International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 4 | ISSN : 2456-3307

Survey on Classification Approach for Text Categorization

Rupali Patil, V. M. Barkade
Department of Computer Engineering Rajashri Shahu College of Engineering, Savitribai Phule Pune University, Pune, India

ABSTRACT

In the area of information retrieval, text categorization has recently become an active research topic. The goal of text categorization is to allot entrics from a set of prespecified categories to a document. Learning in a very high dimensional data space is a key challenge in a text categorization approach. Learning from such high dimensional features may prompt a high computational burden and may even hurt the classification performance of classifiers because of irrelevant and, redundant features. To improve the ”curse of dimensionality” issue and to speed up the learning procedure of classifiers, it is important to perform feature reduction to reduce the size of features. This paper introduces a Bayesian arrangement approach and J48 classifier for automatic text categorization utilizing class-specific features. For text categorization, has the proposed strategy chosen a specific feature subset for every class. The detectable significance of this methodology is that most feature selection criteria, for example, Information Gain (IG) and Maximum Discrimination (MD), can be effectively joined into this methodology. The J48 classifier saves the time and memory. The proposed system also uses Term weighting concept for preprocessing. These methods increase the accuracy of classification and feature selection process and improve the system performance.

Keywords : Text categorization, class-specific features, Feature selection, PDF projection and estimation, dimension reduction, J48, Term weighting.

I. INTRODUCTION

As long as the size of data on the internet and organizations will get maximized there is a need of a technique for managing the large size of data which will filter as well as manage these data categorization. The most important part is to categories the free text document in to one or more categories which are predefined, sorting of emails or files into a folder hierarchies, identification of topic to support topic,particular processing operations, structured searchand/or browsing, or finding documents that match longterm standing interests or moredynamic taskbased interests.

In various contexts professionals are appointed to categories the new items, but this process is very much time consuming as well as will as costly so bounding its applicability apparently there is an more interest in the research and development of the methods for text categorization automatically.

There are several classification as well as machine learning methods has been implemented for categorization of text such asnearest neighbor classifiers, decision trees, Support Vector Machines, rule learning algorithms and so on.

In this survey paper we will list some of the work done by the researchers on the text categorization.

II. RELATED WORK

In paper [1] authors have developed a system which is automatically categorize text by making use of class specific features which is Bayesian classification. The proposed method allows selecting the vital features for every class. Authors have designed a Naive Bayes rule by using Baggenstoss PDF Projection Theorem for applying the class specific features for classification. The major advantage of derived technique is it can make use of present feature selection conditions.

In paper [2], authors have developed a system for finding optimal classification by making use of class-specific features known as the hypothetical establishment also provided examples of its utilization. A new probability density function (PDF) projection hypothesis makes it possible for project probability density functions from a low-dimensional feature space back towards the raw information space. An M-ary classifier is produced by assessing the PDFs of class-specific features, after that the transformation of every PDF back to the raw information space in which they can be fairly studied. Albeit statistical sufficiency is not important, the classifier in such a way created will get to be equal to the optimal Bayes classifier if the features meet adequacy prerequisites exclusively for each class.

In paper [3] authors developed an automatic text categorization method as well as research its application to text retrieval. The categorization technique designed using a combination of 01 a learning paradigm called as instance-based learning as well as an advanced document retrieval technique i.e. retrieval feedback. They showed the ability of developed method using two real world document collections from the MEDLINE database. Next, they investigate the use of programmed automatic categorization to text retrieval.

In paper [4] presents ideas and algorithms of feature selection, surveys existing feature selection algorithms for classification and clustering, groups also contrasts distinctive algorithms and an arranging structure in view of pursuit methodologies, evaluation conditions, and information mining task, uncovers unattempted combinations, and gives rules in selecting highlight choice algorithms. With the categorizing structure, they precede with our efforts toward-creating an incorporated framework for intelligent feature selection.

In paper [5] authors developed a unique idea of segmenting the documents for calculating term weights. Tests with DynaPart-FiLa(Dynamic partitioning of text documents with first and last partitions) suggest that segmenting the documents before estimating term weights helps in improving weighted average F-measure (based on macro-weighted averaging). For all the datasets, F-measure has enhanced for all classifiers with DynaPart-FiLa. From the improved F-measure, authors say that those relevant terms appear from the beginning of the document. Increasing their importance helps in improving the classification outcomes. Finally they predicted that positional significance of terms are good signal of context of the document.

