International Journal for Research in Applied Science & Engineering Technology (IJRASET)

New Search Techniques for Automatic Information Classification

Dr. S. Vijayarani[#1], Ms. N. Nithya[*2]
Assistant Professor[1], M. Phil Research Scholar[2],
[1,2]Department of Computer Science, School of Computer Science and Engineering,
Bharathiar University, Coimbatore, Tamilnadu, India.

Abstract- ***Information categorization has become one of the primary tasks of the text retrieval systems. The main objective of this research work is to classify the information which is available in tables. In this work, two new searching techniques, normal search and indexed search are proposed for categorizing the table information. The synthetic data set is created by using the information available in the Annexure II journal list dataset. The performance factors used here are classification accuracy, search time and outliers. From the experimental results it is observed that indexed search technique has produced better results in classification accuracy and outliers. For search time performance factor, the normal search technique required minimum time compared to index search.***
Keywords: Pre-processing, Classification, Dictionary, Normal search, Indexed search.

I. INTRODUCTION

Data mining can be defined as an extraction of useful knowledge from large data repositories. Major data mining domains are text mining, image mining, sequence mining, spatial mining, web mining, structure mining, graph mining and multimedia mining. Important and most popular data mining techniques are classification, clustering, summarization, time series analysis, outlier analysis and association rule generation. Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose to predict the class of objects whose class label is unknown. Classification may be preceded by the relevance analysis, which attempts to identify attributes that do not contribute to classification process, so those attributes can be excluded [1]. The text classification process consists of two steps; they are model construction and model usage. Model construction is nothing but describing the predefined classes and model usage is about unknown objects for classification. Before classifying the objects certain steps should be taken like data cleaning, relevance analysis, data transformation and reduction. Some of the real time applications of classifications are email classification, text classification, credit card analysis, medical diagnosis and target marketing. The main objective of this research work is to automatically classify the information in the table. Categorizing the information is one of the primary requirements of the text retrieval systems. This research work has two important steps; they are (i) Pre-processing and (ii) Classification. The first step of automatic information categorization is pre-processing. Pre-processing task is essential for text mining which transforms the complex text into information rich text for better knowledge interpretation by machines. Here the information is pre-processed using stemming techniques and stop word removal. The second step is classification; In order to perform classification, dictionaries are created for various disciplines which contains list of keywords. Here, two new searching techniques are proposed, they are (i) Normal search and (ii) Indexed search. These search techniques searches the dictionaries, and then the classification is performed. From the classification results of the proposed methodology it is observed that the Indexed search technique performs better than the normal search technique. The remaining portion of the paper is organized as follows. Section 2 gives the related works. Section 3 describes the proposed methodology. Experimental results and screenshots are discussed in Section 4. Conclusions are given in Section 5.

II. LITERATURE REVIEW

Jian Ma et.al [8] proposed a novel ontology-based text-mining approach to cluster research proposals based on their similarities in research areas. This method is efficient and effective for clustering research proposals with both English and Chinese texts. The method also includes an optimization model that considers the applicant's characteristics for balancing proposals by geographical regions. The proposed method is tested and validated based on the selection process at the National Natural Science Foundation of China. The results can also be used to improve the efficiency and effectiveness of research project selection processes in other government and private research funding agencies.

Shalini Puri et.al [15] proposed a new Fuzzy Similarity Based Concept Mining Model (FSCMM) and it is used to classify a set of text documents into pre - defined Category Groups (CG) by providing them training and preparing on the sentence, document and integrated corpora levels along with feature reduction, ambiguity removal on each level to achieve high system performance. Fuzzy Feature Category Similarity Analyzer (FFCSA) is used to analyze each extracted feature of Integrated Corpora Feature Vector (ICFV) with the corresponding categories or classes. This model used support vector machine classifier (SVMC) for classifying the training data patterns into two groups; i. e., + 1 and – 1, thereby producing accurate and correct results. The proposed model works efficiently and effectively with great performance and high - accuracy results.

