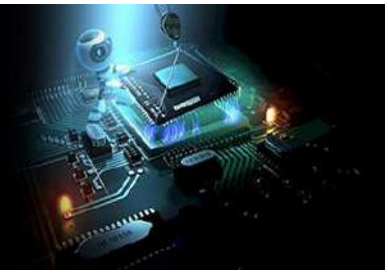
International Journal of Engineering in Computer Science



E-ISSN: 2663-3590
P-ISSN: 2663-3582
IJECS 2020; 2(2): 19-21
Received: 08-04-2020
Accepted: 10-05-2020

Rajendra Prasad Kudumula
GATE College, Tirupati, Andhra Pradesh, India

Correspondence
Rajendra Prasad Kudumula
GATE College, Tirupati, Andhra Pradesh, India

Efficient email classification strategy based on semantic methods

Rajendra Prasad Kudumula

DOI: https://doi.org/10.33545/26633582.2020.v2.i2a.35

Abstract
Emails have emerged as one of the foremost packages in each day life. The continuous increase in the wide variety of email users has led to a huge boom of unsolicited emails, which might be also known as junk mail emails. Managing and classifying this large variety of emails is an important challenge. In this paper, a green email filtering approach based totally on semantic techniques is addressed. The proposed technique employs the Word Net ontology and applies exceptional semantic-based totally strategies and similarity measures for lowering the huge number of extracted textual features, and as a result, the gap and time complexities are reduced. Most of the approaches delivered to remedy this trouble treated the high dimensionality of emails by the use of syntactic feature selection. Moreover, to get the minimal most appropriate features' set, function dimensionality reduction has been integrated using characteristic selection strategies which include the Principal Component Analysis (PCA) and the Correlation Feature Selection (CFS). Experimental results on the usual benchmark Enron Dataset showed that the proposed semantic filtering approach combined with the function choice achieves excessive computational performance at high area and time discount rates. A comparative study for numerous classification algorithms indicated that the Logistic Regression achieves the very best accuracy in comparison to Naïve Bayes, Support Vector Machine, J48, Random Forest, and radial basis function networks. By integrating the CFS characteristic choice technique, the average recorded accuracy for the all used algorithms is above 90%, with more than 90 reductions. Besides, the carried-out experiments showed that the proposed paintings have a highly sizeable overall performance with better accuracy and much less time in comparison to other related works.

Keywords: Word net, dimensionality, naïve Bayes, random forest

Introduction
Electronic mails (Email) have become one of the most important and powerful communication ways in personal lives and business. Some users misuse the Emails by sending computer worms and spams which are unrequested information sent to the Email inboxes [1, 3]. The average spam Email messages sent every day have reached 54 billion messages based on statistics in 2014. Spam Emails cause an overload to the email servers, and consume network bandwidth and storage capacity. Therefore, Email filtering is a very important process to solve these problems. The filtering purpose is to identify and isolate the spam Emails. Many mail server engines are using various authentication mechanisms to analyze Email content and categorize the Emails into white and black lists so; it can be optimized by users [2]. Using white and black lists, the new Email source is compared with a database to know if it is classified as spam before or not. On another side, an alternative approach filters Emails by extracting features from the Email body and using some classification methods, such as Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Neural Networks (NN). Most of the related works classify emails using the term occurrence in the email. Some few works additionally consider the semantic properties of the email text. Integrating semantic concepts and approaches for email classification is expected to add important benefits of enhancing the computational performance, in addition to the accuracy of classification.

Related Work
Supervised class strategies were applied extensively for Email filtering. The pre-described category labels (Ham or Spam) are assigned to documents based at the probability recommended by way of a schooling set of labeled documents. Some of the techniques used in particular for filtering spam Email are: Naive Bayes (NB) [3], Artificial Neural Net-works (ANN), k-Nearest Neighbor (KNN), Logistic Regression, C4.five classifier, RBF Networks,

