# Self Organization of web Document

Shuaib Khan[1], Vijay[2]
[1]Graphic Era University, Dehradun
[2]I.T. Department, Graphic Era University, Dehradun
([1]shuaibkhan.it@gmail.com, [2]Vijay.movfrose@yahoo.com)

***Abstract***- Classification of document is an important area for research, as large amount of electronic documents are available in form of unstructured, semi structured and structured information. Document classification will be applicable for World Wide Web, electronic book sites, online forums, electronic mails, online blogs, digital libraries and online government repositories. So it is necessary to organize the information and proper categorization and knowledge discovery is also important. This paper focused on the existing literature and explored the techniques for automatic documents classification i.e. documents representation, knowledge extraction and classification. In this paper we propose an algorithm and architecture for automatic document collection.

Keywords- **Text mining; Web mining; Information retrieval.**

## 1 INTRODUCTION

Today web is the main resource for the text documents. The amount of textual data available to us is consistently increasing; approximately 80% of the information of an organization is stored in unstructured textual form in the form of reports, email, views and news. Information intensive business processes demand that we transcend from simple document retrieval to "knowledge" discovery. The need of automatically extraction of useful knowledge from the huge amount of textual data in order to assist the human analysis is fully apparent. Market Trends based on the content of the online news articles, sentiments, and events is an emerging topic for research in data mining and text mining community.

We have a set of training records D = {X1, . . . , XN}, such that each record is labeled with a class value drawn from a set of k different discrete values indexed by {1 . . . k}. The training data is used in order to construct a classification model, which relates the features in the underlying record to one of the class labels. So extracting information from many web resources and proper categorization and knowledge discovery is an important area for research.

The classification problem assumes categorical values for the labels, though it is also possible to use continuous values as labels. The latter is referred to as the regression modeling problem. The problem of text classification is closely related to that of classification of records with set-valued features; however, this model assumes that only information about the presence or absence of words is used in a document. In reality, the frequency of words also plays a helpful role in the classification process, and the typical domain-size of text data (the entire lexicon size) is much greater than a typical set-valued classification problem.

## 2. LITERATURE REVIEW

A wide variety of techniques have been designed for text classification. In this chapter, we will discuss the broad classes of techniques, and their uses for classification tasks. We note that these classes of techniques also generally exist for other data domains such as quantitative or categorical data. Since text may be modeled as quantitative data with frequencies on the word attributes, it is possible to use most of the methods for quantitative data directly on text. However, text is a particular kind of data in which the word attributes are sparse, and high dimensional, with low frequencies on most of the words. Therefore, it is critical to design classification methods which effectively account for these characteristics of text. In this chapter, we will focus on the specific changes which are applicable to the text domain. Some key methods, which are commonly used for text classification, are as follows:

### *2.1. Decision Trees*

Decision trees are designed with the use of a hierarchical division of the underlying data space with the use of different text features. The hierarchical division of the data space is designed in order to create class partitions which are more skewed in terms of their class distribution. For a given text instance, we determine the partition that it is most likely to belong to, and use it for the purposes of classification.

### *2.2 Pattern (Rule)-based Classifiers*

In rule-based classifiers we determine the word patterns which are most likely to be related to the different classes. We construct a set of rules, in which the left-hand side corresponds to a word pattern, and the right-hand side corresponds to a class label. These rules are used for the purposes of classification.

### 2.3 SVM Classifiers

SVM Classifiers attempt to partition the data space with the use of linear or non-linear delineations between the different classes. The key in such classifiers is to determine the optimal boundaries between the different classes and use them for the purposes of classification.

### 2.4 Neural Network Classifiers

Neural networks are used in a wide variety of domains for the purposes of classification. In the context of text data, the main difference for neural network classifiers is to adapt these

classifiers with the use of word features. We note that neural network classifiers are related to SVM classifiers; indeed, they both are in the category of discriminative classifiers, which are in contrast with the generative classifiers [9].

2.4. Bayesian (Generative) Classifiers

In Bayesian classifiers (also called generative classifiers), we attempt to build a probabilistic classifier based on modeling the underlying word features in different classes. The idea is then to classify text based on the posterior probability of the documents belonging to the different classes on the basis of the word presence in the documents.

2.5. Other Classifiers

Almost all classifiers can be adapted to the case of text data. Some of the other classifiers include nearest neighbor classifiers, and genetic algorithm-based classifiers. We will discuss some of these different classifiers in some detail and their use for the case of text data.

## 3. PROPOSED WORK

In this proposed methodology, HTML documents are used as an input and converted into text file. Some lines are selected from the text file and forwarded to the Tokenizer. That Tokenizer removes the stop-words from these lines. These tokens again forwarded to word matcher that matches the words with the index files. If 40% and above tokens matched with one index file then document is belongs to concern index topic and get moved to concern topic repository.

In this way we are doing the document classification according to content of HTML pages.

### 3.1 MODULES USED IN PROPOSED METHODOLOGY

In this section, describing the following modules in brief which are used in proposed architecture. They are-

### 3.2 Convertor & Parser

This module is used to convert the HTML documents to text file. The parser will remove the HTML tags from the text file and select the starting eight lines and also the middle eight lines from the same text file.