S. M. Kamruzzaman et.al [9] proposed a new algorithm for text classification using data mining that requires fewer documents for training. Instead of using words, word relation, i.e. association rules from these words is used to derive feature set of pre-classified text documents. The concept of Naïve Bayes classifier is then used on derived features and finally only a single concept of Genetic Algorithm has been added for final classification. A system based on the proposed algorithm has been implemented and tested. The experimental results show that the proposed system works as a successful text classifier.

C. Ramasubramanian et.al [10] analyzed to make an effective Pre-Processing step to save both space and time requirements by using improved Stemming Algorithm. Stemming algorithms are used to transform the words in texts into their grammatical root form. Several algorithms exist with different techniques. The most widely used stemming algorithm is M.F Porter stemming algorithm. However, it still has certain drawbacks of handling Named Entities. This research work has improved its structure by refining with certain constraints, so that it improved the information retrieval system efficiency.

III. PROPOSED METHODOLOGY

Today the area of information technology has tremendously grown but the interpretation of the unstructured data is difficult to handle. Hence, in order to handle these kinds of data the automatic classification methods are essential to develop. The main objective of this research work is to automatically classify the information in the table. Categorizing the information is one of the primary requirements of the text retrieval systems. The proposed methodology of this research work is given in Figure 1.

A. Dataset Description

The synthetic dataset for this research work is created by extracting the information from the Annexure II journal list which is obtained from www.annauniv.edu/research website. This dataset consists of list of journal names with its related information for various disciplines. The number of instances of this dataset is 728 and it has four attributes; they are serial number, source title, ISSN number and country. The source title describes the name of the journals, the ISSN number refers to the International Standard Serial Number of a journal and the country refers to the place where the journal is published. In this research work the source title attribute is considered for automatic classification.

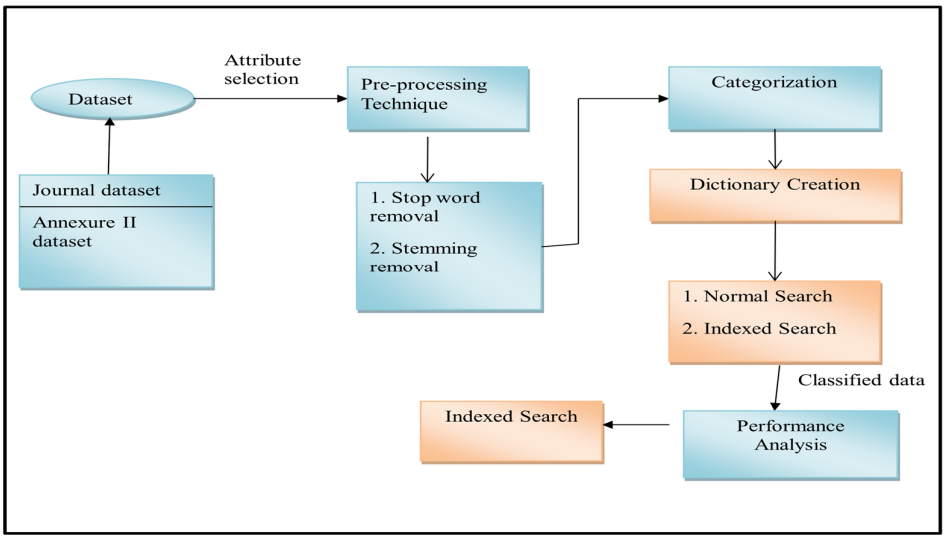


Figure 1: Proposed Architecture

B. Dataset Pre-processing

It is a preliminary processing of text data in order to prepare it for the primary processing or for further analysis. There are two

important steps in pre-processing; they are stop word removal and stemming.