Multi-Layer Perceptron (MLP), Ada Boost, Support Vector Machine (SVM), and Random Forest (RF). In Sharma et al. proposed a way based on ANN the use of Radial-Basis Function networks (RBF). In, an anti-spam filter named SENTINEL is applied. The filter extracts the natural language attributes from Email text which might be related to author stylometry. RF, SVM, and NB, and meta-algorithms referred to as ADABOOSTM1 and bootstrap aggregating (BAGGING) are used and evaluated the use of CSDMC2010 dataset. Comparative evaluation for special type algorithms has been also performed in several works. In a junk mail type approach is supplied to extract capabilities (terms) from Email body and additional readability functions relative to the Email (e.G. word length and document length). The experiments were carried out on 4 datasets; Spam Assassin, Enron-Spam, Ling Spam, and CSDMC2010 [4, 5]. Classification has been implemented the use of NB, RF, SVM, Bagging and Ada Boost. They claimed that that the classifiers generated the use of meta-mastering algorithms perform higher than trees, functions, and probabilistic methods. In another comparative evaluation the usage of 4 classifiers (NB, Logistic Regression, Neural Network, and RF) has been offered on Enron dataset, with tremendous performance been recorded for the RF classifier. Moreover, as compared 4 classifiers; BayesNet, J48, SVM and LazyIBK. Their end result showed that the BayesNet and J48 classifiers perform higher than SVM. Some paintings also considered ensemble classifiers, such as in where junk mail filtering is completed the usage of a couple of classifiers. The previously stated related works for Email classification do not think about the problem of high dimensionality of and the related complexity of filtering method. Subsequently, this hassle attracted some researchers [6], [7] to address it in some work. In a content based junk mail filtering method has been provided for classifying the unsolicited mail and ham Emails, with function choice techniques been implemented, specifically PCA (Principal Component Analysis) and CFS (Correlation Feature Selection). The presented technique has been tested on Enron corpus. The results display that the use of CFS saves the time for classifiers than PCA and SVM has the nice prediction accuracy. In an unsolicited mail detection approach the usage of RF classifier has been used on the Spambase Dataset, which permits feature choice and parameter optimization. In feature selection algorithms based totally on similarity coefficients is implemented on Spambase dataset to beautify the detection charge and improve the class accuracy. In improved mutual information model is proposed mixed with the phrase frequencies to calculate the correlation between Email functions and their classes. The experiments have been conducted on English corpus named PU10s and Chinese corpus E-mail dataset.

Proposed Method

The proposed method consists of several additives for reducing the feature dimensionality to filter the E-mails into two classes: Ham and Spam. The proposed architecture is confirmed. It has phases which are education and checking out. The schooling phase consists of 4 important modules: Pre-processing; Feature Weighting; Feature Reduction and Classification. The trying out phase consists of Pre-processing, Feature Weighting and Classification. The most considerable part in the structure is a reduction module. This module consists of 3 proposed processes: Semantic discovery, Weight Function, and Feature Selection. The reduction module is an enhancement to a previous work brought both the pre-processing and the characteristic weighting modules will describes the function reduction module in more detail.

Pre-processing

In the pre-processing module, Tokens are extracted and the beside the point tokens which include numbers and logos are removed. Tokens are extracted from each the frame and difficulty line of the Email. After that, the stop phrases are eliminated. For extra data cleaning, by way of consulting the WordNet as an English dictionary, only the significant phrases are taken into consideration. WordNet is interpreted as massive lexical database for English Language, which businesses English words into units of synonyms known as synsets. In this module, stemming isn't applied. This is applicable in order to keep the meaning of the features. Instead of stemming, morphology is applied the use of WordNet. In morphology, all the features with the equal root are considered as a token using.

Once an email document is pre-processed, the Email document‘d’ can be represented by:

$$d_i = \{(t_1, w_1), (t_2, w_2), \ldots., (t_n, w_n)\}$$

Where, each term‘t’ is considered as a feature which has a corresponding weight ‘w’ in a given d.

