International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

AUTO E-MAILS CLASSIFICATION USING BAYESIAN FILTER

G. Bhagyashri, Department of Technology, Kolhapur. H. Pratap, D.Y.Patil College of Engineering & Technology, Kolhapur

Abstract

Now-a-days, email becomes a powerful tool for communication as it saves a lot of time and cost. Like every powerful medium, however, it is prone to misuse. One such case of misuse is the blind posting of unsolicited e-mail messages, also known as spam, to very large numbers of recipients. Spam can be defined as unsolicited (unwanted, junk) email for a recipient or any email that the user does not want to have in his inbox. These junk mail not only wastes user time, but can also quickly fill-up file server storage space, especially at large sites with thousands of users who may all be getting duplicate copies of the same junk mail. As a result of this growing problem, automated methods for filtering such junk from legitimate E-mail are becoming necessary. This paper described spam filter implemented is used to block spam. It uses Bayesian filtering to block the spam. Classification using Bayesian filter is done according to the method defined by Paul Graham. The general idea is that some words occur more frequently in known spam, and other words occur more frequently in legitimate messages. Using well-known mathematics, it is possible to generate a "spam-indicative probability" for each word.

Keywords: Spam, Feature selection, Bayesian filtering, Data pre-processing, Feature weighting

1. Introduction

Electronic mail is an efficient and increasingly popular communication medium. Like every powerful medium, however, it is prone to misuse. One such case of misuse is the blind posting of unsolicited e-mail messages, also known as spam, to very large numbers of recipients. Spam is an unfortunate problem on the internet. Spam emails are the emails that we get without our consent. They are typically sent to millions of users at the same time. Spam can be defined as unsolicited (unwanted, junk) email for a recipient or any email that the user does not want to have in his inbox. It is also defined as "Internet Spam is one or more unsolicited messages, sent or posted as a part of larger collection of messages, all having substantially identical content." E-mail spam has steadily grown since the early 1990s. Botnets, networks of virus-infected computers, are used to send about 80% of spam. [2]

Spammers collect e-mail addresses from chat rooms, websites, customer lists, newsgroups, and viruses which harvest users' address books, and are sold to other spammers. Since the cost of the spam is borne mostly by the recipient, many individual and business people send bulk messages in the form of spam. The voluminous of spam emails a strain the Information Technology based organizations and creates billions of dollars lose in terms of productivity. In recent years, spam emails lands up into a serious security threat, and act as a prime medium for phishing of sensitive information Addition to this, it also spread malicious software to various user. An average user on the internet gets about 10-50 spam emails a day and about 13 billion pieces of unsolicited commercial e-mail are sent each day, which represents about half of all e-mail sent.[15]

It was reported an American received 2200 pieces spam e-mail on average in 2002. Increasing by 2% per month, it will reach 3600 pieces spam e-mail in 2007. A survey by CNNIC found that every email user in China received 13.7 piece emails per week in 2004, including 7.9 piece spam emails. In America, spam emails cost enterprises up to 9 billions per year. [17] A study reported that spam messages constituted approximately 60% of the incoming messages to a corporate network. Without appropriate counter-measures, the situation will become worse and spam email will eventually undermine the usability of email Anti-spam legal measures are gradually being adopted in many countries. In China, some experts advocated that an effective anti-spam e-mail measure should be carried out as early as possible. In 2003, AOL, Microsoft, EarthLink and Yahoo sued hundreds of marketing companies and individuals for sending deceptive spam using a new federal law called the CAN-SPAM Act, which prohibits such activities. But these legal measures have had a very limited effect so far due to Internet's open architecture. Hence, apart from legal measures, we should make use of some effective anti-spam e-mail technological approaches too. At present, most anti-spam e-mail approaches, which are too simple to stop spam e-mail efficiently, block spam messages by blacklist of frequent spammers. [16]

With the proliferation of direct marketers on the Internet and the increased availability of enormous Email address mailing lists, the volume of junk mail (often referred to colloquially as spam") has grown tremendously in the past few years. As a result, many readers of E-mail must now spend a non-trivial portion of their time on-line wading through such unwanted messages. Moreover, since some of these messages can contain offensive material (such as graphic pornography), there is often a higher cost to users of actually viewing this mail than simply the time to sort out the junk. Lastly, junk mail not only wastes user time, but can also quickly fill-up file server storage space, especially at large sites with thousands of users who may all be getting duplicate copies of the same junk mail. As a result of this growing problem, automated methods for filtering such junk from legitimate E-mail are becoming necessary. [3] Automatic email spam classification contains more challenges because of unstructured information, more

number of features and large number of documents. As the usage increases all of these features may adversely affect performance in terms of quality and speed. Many recent algorithms use only relevant features for classification.