1) *Stop word removal* - Stop words are language specific functional words which carry no information. It may be of the following types such as pronouns, prepositions, conjunctions. These kinds of word carry less meaning, so these words are filtered out in pre-processing technique.
2) *Stemming* - It is the process for reducing a word to their word stem i.e. base or root form. The stem is not to be identical to the morphological root of the word. This pre-processing technique is a common requirement in the areas of information retrieval systems and natural language processing.

1) *Stop word Removal Method:* Many of the most frequently used words in English are useless in Text Mining and Information Retrieval (IR). These words are called 'Stop words'. Stop-words, which are language-specific functional words, are frequent words that carry no information (i.e., pronouns, prepositions, conjunctions). In English language, there are about 400 to 500 Stop words. Examples of such words are 'the', 'of', 'and', 'to'. The first step of pre-processing is to remove these stop words [1]. Four types of stop word removal methods are available and these methods are used to remove stop words from the files [5].

   a) *The Classic Method:* The classic method is based on removing stop words obtained from pre-compiled lists [7].
   b) *Methods based on Zipf's Law (Z-Methods):* The important methods of Zipf's law are removing most frequent words (TF-High) and removing words that occur once, i.e. singleton words (TF1). It also considered removing words with low inverse document frequency (IDF) [7, 8].
   c) *The Mutual Information Method (MI):* The mutual information method is used for stop word removal that works by computing the mutual information between the given term and a document class. Low mutual information suggests that the term has low discrimination power and hence it should be easily removed [7, 8].
   d) *Term Based Random Sampling (TBRS):* This method was first proposed by Lo et al. (2005) to manually detect, stop words from web documents. The method works by iterating over separate chunks of data randomly selected. It then ranks terms in each chunk based on their format values using the Kullback-Leibler divergence measure.

In this research work, we have applied classic pre-processing method.

2) *Stemming Method:* Stemming techniques are used to find out the root/stem of a word. Stemming converts words to their stems, which incorporates a great deal of language-dependent linguistic knowledge. Behind stemming, the hypothesis is that words with the same stem or word root mostly describes same or relatively closer concepts in the text and so words can be conflated by using stems. For example, the words, user, users, used, using all can be stemmed to the word 'USE' [1].

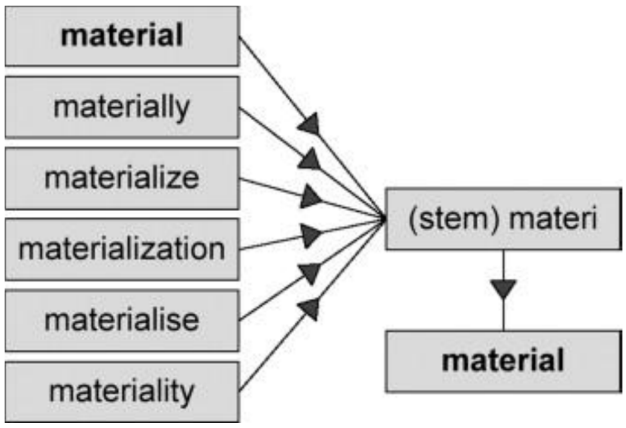


Figure 2: Stemming technique

Figure 2 represents the stemming process; it helps to identify the root of words. There are basically two types of stemming technique, one is inflectional other one is derivational stemming. Inflection Stemming is nothing but the form variation of a word under certain grammatical condition. Derivation stemming refers to the combinational affixes to an existing root or stem to form a new word.

3) *Tokenization:* Tokenization is the process of breaking up a sequence of strings into pieces such as words, keywords and phrases called tokens. Tokens can be in the form of individual words or phrases. In the process of tokenization, some

characters like punctuation marks are discarded.

a) *Text Categorization Methods:* Text categorization is a process that group text information into one or more predefined categories based on their title [1]. It has wide applications, such as email filtering, category classification for search engines and digital libraries. In this research work the text categorization is done using two methods, namely (i) normal search and (ii) indexed search. These two methods have been proposed and these methods reduced the level of difficulty when this information is classified manually.

