International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

# Document Organization using Naive Bayes Related Classifier

**R.Sathish Babu, R.Nagarajan**

***Abstract*: Document organization is necessary for better utilization of documents. The major problem of organization online documents is so complex because documents should be grouped into its appropriate group during its appearance on the web. Classification is one of the best solutions to organize the documents.**

***Naive Bayes categorization is playing a vital role in document organization. It is one of the simplest probabilistic Bayesian categorization and assumption that the effect of an attribute value on a given category is independent of the values. The document classification is the essential task of organization and necessary for efficient control of textual fact systems. The files may be classified as unconfirmed, supervised and semi supervised methods. In this paper, to review and study of various types of document organization approach using naive Bayesian classification and other related existing document organization methods.***

***Keywords*: Bayes, Document, Classification, Organization, web classification and neural network.**

## I. INTRODUCTION

The document classification is a supervised wisdom task, clear as conveying class labels to new documents based on chance suggested by a working outset of labeled documents. The document Categorization is the automatic classification of text under predefined categories or classes. Information Retrieval (IR) and Machine Learning (ML) techniques are used to assign keywords to the documents and classify them into specific categories. Machine gaining knowledge of avails us to categorize the files automatically. Information Retrieval avails us to symbolize the textual content as an attribute.

With the continuous growth in the web, locating relevant information becomes an increasingly difficult process. There exist about 30 billion pages on the Web which represent documents on different topics or different aspects of the same topic and introduce massive volume of online unstructured or semi-structured text with diverse information sources.

**Revised Manuscript Received on February 28, 2020.**
***** Correspondence Author
**R.Sathish Babu***, Assistant Professor, Department of Computer and Information Science, Annamalai University, Annamalai Nagar, India. Email: yeseswini.s@gmail.com.
**Dr. R. Nagarajan**, Assistant Professor, Department of Computer and Information Science, Annamalai University, Annamalai Nagar, India. Email: rathinanagarajan@gmail.com.



Manually organizing huge report bases is extremely tough, time ingesting, error prone, luxurious and is frequently no longer viable. Automated text categorization is a viable option for larger organizations which has got time and money as the main constraints. The document indexing, spam filtering, populating the hierarchical catalogues of web resources, document genre identification, automated essay grading, and categorizing newspaper ads are some of the important applications of Text Categorization in the field of science and technology. It is also used in the fields of finance, sports and entertainment and medical sciences. This paper deals with the comparative study on document categorization. It involves categorizing various document collections using Naive Bayes based on Bayes theorem, K- Nearest Neighbor methods and other related document classification algorithms, well known data mining techniques.

## II. EXISTING MODELS FOR DOCUMENT ORGANIZATION & CLASSIFICATION

Naive Bayes procedure intends the supreme and required plausible prospects for an official record to accord to a league. K-Nearest Neighbor method acquires the adjacent bystander that permeates to the corresponding tier by manipulating the Euclidean distance quantum. In this chunk, we categorized into three approaches like classification approach, web classification approach and hybrid approaches.

### 2.1 Classification Approach

#### 2.1.1 Naïve Bayesians Approach

Naive Bayes Approach (NBA) is a very transparent classifier which accumulates very prudent on numerical and textual data. It is paltry to actualize and computationally slashed when correlated to several spare allocation algorithms. One of the leading impediments of the classifier, it executes defectively when lineaments are highly tied in. So, it mistakes to examine the frequency of phrase contingency in the feature vector. The duct detriment of the Naive Bayes category method and its comparably low type of entire is comparable to other discerning procedures.

NBA demonstrated its progress in many exploration endeavors (Friedman, Sahami, McCallum and Nigam, Mladenic, Craven, Nigam ) was used as a yardstick in others (Twycross and Cayzer 2002), (Yang et al. 2002).

$$c_{NB} = arg^{max}_{c_i \in C} P(C_i) \prod_i P\left(\frac{f_i}{c_i}\right) for\ all\ c_i \in C \quad (1)$$

Applying an NBA, the operation opens by forging a probability distribution practicing the guidance documents set. This transportation is called a prior distribution.

When a recent test testimony is weighed, equation (1) is enforced to catch its division and the preceding dispensation is refurbished into a posterior distribution.

Suppositional, NBA has the merest error betwixt all the alternative classifiers. Withal, in convention one cannot consistently estimate appearance and class conditional independence.