This paper described spam filter which is implemented using Bayesian filter approach. Classification using Bayesian filter is done according to the method defined by Paul Graham. First of all the program has to be trained using set of spam and non-spam mails. These put in a database (i.e. in spam & ham lists). The Performance increased with the number of training it gets. When new mail comes it is tokenized and probability of each word is found by looking into database. The total probability is found out and if it is greater than 0.5 it is marked as spam.

Outline of this paper:

Section 2 presents related works on email spam classification, Section 3 presents framework of the proposed system, Section 4 presents Implementation of Bayesian filter, Section 5 gives result & analysis. Finally Section 6 presents conclusion and future work.

2. Related Work

In addressing the growing problem of junk E-mail on the Internet, Mehran Sahami & Susan Dumaisy examine methods for the automated construction of filters to eliminate such unwanted messages from a user's mail stream. By casting this problem in a decision theoretic framework, they are able to make use of probabilistic learning methods in conjunction with a notion of differential misclassification cost to produce filters. In order to build probabilistic classifiers to detect junk E-mail, they employ the formalism of Bayesian networks. There experiments also show the need for methods aimed at controlling the variance in parameter estimates for text categorization problems. [3] Vikas P. Deshpande, Robert F. Erbacher, proposed An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques in which efficient anti-spam filter that would block all spam, without blocking any legitimate messages is a growing need. To address this problem, they examine the effectiveness of statistically-based approaches Naïve Bayesian anti-spam filters, as it is content-based and self-learning (adaptive) in nature. Additionally, they designed a derivative filter based on relative numbers of tokens. They train the filters using a large corpus of legitimate messages and spam and also test the filter using new incoming personal messages. [5] Ahmed Obied proposed Bayesian Spam Filtering in which he describes a machine learning approach based on Bayesian analysis to filter spam. The filter learns how spam and non spam messages look like, and is capable of making a binary classification decision (spam or non-spam) whenever a new email message is presented to it. The evaluation of the filter showed its ability to make decisions with high accuracy. [4] Raju Shrestha and Yaping Lin present the new approach to statistical Bayesian filter based on co-weighted multi area information. This new algorithm co-relates the area wise token probability estimations using weight coefficients, which are computed according to the number of occurrences of the token in those areas. Experimental results showed significant improvement in the performance of spam filtering than using individual area-wise as well as using separate estimations for all areas. Future developments may include integrating their approach with phrase-based and/or other lexical analyzers and with rich feature extraction methods which can be expected to achieve even better performance. [11] Denil Vira, Pradeep Raja & Shidharth Gada present An Approach to Email Classification Using Bayesian Theorem. They propose an algorithm for email classification based on Bayesian theorem. The purpose is to automatically classify mails into predefined categories. The algorithm assigns an incoming mail to its appropriate category by checking its textual contents. The experimental results depict that the proposed algorithm is reasonable and effective method for email classification. [12]

Michal Prilepok1, Jan Plato proposed Bayesian Spam Filtering with NCD in which a novel variant of Classic Bayesian filter with combination of Normaliced Compressed Distance was described. This combined filter was tested as filter for spam identification. In addition to Classical implementation of Bayesian filter, two versions of combination with NCD were implemented. The first version uses NCD for all emails which have spamcity higher than 0.5. The second version uses NCD only, when the spamcity was in the interval from 0.5 to 0.75. The second version is much faster than the first version and its speed is almost the same as speed of Classical Bayesian filter. Both new developed versions have worse efficiency in successful marking of non spam emails. The overall efficiency of both new algorithms was better than the original filter. [13]Georgios Paliouras, Constantine D. Spyropoulos, Panagiotis Stamatopoulos, Georgios Sakkis & Vangelis Karkalets are present Learning to Filter Spam E-Mail A Comparison of a Naïve Bayesian and a Memory-Based Approach in which they investigate the performance of two machine learning algorithms in the context of anti-spam Filtering. They investigate thoroughly the performance of the Naive Bayesian filter on a publicly available corpus, contributing towards standard benchmarks. At the same time, we compare the performance of the Naive Bayesian filter to an alternative memory based learning approach, after introducing suitable cost-sensitive evaluation measures. Both methods achieve very accurate spam filtering, outperforming clearly the keyword-based filter of a widely used e-mail reader. [14] Zhan Chuan, LU Xian-liang proposed An Improved Bayesian with Application to Anti-Spam Email in which they presents a new improved Bayesian-based anti-spam e-mail filter. They adopt a way of attribute selection based on word entropy, use vector weights which are represented by word frequency, and deduce its corresponding formula. It is proved that their filter improves total performances apparently. [16] R. Malathi proposed Email Spam Filter using Supervised Learning with Bayesian Neural Network in which he describes a new Spam detection method using Text Categorization, which uses Rule based heuristic approach and statistical analysis tests to identify “Spam”. The initial goal of this paper needed to be determined if the detection of spam precursors could be used to create a system running in real-time that could identify the imminent arrival of spam and block it at the network gateway. Analysis of