Proposed Architecture of Automatic document collection

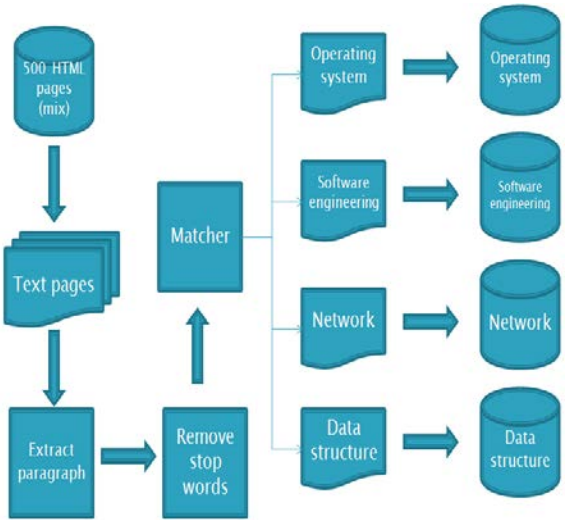


Figure 1: Proposed architecture

```
Parser ()
{
  If(HTML page)
  {
    Txt file  ←  HTML file;
  }
  If(txt File)
  {
    While(EOF)
     {
       Text file ← Remove HTML tags;
     }
  }

If (Text file)
{
   While (EOF)
   {
      Line_no ++;
    }
  Mid_no = (1+line_no)/2;
  While(temp<8)
   {
     String1 ← Lines;
   }
     Temp=mid_no;
  While(temp<temp+8)
   {
     String2←Lines;
   }
  FString=String1+String2;
}
Return(Fstring)
}
```

3.1.1 Tokenizer

This module will remove the stop-words from the lines. After removing stop-words it generates the tokens.

3.1.2 Word Matcher

This is the main module of proposed architecture. It matches the tokens with the indexes available. If the matched tokens equal to or greater than 40% of index words then it belongs to concern index repository.

Proposed algorithm for Word-matcher()

```
Word_matcher()
{
  If (token)
  {
    While(token == word dictionary)
    {
      calculate match token percentage;
    }
  }
    If (match>= 40%)
    {
       Store document in specific domain cluster;
    }
     Else
     {
       Parse whole document to generate token;
     }
}
```

3.2 Parse whole file

If the matched tokens are less than 40% of index words then the whole text file will be recycled. And send to parser then tokens will be generated from the whole text file.

3.3 Index

In this module indexes are available for different topics, which get matched with tokens.

3.4 Domain specific clustering

When matching phase is completed then 40% matched or greater matched document moved to concern specified domain name repository.

4. CONCLUSION

In this paper, we have presented an approach that works as self organization of documents. It classifies the HTML documents on content based technique. It extracts the paragraphs from web pages and then generates the tokens. Our preliminary experimental results demonstrate that classification on information extraction based technique working well.

Text mining is a relatively new research area at the intersection of natural-language processing, machine learning, data mining, and information retrieval. By appropriately integrating techniques from each of these disciplines, useful new methods for discovering knowledge from large text corporation can be developed

5. REFERENCES

[1] Aurangzeb Khan, Baharum B. Bahurdin, Khairullah Khan, "An Overview of E-Documents Classification", 2009 International Conference on Machine Learning and Computing IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

[2] S. Gopal, Y. Yang. Multilabel classification with meta-level features. ACM SIGIR Conference, 2010.

[3] M. James. Classification Algorithms, Wiley Interscience, 1985.

[4] A. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/~mccallum/bow, 1996.

[5] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

[6] S. Chakrabarti, B. Dom. R. Agrawal, P. Raghavan. Using taxonomy, discriminants and signatures for navigating in text databases, VLDB Conference, 1997.

[7] B. Liu, L. Zhang. A Survey of Opinion Mining and Sentiment Analysis. Book Chapter in Mining Text Data, Ed. C. Aggarwal, C. Zhai, Springer, 2011.

[8] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz. A Bayesian approach to filtering junk e-mail. AAAI Workshop on Learning for Text Categorization. Tech. Rep. WS-98-05, AAAI Press. http://robotics.stanford.edu/users/sahami/papers.html.

[9] A. Y. Ng, M. I. Jordan. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. NIPS. pp. 841- 848, 2001.

[10] J. R. Quinlan, Induction of Decision Trees, Machine Learning, 1(1), pp 81–106, 1986.

[11] P. Long, R. Servedio. Random Classification Noise defeats all Convex Potential Boosters. ICML Conference, 2008.

[12] S. A. Macskassy, F. Provost. Classification in Networked Data: A

**First A. Author** Mohd Shuaib Khan pursuing M. Tech. from Graphic Era University, Dehradun in 2010-2013. His research area includes Web Documentation. . He has five research publications to his credit in International and National journals/conferences of repute. Presently he is working as an Assistant Professor in the DBIT, Saharanpur.

**Second B. Author** Vijay received his M.Tech. in Software Engg. (First class with honours), He has the privilege of guiding undergraduate and postgraduate students for their research work. He has more than thirty research publications to his credit in International and National journals/conferences of repute. Presently he is working as Assistant Professor in the Department of Information Technology and Computer Engineering, Graphic Era University, Dehradun. His area of interest includes web information retrieval and system programming.