### 2.1.2 Multivariate Bernoulli Model

The multivariate Bernoulli advent model speculates that a detail is engendered by a streak of VI Bernoulli experiments, one for exclusive words 'wt' in the glossary V. The fallout of any experiment regulates whether the analogous word will be comprised at least once in the script.

The document '$d_i$' can be drawn in a binary component vector of length IV. Here 't' is the each dimension of the vector, denoted as Bit c {0, 1}, illustrates whether the term 'wt' occurs at least once in 'dz'. The Naive Bayes hypothesis speculates that the V trials are self-reliant of each other. By molding the Naive Bayes speculation, we can figure out the feasibility of a credential given a class from the contingencies of the words given in the class.

### 2.1.3 Multinomial Model

The multinomial proceeding model deduces that a form 'di' of length 'di' is provoked by a concatenation of I d7 term events, where the end product of individual event is a term taken away the vocabulary V. Following McCallum and Nigam (1998), we estimate that the script length circulation P (Id, I) does not confide on the grade. Thus a document 'di' can be characterized as a track of length IV, where respective magnitude t of the angle, mark as Nit > 0, is the general word w t takes place in 'di'. The Naïve Bayes assumption predicates that the 'di' probations are sovereign of the individual. By making the Naïve Bayes assumption, the conceivability of a document inclined a sharp is the most nominal dissemination.

### 2.1.4 K-Nearest Neighbor

A contrast to NBC, which is treated a thirst trainee, k-NN is advised lazy classifier where the allocation work takes place when a new test document is encountered rather than when the training documents set is handled. The radical assumption of a k-NN classifier is adopting a fitting resemblance allowance, if a document *d* is culmination to *k* documents, and then it is confidential as kinship to the class of the preponderance of these documents. k-NN has been auspiciously practiced to many Web classification responsibilities. (Yang et al. 2002).

There are two underlying guidelines that involve the conduct of k-NN classifier particularly, *the similarity measure* and *the parameter k. The similarity measure* is used to regulate how analogous a test document is for a teaching one. A candid choice is the significance of the disparity between the countenance vectors of the two evidences.

$$cosSim(d_1, d_2) = \frac{d_1 . d_2}{|d_1||d_2|} \qquad (2)$$

Mostly k-NN occupying allocates fixed the value of *k*, however, classification will be tendentious concern classes with extra practicing documents exemplifications. Baoli et al. (2003) propose a k-NN algorithm that the usage divergent value of *k* for various classes slightly than a rigid *k* value for all classes. To evade bias blend with a fixed value of *k*, their approach does proportional savor; the original k-NN algorithm is used to first rescue *k* similar documents. Then it enumerates the contingency that a document resides in a class by accepting only few top *n* adjacent acquaintances for that class. The routine conclusion on the use of various numbers of adjoining neighbors for different classes.

### 2.1.5 Support Vector Machines

Support Vector Machines (SVM) where most prospective by Vapnik in 1979, they are based on the Structural Risk Minimization Principle (Vapnik 1995), SVMs attained familiarity only latterly as a dynamic machine training gizmo. A linear SVM is a hyper stratum that aspiration at detachment a turn of absolute learning patterns from a set of pessimistic studying ones with greater viable margin. Distending the edge can be exemplified as an accretion obstacle as exposed in figure 1.

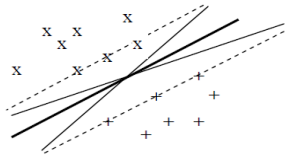

**Figure 1** the set of solid lines represent separating hyper-planes, the bold hyper-plane is the optimal one, and dashed lines represent support vectors.

Regularly, given a set of n linearly detachable points S = {x $\epsilon$Rn|i = 1, 2,…, n}, each point xi resides in one of the two assorts, characterized as $c_i \in \{+1,-l\}$. A slender SVM directs at discovering a hyper plane H that cuts S into two sectors, each side enclosing points with the equivalent division label only.

***High dimensional input space***: When gaining knowledge of text classifiers, one has to transact with large number of countenance. Due to the fact SVM use over fitting protection, which does not routinely confide on the range of fellows, they have got the opportunity to handle those immense factor spaces.

Most textual content categorization trouble is linearly separable:All leagues are linearly detachable and so there are among the Reuters duties. The concept of SVMs is to locate such linear separators.