gateway audit log files identified potential spam precursor activity, but unfortunately this activity was found in such low proportions that it was deemed unsuitable for use in a real-time spam prevention system. With network precursor detection deemed unsuitable, it was then proposed that spam messages themselves could be used as precursors, allowing the system to identify the current IP addresses of spammers and block them from accessing the network. In order to provide protection against legitimate emails being blocked by the system, a system of IP suspicion was implemented, with IP addresses being classified as malicious or benign based on the collection of supporting evidence. Analysis of the timeliness of the system led to the conclusion that it executed with sufficient efficiency to make real-time operation viable. The speed of the system is directly linked to the amount of time that is looked back in the audit logs when searching for precursor spam activity. This amount of time also influences the effectiveness of the system, as accuracy is lowered when the log look back amount is too small. It was determined that a look back amount of three days provided a good balance between the system timeliness and effectiveness. [18]

3.Framework of the Proposed System

The overall design of the proposed system is given in Fig1. Collections of emails are dataset required for training & testing purposes retrieved from following website:

http://spamassassin.apache.org/publiccorpus

All these emails must be converted into text file format first before being used in processing stage.Proposed System consist of different steps such as Data Pre-processing, Feature Selection, Bayesian Filter, Final Feature Weighting, Calculate total spam score & finally it shows filtering result that is mail is spam or non-spam. These steps are shown briefly as follows:

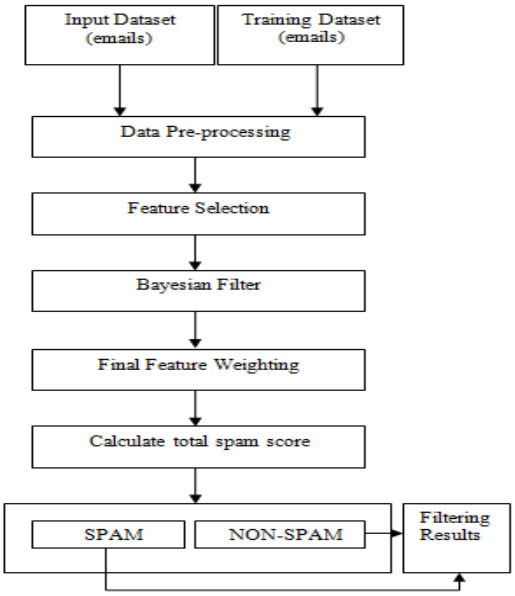

Fig.1 Proposed System Architecture

1) Data pre-processing:-Data preprocessing [1] involves transformation of the data into a format suitable for Bayesian filter. There are three steps in preprocessing task for email classification, which are tokenization, stop word removal and stemming. First step used is tokenization. In tokenizing process, all symbols (@, #, %,$), punctuations and numbers will be removed. The remaining strings will be split up into tokens. Second step is stopword removal. Many of the most frequently used words in English are useless in Information Retrieval (IR) and text mining. 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 this step, the common words, which are the most frequent words that exist in a document like ‘we’, ‘are’, ‘is’ and etc are removed. In English language, there are about 400-500 Stop words. Stop word list is based on word frequency. This process will identified which words those match with the stop word lists by comparing both of them. Removing these words will save spaces for storing document contents and reduce time taken during the searching process. Third step is stemming, which is done to eliminate suffix and prefix of a word, in other meaning to get only a root word in each term that occurred in email. 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 describe same or relatively close concepts in 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'.