b) *Dictionary Creation:* Generally dictionary is a macrostructure which consists of vocabularies in alphabetical list. The microstructure comprises various kinds of linguistic information on each word. In this research work multiple dictionaries are created and used for searching. The dictionary consists of 17 sub dictionaries of various disciplines namely accounts, astronomy, chemistry, civil engineering, commerce, computer science, english, environmental science, geology, maths, medicine, nanotechnology, physical education, physics, electronics, mechanics and cell biology. The total numbers of words in these dictionaries are 12830.

c) *Normal Search technique* In this research work the normal search verifies the tokenized words are found in the dictionaries or not. Typically, a simple function is applied to the key to determine its place in the dictionary. The instance S_i is taken one by one and each token T_m in S_i is compared with D_n dictionaries. The string $S_i(T_m)$ goes to the storage buffer where the dictionaries $D_1,D_2,D_3,\ldots D_n$ are stored and it searches the keyword K_i. If $S_i=K_i \in D_n$ then the particular S_i is labeled with the relevant class. If the keyword is not found then the instance is considered as the outliers O_n.There are two types of cases namely best case and worst case are considered for normal search.

i. *Best case:* The best case occurs when the search term is found in any one of the dictionaries from $D_1, D_2,\ldots D_n$.

ii. *Worst case*: The worst case occurs when the search term is not found in any kind of dictionaries from $D_1, D_2,\ldots D_n$.

```
Input : Source title S1,S2, .....Sn
        Tokens Tm = T1, T2, .....Tm
        Dictionary D1 , D2, .....Dn
Output: Classified Instances
          1. Discipline
          2. Outlier

Step 1: K=0;
Step 2: for (I = 1 to n)
        }
Step 3: Consider the source title one by one Si
Step 4: for (r = 1 to m)
{
Step 5: verify if ( S (Tr) is in Dj )
              {
              Assign a class label, (Disciplinej) into Si
              Go to step 4
             }
             }
Step 6: for (j=1 to n)
              {
Step 7: Verify if (Si in Dj )
               {
                 Assign a class label, (Disciplinej) into Si
                 Go to step 2
               }
               }
 Else {
            Assign a class label (outlier) into Si
            K = K+1 }
Step 8: // Outlier handling
Consider the outlier instances Oi
              For(i=1 to k)
            {
              Get(Oi) discipline from the user
              Assign a class label (Disciplinej) into Oi
             }
Step 9 : Stop the process
```

Figure 3: Pseudo code for Normal search technique

d) Indexed Search : Indexed Searching Algorithm indicate each array element with a particular index value and has that index value to the function in the form of M mod N (M is the index value and N is the user defined number). In this method, the list of significant keywords of various disciplines are stored in a single dictionary I_n. The pre-processed information S_i is compared with the indexed dictionary I_n which consists of the words arranging from W_1, W_2,…W_n. If S_i is not found in this indexed searching dictionary, then the normal search method is used D_1,D_2,…..D_n and searches for the keyword. If $S_i \in D_n$, then the information in the table is grouped into their relevant classes. If the normal search dictionary process fails to do this, then the instance is considered as outliers.

```
Input : Source title S1,S2, .....Sn
        Indexed table which contain important keywords with the max
entries of indexed table with discipline code
        Dictionary D1 , D2, .....Dn
Output: Classified Instances
          1. Discipline
          2. Outliers
Step 1: K=0;
Step 2: for (i = 1 to n)
Step 3: Consider the source title one by one Si
Step 3: for (j = 1 to Max_entry)
        {
Step 4:Verify if (Si in Ind_Tabj) then consider its relevant  discipline code
and put Si into that discipline
        }
        Else {
Step 5: for (j=1 to n)  {
Step 6: Verify if (Si in Dj )
        { Assign a class label, (Disciplinej) into Si
          Go to step 2   } }
          Else {
            Assign a class label (outlier) into Si
            K = K+1 }
             }
Step 6: // Outlier handling
          Consider the outlier instances Oi
       For(i=1 to k)
           {
       Get(Oi) discipline from the user
       Assign a class label (Disciplinej) into Oi
           }
Step 7 : Stop the process
```

Figure 4: Pseudo code for Indexed Search

IV. Experimental Results

e) Performance Measure: Performance measurement is generally defined as regular measurement of outcomes and results, which generates reliable data on the effectiveness and efficiency of programs. The performance measure in this research work is used to identify the best method for classifying the information. In order to measure the performance of the proposed search methods four different criteria are used; they are correctly classified instances, incorrectly classified instances, number of outliers and time taken.