### 2.1.5 Decision Trees

A Decision Tree (DT) is a tree wherein inner hubs are labeled with terms, divisions isolating from they are set apart by the weightlessness that the referent has in the content patent and leaves are marked by accumulation. Choice Tree corduroy utilizing 'separate and vanquish' methodology. Every hub in a tree is pondered with set of cases. This wheeze checks whether all the penetrating point of reference have a similar shame and on the off chance that not, at that point select a terminal setting from the reservoir classes of archives that have precisely indistinguishable qualities for term and pack each such class in a confined sub tree. Advantages over the current situation with the-workmanship.

Content arrangement is multiclass given.

Preparing is totally overseen content classifiers hypothesizes huge amount of labeled information whose explanation can be important. As an outgrowth there has been included it in utilizing SSL systems for content classification.

### *2.2 WEB CLASSIFICATION APPROACH.*

Diverse perspectives to computerized Web document codification have been advanced with distinct extends of victory, each utilizes one or amalgamation of proficiencies to inscribe the hurdles as validly as viable. We sort these proposals into the consecutive:

#### 2.2.1 Text-Only Approach

This Text only approach (TA) confides on text countenance only which are culled from captivate of the Web document. First, a database of secret sign for each division is figured out from a group of pre-segregated Web documents (training documents), when a current document is to be grouped (test document) its text composition is separated, stop-words are evacuated and tarrying text is conferred as a component vector to the classifier. This class of access verified incapability in Web classification as more Web pages do not encompass plentiful text to classify its class label (e.g. Home pages). Wong and Fu (2000) recognized that about 95% of Web pages accommodate less than 500 distinct words.

Mladenic (1998) advanced an advent based on the Yahoo! Pecking order using an NB classifier where Web documents are pictured as feature-vectors applying the bag-of-words depiction capable of 5 words (1-grams, 2-grams, . . . 5-grams) transpire in a text as an arrangement, feature shearing is done by getting rid of stop-words and applying feature verge Experimental appraisal on real-world data displays that the suggested advent provides great results. For aggrandized than a partly verifying illustration a perfect category is betwixt the 3 grades with the apical anticipated probability.

#### 2.2.2 Hypertext Approach

Furnkranz (1999) advanced a Hypertext Approach (HP) that is built upon the guess is that it is over and over show to convey a hypertext page receiving guidance gave on pages that point to it rather by utilizing guidance that is managed on the page itself. The advanced access epitomizes a site page with a countenance borrowed from instruction of pages that imprint to that page, then it conceal each hyperlink denoting to a page with its comfort text, the description structurally leading it, and the text of the feature in which it appears, and then it reviews a set of allocation guidelines using the inaugural rule learning algorithm RIPPER (Cohen 1995). The anticipation of connections assigning to a similar page is then blended to yield a guess for this page testing one of these democratic plans: voting, restricted voting, weighted sum, and maximum tenacity. It programs reminiscence and rigor percentages using bower text, headings, paragraph and consolidation of them as a basic source of instruction for designation and appropriate different voting strategy for prognosis.

This tier of accessions makes benefit of of both climates just as substance (content) lineaments of Web archives. Setting demeanors are misused to improve the quirk of the guideline; they are removing either from the report itself, for example, HTML labels (eg. title, headings, table, picture, and Meta labels) or from connecting archives (for example anchor text). Esposto et al (1999) and Chakrabarti et al (1998) gave two particular advances that depend on the usage of hypertext highlights referenced previously.

Glover et al. (2002) researched the effort of the content in demonstrating archives close to the representation for grouping. Result show that the content in supposing records, when empty. Regularly has more noteworthy specific and smooth power than the content in the goal report itself. The endeavor aligns the adequacy of utilizing only a website page's full-content, grapple content, and what they call extended stay message (the words and expressions come to pass close to a connect to an objective page). The built upon a reasonable four stage technique: First, obtain a lot of positive and negative managing reports. Second, elicitate every possible element from these records (an element in this situation is a word or expression). Third, perform dimensionality reduction using threshold and entropy-based. Fourth, train an SVM classifier.

### 2.2 HYBRID APPROACHES

#### 2.2.1 Neural Networks Initialized with Decision Trees (NNIDT)

This is a bastard access can be enforced to the obstacle of content distribution and to test its pursuance contingent to a number of auxiliary text distribution algorithms. This entrance goes before the utilization of a miscegenation decision tree and aural system execution to the issue of content dispersion, since half breed gets to choice tree, preparing is utilized to do an emotional examination and visual learning is utilized to do resulting noticeable investigation.