2) Feature selection: - Feature selection involves analyzing data (such as a bunch of average emails) and determines which features (words) will help the most in classification, which can then be used to train a classifier. [7] TF (Term Frequency) method is used, which is one of the independent feature selection method, in order to select the best attributes to be used in probability calculations. Term frequency of each word in a document (TF) is a weight which depends on the distribution of each word in documents. It expresses the importance of the word in the document. [10]

3) Bayesian Filter: - Once the features have been selected then Bayesian filter is used to classify data. Classification using Bayesian filter is done according to the method defined by Paul Graham. According to Paul Graham, Bayesian spam filter recognize spam by looking at the words (tokens) in the messages, based on learning characteristics of spam versus ham. The filtering process starts with two sets of emails which are spam and legitimate. It will examine the content in both sets of emails and calculate spam probabilities based on the proportion of spam occurrences. Bayesian filter learns to distinguish spam from legitimate mail by looking at the actual mail received by each user.

4) Final Feature weighting:-Feature weighting, which seeks to estimate the relative importance of each feature with respect to the classification task and assign it a corresponding weight. It is used for improving classification robustness. [8]

5) Calculate total spam score:-Calculating the total spam score which is the minimum score required to mark a message as spam.

4. Implementation

Bayesian spam filtering is a statistical technique of E-mail filtering. It makes use of a naive Bayes classifier to identify spam e-mail. The classification using Bayesian filter is done according to the method defined by Paul Graham in [6] and contains the following steps:

1) the existence of a collection of ham and spam emails;
2) these collections are divided into words, named by Graham tokens, with the help of Predefined separators;
3) contouring the appearances of every word in those two collections;
4) The results offered by the stages presented above are consisted in 2 lists of word, one for the ham set and the other one for the spam, along with the number of appearances of each word in the lists mentioned.
5) This stage consists in calculating the spam probabilities for every word with the help of the 2 number of appearances.

Bayesian email filters take advantage of Bayes' theorem. Bayes' theorem is used several times in the context of spam. first time, to compute the probability that the message is spam, knowing that a given word appears in this message; a second time, to compute the probability that the message is spam, taking into consideration all of its words (or a relevant subset of them); Sometimes a third time, to deal with rare words.

Computing the probability that a message containing a given word is spam:

The formula used by the software to determine that is derived from Bayes' theorem

$$P_r(S|W) = \frac{P_r(W|S) \cdot P_r(S)}{P_r(W|S) \cdot P_r(S) + P_r(W|H) \cdot P_r(H)} \qquad (1)$$

Where

1) Pr(S|W) is the probability that a message is a spam, knowing that the word replica(Let's suppose the suspected message contains the word "replica".) word is in it;
2) Pr(S) is the overall probability that any given message is spam;
3) Pr(W|S) is the probability that the word "replica" appears in spam messages;
4) Pr(H) is the overall probability that any given message is not spam (is "ham");
5) Pr (W|H) is the probability that the word "replica" appears in ham messages.

The Spamicity of a word:

Most Bayesian spam detection software makes the assumption that there is no a priori reason for any incoming message to be spam rather than ham, and considers both cases to have equal probabilities of 50%.

$$P_r(S) = 0.5 \; ; P_r(H) = 0.5$$

The filters that use this hypothesis are said to be "not biased", meaning that they have no prejudice regarding the incoming email. This assumption permits simplifying the general formula to:

$$P_r(S|W) = \frac{P_r(W|S)}{P_r(W|S) + P_r(W|H)} \qquad (2)$$

This quantity is called "spamicity" (or "spaminess") of the word.

1) Pr(W|S) used in this formula is approximated to the frequency of messages containing "replica" in the messages identified as spam during the learning phase.
2) Pr(W|H) is approximated to the frequency of messages containing "replica" in the messages identified as ham during the learning phase.

Combining individual probabilities:

$$p = \frac{p_1 \; p_2 \ldots\ldots\ldots\ldots \; p_N}{p_1 \, p_2 \ldots\ldots\ldots \; p_N + (1 - p_1)(1 - p_2) \cdots\cdots (1 - p_N)} \qquad (3)$$

Where

1) p is the probability that the suspect message is spam;
2) p1 is the probability p(S|W1) that it is a spam knowing it contains a first word (for example "replica");
3) p2 is the probability p(S|W2) that it is a spam knowing it contains a second word.
4) p_n is the probability $p(S|W_n)$ that it is a spam knowing it contains an *N*th word.