In paper [6] author has developed a system which is automatically categorize text of Marathi files depending on the profile of the user that has browsing history of the user. Vector Space Model provides good outcomes compared with the present Probabilistic Models. The precision of the outcomes related to the system is way good compared with the Tamil language. LINGO algorithm provides better cluster quality compared with different clustering methods.

In paper [7] author searched as well as observed that, a discussion in the various developers related the pros of utilizing the stemming in categorization of Arabic text. Due to this, authors have conducted the analysis of feature reduction technique for clearing the effect of this famous technique in the mining of text as well as classification of files. They also few Arabic text condition to refuse the use of stemming in Arabic text categorization.

Table 1. Survey Table

Sr. No.	Paper Name	Authors Name	Technique Used	Advantage
1	A Bayesian Classification Approach Using Class-Specific Features for Text Categorization	Bo Tang, Haibo He, Paul M. Baggenstoss , and Steven Kay	Bayesian Classification Approach	Many existing feature selection criteria can be easily incorporated
2	The pdf projection theorem and the class-specific method	Paul M. Baggenstoss	pdf projection theorem	Possible to automate the feature and model selection process
3	Automatic text categorization and its application to text retrieval	W. Lam, M. Ruiz, and P. Srinivasan,	Automatic Text Categorization	improvements in the performance
4	Toward integrating feature selection algorithms for classification and clustering	H. Liu and L. Yu	Feature selection algorithms	Feature selection increases classification accuracy
5	Term weighting using contextual information for categorization of unstructured text documents	A. Kulkarni , V. Tokekar and P Kulkarni ,	DynaPart-Fila (Dynamic partioning of text documents with first and last partitions)	Has good classification accuracy
6	Automatic text categorization: Marathi documents	J. J. Patil and N. Bogiri	Lingo Algoritm for Marathi Categorization	LINGO algorithm provides better cluster quality
7	Stemming impact on Arabic text categorization performance	F. S. Al-Anzi and D. AbuZeina,	Arabic text categorization using stemming	Stemming performs better in some cases

III. PROPOSED SYSTEM

This paper proposes a Bayesian arrangement approach for automatic text categorization utilizing class-specific features. Not at all like the conventional methodologies for text categorization, has the proposed strategy chosen a specific feature subset for every class. To apply these class-dependent features for classification, we follow Baggenstoss' PDF Projection Theorem to reconstruct PDFs in raw data space from the class-specific PDFs in low-dimensional feature space, and assemble a Bayes classification rule. The detectable significance of this methodology is that most feature selection criteria, for example, Information Gain (IG) and Maximum Discrimination (MD), can be effectively joined into our methodology. System assesses this technique's classification performance on several real-world benchmark data sets, contrasted with the state-of-the-art feature selection approaches. In contribution we use Weighted J48 classifier for classification and Chi square method for feature selection. These methods increase the accuracy of classification and feature selection process and improve the system performance.

System assesses this technique's classification performance onseveral real-world benchmark data sets, contrasted with thestate-of-the-art feature selection approaches. To remove the unnecessary data, we use the term frequency concept with TF-IDF. According to this method, term frequency is calculatedfor each word to generate the training file. This training fileis provided to Feature selection process. In contribution weuse J48 classifier for classification. These methods increasethe accuracy of classification and feature selection processand improve the system performance. Also system used Termweighting concept for categorization of unstructured text documents.During the categorization of text documents, termweighting assigns appropriate weights to different terms. It helps in to improve the categorization result.

Module Description

1) Input and Read Dataset

In this module user provide the 20-newsgroup dataset with different topics and read the dataset.

2) Preprocessing approach

Certain preprocessing tasks are usually performed before the dataset is used for retrieval .In the dataset number of words present means it content numerous features which can be hurt classification performance are removed using stemming and Stopward operations.

Stemming : Stemming refers to the process of reducing words to their stems or roots.

Stopwards : They are frequently occuring and insignificant words in a language that help to construct sentences but do not represent any content of the documents.