The normal approach for content gathering task builds the supply route by straightforwardly mapping decision hubs or point of reference to the sound-related units and coagulate the system by nullify unimportant and unnecessary units and alliance. This methodology exhibits that compound decision tree and neural system proposition improved veracity in content demeanor task and are similarly better than isolated decision tree or neural system boot aimlessly message classifiers introduction proportionate to point

#### 2.2.2 TF*IDF for text representation

The TF*IDF is developed from the IDF, which is projected by Sparck Jones (1972, 2004) with a heuristic hunch that a concern term which appears in many documents is not a pleasant segregate, and should be provided less mass than one which takes place in few documents.

The fundamental design of TF*IDF belongs to the hypothesis of words displaying that the conditions in a likely document can be segregated from below categories: these words with relatedness and those words without relatedness (Roberston, 2004), i.e., whether or not a term is admissible with the subject matter of an issued document. Further, the relatedness of a term for a accustomed document can be judged by TF and IDF and in TF*IDF formulation, it is adopted to limit the prominence of a term in the document assemblage.

Therefore, there are a few assessments of applying TF*IDF for textual depiction. The first one is that TF*IDF is too 'ad hoc' due to the fact it is not exactly derived from an analytical model, even even though generally it is elucidated by Shannon's statistical concept (Caropreso, Matwin, & Sebastiani, 2001).

The second criticism arrives from that the dimensionality (size of the character set) in TF*IDF for textual testimony is the size of the vocabulary traversing the entire dataset, resulting in that it contributes about a massive computation on weighting all these phrases (Christopher & Hinrich, 2001).

### 2.2.3 LSI for text representation

A basic defect of gift data healing is that dispute the mortal adopts often aren't constant as those words, by that the instruction they appear for the indexed (Berry, Dumais, & O'Brien, 1995). There is Unit completely 2 surfaces to the current matter: synonyms and ambiguity. Totally {different completely different} users in contexts or with different desires, knowledge, or linguistic habits can characterize similar data adding completely different terms. The ambiguity is a fact; most words acquire multiple distinct which means. Thus, they utilize a term during a search question doesn't inevitably mean a document is accommodated by a similar term.

LSI (Latent Semantic Indexing) (Deerwester, Dumais, Landauer, Furnas & Harshman, 1990) could be a far-famed linear pure mathematics classification technique to introduce low dimensional delegations by word co-occurrence. The fundamental idea following LSI is to take dominance of contained higher-order construction in accompany with documents (''semantic structure") in order to boost the disclosure of applicable documents, in accordance with the terms form in queries.

LSI directs to sight the most effective house|topological space} closeness of the first document space within the sense of reducing the world rehabilitation error (the distinction of Frobenius criterion at the intervals the first matrix and its approximation matrix). It's established on SVD (Singular worth Decomposition) and estimates the document vectors into a resembled mathematical space, so affinity will precisely represent linguistics closeness.

## III. METHODOLOGY

In this section, comparative study on some significance of the diverse document institution and categorization of using bays and other relevant models. Table 1 shows recall and accurate percentages categorization was set up using only the pupil (558), staffs (153), courses (245) and project (84) category with different features combinations such as using URL text (U), anchor text (A), headings (H), paragraph (P), Page Text (Pt), Title Text (T) and combinations of them as a basic source of information for classification and utilizing different voting schemes for predictions.

***Data set:*** The categorization task using the ILP98 WebKB dataset. The experiments were on a subset of the WebKB (the ILP 98 data set) (Slattery and Craven, 1998), containing 4,167 pages, in which all documents are related to its anchor, words (text from the hyperlinks that time to the page). The predictions of links inform to identical page are then combined to yield a prediction for this page exploitation one amongst these pick schemes: pick, restricted pick, weighted total, and most confidence.