C. Correctly classified instance

They are also called as true positive rate or the recall rate, which measures the proportion of actual positives which are correctly identified instances.

True positive = correctly identified

$$TPR = \frac{TP}{P} = \frac{TP}{(TP + FN)}$$

B. Incorrectly classified Instances is the false positive rateor error in the predicted data for the test set.

False positive = incorrectly identified

$$SPC = \frac{TN}{N} = \frac{TN}{(FP + TN)}$$

$$FPR = \frac{FP}{N} = \frac{FP}{(FP + TN)} = 1 - SPC$$

Where **P** is positive instances and **N** is negative instances

D. Outliers

Outliers in this research work refers to source title S_i which never falls into any one of the classes arranging from $C_1, C_2, \ldots C_{17}$. Outliers is the examination used to find the number of S_i which are unlabelled after classification process. These unlabelled instances are considered as outliers. These instances are stored in particular location S_x, from which the user can select the class for each instance S_i manually from the given classes C_n or he can also specify the class which he wishes.

E. Search time

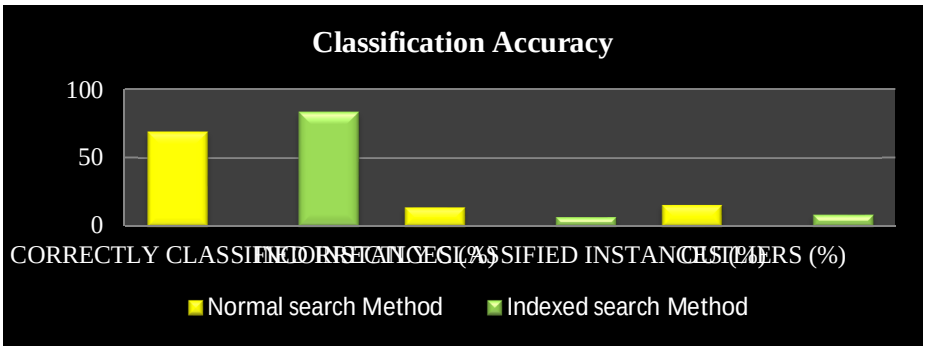
The time complexity of an algorithm quantifies the amount of time taken by an algorithm to run as a function of the length of the string representing the input. Time complexity is commonly estimated by counting the number of elementary operations performed by the algorithm, where an elementary operation takes a fixed amount of time to perform. Thus the amount of time taken and the number of elementary operations performed by the algorithm differ by at most a constant factor. Search time measures the amount of time required for searching and classifying the information.In order to find the efficiency of the proposed searching techniques the classification accuracy and the execution time performance factors are used. The classification accuracy is a measure to find how close the results of the classification matches the true categories of the information. Accuracy is estimated by applying the classifier to the dataset.

Table 1: Classification Accuracy Measure in %

METHOD	CORRECTLY INSTANCES (%)	INCORRECTLY CLASSIFIED INSTANCES (%)	OUTLIERS (%)
Normal search Method	69.50	14.28	16.20
Indexed search Method	83.93	7.28	8.79

Table 1 depicts the percentage of accuracy for classification task using the two proposed search techniques.

Figure 5: Classification accuracy



From the figure 5, it is analyzed that the Indexed search performance is better than the normal search technique. Therefore the Indexed search performs well because it attains lowest number of incorrectly classified instance and outliers when compared to normal search algorithms.