Feature weighting

This module calculates the weight of the extracted feature, where each term ‘t’ is weighted by a weight ‘w’ using term frequency/ inverse document frequency method (TF-IDF). The Term frequency calculates the number of times the term‘t’ appears in the Email document ‘d’ as shown in the Eq. (1).

$$if(t, d) = (f_d(t))/(\max[f_d(t)])$$

Where fd(t) is frequency of term ‘t’ in Email ‘d’. The Inverse Document Frequency (IDF) estimates the importance of a given term is. It measures how rare a given term in the whole document, using equation (2):

$$IDF(t) = \log(N/df_t)$$

Where dft is the no. of Emails with term‘t’, and ‘N’ is the total number of Emails. Finally, the TF-IDF is computed as the result of multiplication of Eqs. (1) and (2):

$$W = tf(t, d)IDF(t)$$

Results and Discussions

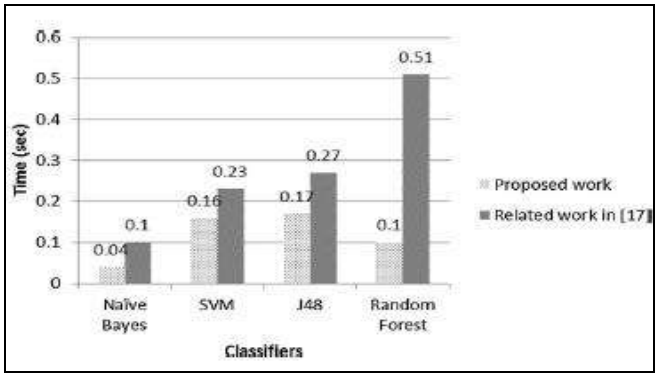


Fig 1: Comparing the time and classifiers

Comparing the time after using CFS feature selection technique against related work between the classifiers Naïve Bayes, SVM, J48, Random Forest.

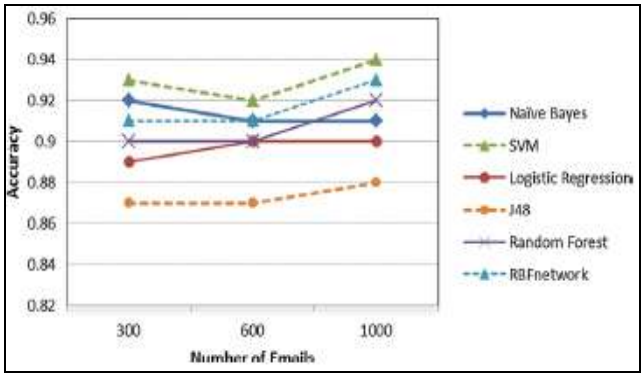


Fig 2: Accuracy vs. No of Emails

Accuracy performance of different classifiers using different sizes of datasets.

Conclusion

In this work, an approach for E-mail filtering is introduced, targeting both accuracy and complexity performance enhancements. It is based on introducing semantic modeling to solve the problem of high dimensionality of features by considering the semantic properties of words. The semantic modeling makes use of semantic relations and semantic similarity measures to compress features in their dimensional space. Feature selection reduction techniques have been moreover for further reduction to achieve optimal feature compression. A set of different classifiers have been studied to test their performances to segregate Emails as spam or ham experiments on the Enron dataset. It has been shown from experiments that the path similarity measure performs the best. Introducing CFS as a technique for feature selection enhanced the accuracy of some classifiers compared to employing the semantic similarity only. Classifiers like Random Forest and RBF Network managed to reach accuracy values 92% and 93% respectively.

References

1. Internet Threats Trend Report, 2014.
2. An interactive hybrid system for identifying and filtering unsolicited e-mail. Springer, Berlin Heidelberg, 2006.
3. In: The Naïve Bayes model for unsupervised word sense disambiguation, 2013.
4. Khan Aurangzeb *et al*. A review of machine learning algorithms for text documents classification.
5. Islam MS. Machine learning approaches for modeling spammer behavior. Information Technology. Heidelberg: Springer Berlin, 2010
6. Blanzieri E, Bryl A. A survey of learning-based techniques of email spam filtering. Information Engineering and Computer Science Department, 2008.
7. http://www.cs.cm.edu/~tom/NewChapters. Html, 2005.
8. Surya PL. Spam classification based on supervised learning. In: International conference on process automation, control and computing (PACC). IEEE, 2011.
9. Shi L, Wang Q. Spam email classification using decision tree. J computer Inform Syst. 2012; 8(3):949-56.
10. Islam M, Zhou W. Architecture of adaptive spam filtering based on machine learning algorithms. Berlin Heidelberg: Springer, 2007.