## IV. RESULT

The following table is concern that the comparative study of the ïve Bayes classifier to find the document organization based on existing methods

**Table. 1: Comparative Study on Various Document organizations**

| Model | Classification | Recall | Precision | | | | |
|---|---|---|---|---|---|---|---|
| | | | Vote | Nor | Wt | Max | Mac.Avg |
| NBA | A | 40.76 | 83.64 | 82.30 | 82.48 | 83.88 | 16.77 |
| | H | 74.10 | 88.05 | 88.19 | 88.95 | 88.95 | ***17.79*** |
| | P | 46.67 | 75.71 | 75.91 | 75.92 | 74.49 | 14.89 |
| | U | 41.70 | 21.30 | 32.90 | 16.90 | 37.00 | 07.40 |
| | T | 39.21 | 12.62 | 54.88 | 16.01 | 67.26 | 13.45 |
| | Pt | 55.12 | 41.64 | 61.33 | 12.22 | 66.55 | 13.31 |
| K-NN | A | 41.16 | 83.12 | 82.48 | 82.30 | 83.64 | 16.72 |
| | H | 64.10 | 88.24 | 88.95 | 88.19 | 88.05 | 17.61 |
| | P | 56.67 | 74.94 | 75.92 | 75.91 | 75.71 | 15.14 |
| | U | 44.76 | 37.10 | 16.92 | 32.91 | 21.34 | 04.26 |
| | T | 41.21 | 67.20 | 16.02 | 54.81 | 12.65 | 02.53 |
| | Pt | 56.14 | 66.52 | 12.22 | 61.31 | 41.66 | 08.33 |
| SVM | A | 30.76 | 73.14 | 79.30 | 82.48 | 73.88 | 14.77 |
| | H | 64.10 | 78.25 | 81.19 | 88.95 | 78.95 | 15.79 |
| | P | 56.67 | 74.31 | 71.91 | 75.92 | 64.49 | 12.89 |
| | U | 51.71 | 21.33 | 32.90 | 16.92 | 37.02 | 07.40 |
| | T | 41.22 | 12.12 | 54.84 | 16.08 | 67.26 | 13.45 |
| | Pt | 57.11 | 41.34 | 61.30 | 12.29 | 66.51 | 13.30 |
| DT | A | 41.21 | 75.71 | 66.52 | 41.16 | 56.67 | 11.33 |
| | H | 44.76 | 21.34 | 73.14 | 64.10 | 44.70 | 08.94 |
| | P | 56.14 | 12.65 | 78.25 | 56.67 | 41.20 | 08.24 |
| | U | 32.90 | 41.66 | 74.31 | 44.76 | 56.10 | 11.22 |
| | T | 54.84 | 73.88 | 21.33 | 41.21 | 40.76 | 08.15 |
| | Pt | 57.11 | 78.95 | 12.12 | 56.14 | 74.10 | 14.82 |
| TA | A | 21.37 | 82.30 | 82.48 | 83.88 | 61.31 | 12.26 |
| | H | 41.16 | 88.19 | 88.95 | 88.95 | 82.48 | 16.49 |
| | P | 64.10 | 75.91 | 75.92 | 74.49 | 88.95 | ***17.79*** |
| | U | 56.67 | 32.91 | 16.92 | 37.08 | 75.92 | 15.18 |
| | T | 44.76 | 54.88 | 16.01 | 67.26 | 16.92 | 03.38 |
| | Pt | 41.26 | 61.33 | 12.22 | 66.51 | 16.08 | 03.21 |
| HP | A | 56.16 | 82.48 | 83.64 | 32.90 | 40.76 | 08.15 |
| | H | 64.10 | **88.95** | 88.05 | 54.88 | 74.10 | 14.82 |
| | P | 56.67 | 75.92 | 75.71 | 61.33 | 46.67 | 09.33 |
| | U | 44.73 | 16.92 | 21.32 | 82.48 | 41.70 | 08.34 |
| | T | 41.23 | 16.02 | 12.61 | 88.95 | 39.21 | 07.84 |
| | Pt | 56.13 | 12.22 | 41.63 | 75.92 | 61.33 | 12.26 |
| NNIDT | A | 75.71 | 78.25 | 74.31 | 44.76 | 73.14 | 14.62 |
| | H | 21.34 | 74.49 | 88.19 | 83.64 | 82.48 | 16.49 |
| | P | 12.65 | 37.08 | 75.91 | 88.05 | 88.95 | ***17.79*** |
| | U | 78.25 | 67.26 | 32.99 | 75.71 | 75.92 | 15.18 |
| | T | 74.31 | 54.88 | 54.88 | 21.33 | 16.92 | 03.38 |
| | Pt | 21.34 | 61.33 | 61.33 | 12.61 | 16.02 | 03.20 |
| TF*IDF | A | 79.30 | 66.56 | 41.63 | 12.22 | 61.33 | 12.26 |
| | H | ***81.19*** | 82.48 | 73.14 | 74.49 | 74.49 | 14.89 |
| | P | 71.91 | ***88.95*** | 78.25 | 37.00 | 37.08 | 07.41 |
| | U | 21.30 | 75.92 | ***88.96*** | 67.26 | 67.26 | 13.45 |
| | T | 12.62 | 61.33 | 67.20 | ***89.19*** | 75.71 | 15.14 |
| | Pt | 41.64 | 16.96 | 66.52 | 71.91 | ***89.01*** | 13.45 |
| LSI | A | 40.76 | 12.22 | 73.14 | 32.90 | 66.51 | 13.30 |
| | H | 74.10 | 88.05 | 78.25 | 66.52 | 75.91 | 15.18 |
| | P | 46.67 | 75.71 | 75.91 | 73.14 | 61.33 | 12.26 |
| | U | 41.70 | 21.31 | 32.93 | 78.25 | 82.48 | 16.49 |
| | T | 39.21 | 54.88 | 54.88 | 74.31 | 88.95 | ***17.79*** |
| | Pt | 55.12 | 61.33 | 61.33 | 75.71 | 75.92 | 15.18 |

