Dhilipan” Text categorization

16] Zhihang Chen (2006), “Neural network approaches to text documents categorization” july16-21(2006).

Ms. Sukhdeep kaur is a Student of M-Tech (CSE) in Lovely Professional University (India).She completed her B-Tech (CSE) in **L.P.U** itself. She is doing specialization in Internet Technologies. Her area of interests is:-Data mining, Artifical Intelligent Systems.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)