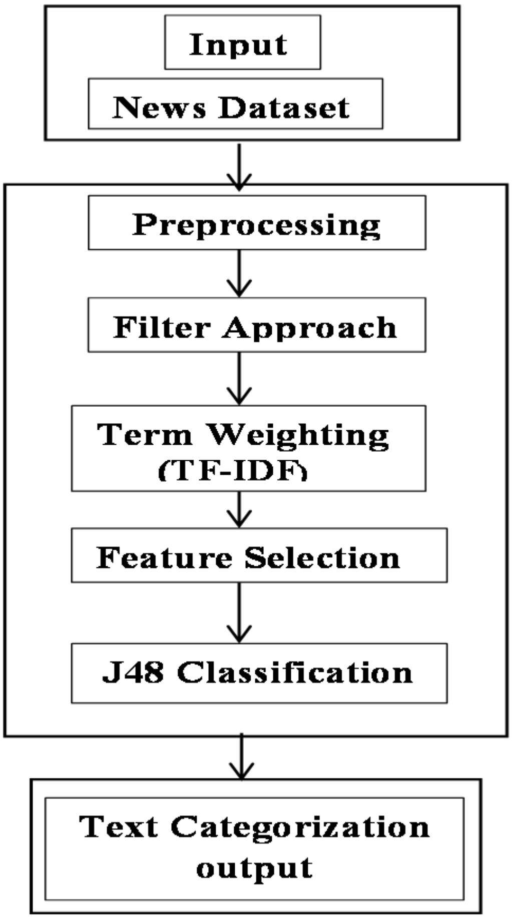

Figure1 : System Architecture

3)Filter Approach

In this module calculate the weight or energy of an each words from dictionary, which get from Stemming and stopword removal(Preprocess Module). After that create Arff File for each Words.

4) **Term Weighted Concept**

Term weighting determines the importance of a term for a document. Term weighting assigns appropriate weights to different terms. According to this method, term frequency is calculatedfor each word to generate the training file. During thecategorization of text documents, term weighting assignsappropriate weights to different terms.

5) Feature Selection

In feature selection 'subset of original features' are selected. Feature selection is important in text categorization to reduce the size of features. Feature selection also speed up learning process of classifiers. This module takes the important words that are features as an input and select the features by applying feature selection method.

6) Classification

After selecting the features using feature selection these features are used for classification. For classification this system used J48 classifier which have high accuracy and get the text categorization output.

IV. CONCLUSION

Text categorization is very important process in various applications of information retrieval, machine learning and text mining. This paper presents the recent text categorization and feature selection techniques used for text document classification. Also make comparative analysis of all these approaches, according to technique used and their respective advantages and limitations. Some Feature selection methods are also discussed; use to reduce the dimensionality problem in text categorization process.

V. REFERENCES

[1]. Bo Tang, Haibo He, Paul M. Baggenstoss, and Steven Kay, "A Bayesian Classification Approach Using Class-Specific Features for Text Categorization", 1041-4347 (c) 2015 IEEE, Transactions on Knowledge and Data

[2]. Paul M. Baggenstoss, "The pdf projection theorem and the class-specific method", IEEE Transactions on Signal Processing, vol. 51, no. 3, pp.672-685, 2003.

[3]. W. Lam, M. Ruiz, and P. Srinivasan, "Automatic text categorization and its application to text retrieval", IEEE Transactions on Knowledge and Data Engineering, vol. 11, no. 6, pp. 865-879, 1999.

[4]. H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering", IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491-502, 2005.

[5]. A. Kulkarni, V. Tokekar and P. Kulkarni, "Term weighting using contextual information for categorization of unstructured text documents," 2015 Annual IEEE India Conference (INDICON), New Delhi, 2015, pp. 1-4.

[6]. J. J. Patil and N. Bogiri, "Automatic text categorization: Marathi documents", 2015 International Conference on Energy Systems and Applications,Pune, 2015, pp. 689-694.

[7]. F. S. Al-Anzi and D. AbuZeina, "Stemming impact on Arabic text categorization performance: A survey", 2015 5th International Conference on Information Communication Technology and Accessibility (ICTA),Marrakech, 2015, pp. 1-7.