It is observed from the table 1, the performance of the document institution and categorization of using bayes and other relevant related models. The work of classifying a latest document depends on the phrase sets generated from preparation files. So the range of training documents is critical in pattern of word sets used to decide the class of a new file.

The greater wide variety of word sets of training files reduces the opportunity of breakdown to categorize a new report.

In this approach implemented as discussed. The experiments are finished with the dataset referred to above, along with the Naive Bayes classifier learning algorithm. For performance evaluation, the Accuracy, Precision and Recall metrics are used that were presented in the preceding sections.

It is noted that document categorization of using bayes and other relevant models procedure gives the approximately equal. But some advanced method brings some relax compared to other methods. Especially all points accumulating in TF*IDF method compared to other existing methods. Its observed the above data set the TF*IDF Naive Bayes document organization is performed better compared to other related existing methods.

## V. DISCUSSION & CONCLUSIONS

The document organization using classification, Web and hybrid document approaches are analyzed in this paper. As initially discussed the importance of feature selection and extraction to classification quality, and then presented the issue of similarity measures. Further discussed the common classification techniques proved promising in the field such as Naïve Bayesian, k-NN, SVM and etc., Although Naïve Bayesian and k-NN is one of the most accurate classifiers available currently, and it is familiar due to their straightforwardness and computational competence

The web classification approaches, namely, *Text-Only* approach and *Hypertext* approach are also analyzed. Finally, compare to the other document organization and classification naïve and naïve related method give better performance to other methods.

Research plan has been introduced to reach a generalized framework for document organization and classification. Taking into account the all categories mentioned beforehand, the framework will utilize different contextual features while considering domain knowledge to decide which features or combination of features are more suitable.

## REFERENCES

1. Irina Rish, (2001), An Empirical Study of the Naïve Bayes Classifier, Proc. of the IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence.
2. Baoli, L., and et al. (2003), An Improved k-Nearest Neighbor Algorithm for Text Categorization, CPOL.
3. Yang, Y., and et al.(2002), "A study of approaches to hypertext categorization", Journal of Intelligent Information Systems, pp. 219-241.
4. Songbo, T., and et al. (2005), A novel refinement approach for text categorization, Proc. of 14th ACM International Conference on Information and Knowledge Management, pp.469-476.
5. Taeho Jo (2010), NTC (Neural Text Categorizer): Neural Network for Text Categorization, International Journal of Information Studies, vol. 2, no.2.
6. J. Kaur and S.Bhagal (2016), New Classification using Naïve Baye's Classifier, International Journal of Advanced Research in Computer Science and Software Engineering,vol. 6, no.4, pp. 698-702.
7. T. Santoso and Setiono R., (2002) "A comparative study of centroid – based, neighborhood-based and statistical approaches for effective document categorization", ICPR '02 Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02), vol.4, no. 4, pp.235–238.

## AUTHORS PROFILE

**R.Sathish Babu** is working as Assistant professor, Department of computer and Information Science, Annamalai University. He has 18 years of experience in teaching and involved in research activities. His area of specialization includes Data Mining and Document Classification.

**Dr.R.Nagarajan,** is working as AssistantProfessor, Department of Computer and Information Science, Annamalai University. He has published 12 research papers in International Journals. He has 21 years of experience in Application Programming. He is involved in research activities for the past 12 years. His area of specialization includes Data Mining, Document Clustering., Cloud Computing and NLP.