SEARCHING TIME

Table 2 shows the time utilization of the proposed searching techniques. Table 2: Time taken for searching

Method	Time taken(msec)
Normal Search Method	156
Indexed Search Method	208

Figure 6: Time taken

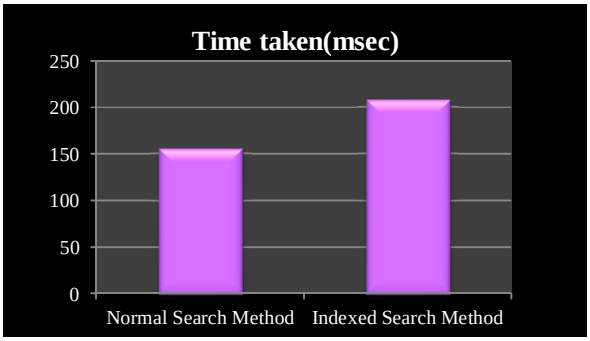


Figure 6 shows the search time for searching keyword by extracting the attribute source title. From the results, it is observed that the Normal search occupies minimum search time than the Indexed search.

V. CONCLUSION

The text databases consist of a huge collection of text documents, from several resources like news articles, digital libraries, web pages, emails and books. Due to the enormous size of the text data it is difficult to understand. Therefore, text mining techniques were applied to solve the arriving issues. In this research work intelligent methods are applied in order to extract the data patterns from the table and perform prediction to know the unknown classes. The research work analyzed the data extracted from Annexure II dataset using the two proposed methodology normal search and indexed search. The main objective of this work is to classify the information in the table accordingly to their classes. From the experimental results we can interpret that the Indexed search method performs better classification accuracy than the normal search. But as far as the time is concerned it is observed that the normal search method takes less time when compared to the indexing method. In future other classification algorithms will be used to perform the classification.

REFERENCES

[1] David D. Lewis and Yiming Yang (2004), RCV1: A New Benchmark Collection for Text Categorization Research.
[2] Deepajothi S, Dr.S.Selvarajan, A Comparative Study of Classification Techniques On Adult Data Set.
[3] Harish B S, S Manjunath and D S Guru, Text Document Classification: An Approach Based on Indexing, 2012.
[4] Hemlata Sahu, Shalini Shrma, Seema Gondhalakar, A Brief Overview on Data Mining Survey, Volume 1, Issue 3.
[5] http://en.wikipedia.org/wiki/Data_pre-processing
[6]Izzat Alsmadi, Izzat Alsmadi, Documents Similarities Algorithms for Research Papers Authenticity, 2012.
[7] James W. Cooper, Anni R. Coden, Eric W. Brown, Detecting Similar Documents Using Salient Terms, 2002.
[8] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu, An Ontology- Based Text-Mining Method to Clusters Proposals for Research Project Selection, IEEE transactions on systems, man and cybernetics- Part A: Systems and Humans, Vol 2, No.3, May 2012.
[9] Kamruzzaman S M, Farhana Haider, Ahmed Ryadh Hasan, Text Classification Using Data Mining, ICTM 2005.
[10] Ramasubramanian C, Ramya R, Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013.
[11] Robert S. Boyer & J. Strother Moore, A fast string Searching Algorithm, Communication of the ACM, October 1977.
[12] Samaneh Chagheri and Sylvie Calabretto, Document Classification Combining Structure and Content. Shalini Puri, A Fuzzy Similarity Based Concept Mining Model for Text Classification, 2011.
[13] Sindhiya B and N Tajunisha (2014), Concept and Term Based Similarity Measure For Text Classification And Clustering, Vol. 3, No. 1, February 2014.
[14] Shalini Puri, A Fuzzy Similarity Based Concept Mining Model for Text Classification, 2011.
[15] Srividhya V, Anitha R, Evaluvating preprocessing techniques in Text Categorization, International Journal of Computer Science and Application Issue, ISSN 0974-0767, 2010.