5. Results & Analysis

The Table I show total emails taken for testing the system as well as how many emails are spam & non spam emails out of them & the Table II shows after testing the system we found result of spam & non-spam emails.

Table1. E-mails Taken Before Testing the System

Total emails	Spam emails	Non-spam emails
80	30	50

Table2. After Testing the System Result of Spam & Non-Spam E-mails

Spam & Non-spam emails	Total of emails
Emails are spam emails and identified as spam	28
Emails are spam emails but identified as non-spam	2
Emails are non-spam emails and identified as non-spam	44
Emails are non-spam emails and identified as spam	6

After the testing system, various performance measures such as the precision, recall & accuracy were observed as follows:-

Measure	Defined as	Value (%)
Accuracy =	(TP + TN) / (TP + FP + FN+ TN)	= 90
Precision =	TP / (TP + FP)	= 82.35
Recall =	TP/ (TP + FN)	= 93.33

The Fig.2 shown below is a graph of system tested result of spam or non-spam emails which shows 35% are spam emails & 55% are non-spam emails. The Fig.3 gives graph of manually & system generated result of emails. For manually (actual) result it shows 80% are total emails, 37.5% are spam emails & 62.5% are non-spam emails. For system generated result of emails it shows 80% are total emails, 35% are spam emails & 55% are non-spam emails.

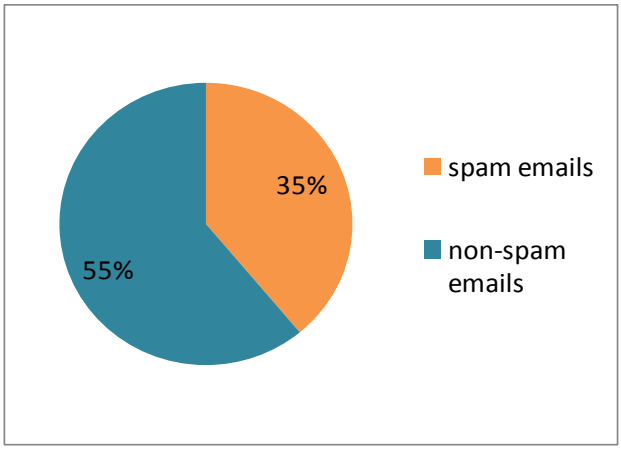

Fig.2 Graph of System Tested Result of Spam or Non-Spam emails

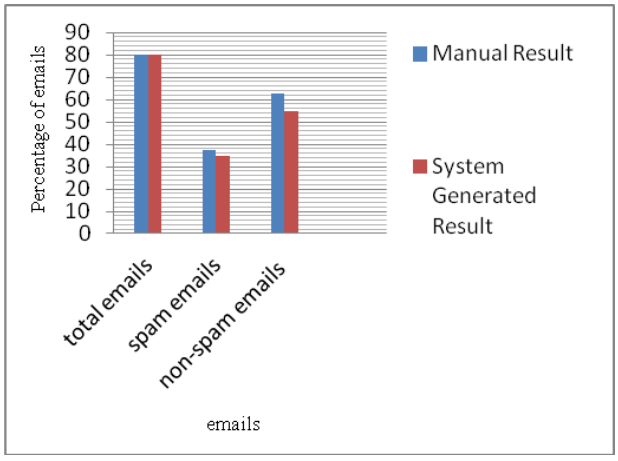

Fig.3 Manually & System Generated Result of emails

6. Conclusion & Future Work

Email spam classification has received a tremendous attention by majority of the people as it helps to identify the unwanted information and threats. Therefore, most of the researchers pay attention in finding the best classifier for detecting spam emails. In this paper spam filter is implemented to efficiently detect the spam emails using Bayesian filter approach. Classification using Bayesian filter is done according to the method defined by Paul Graham. The advantage of Bayesian spam filtering is that it can be trained on a per-user basis. The spam that a user receives is often related to the online user activities. The word probabilities are unique to each user and can evolve over time with corrective training whenever the filter incorrectly classified an email. As a result, Bayesian filter accuracy after training is often superior to pre-defined rules. However there are disadvantages of this technique i.e. with Paul Graham's scheme only the most significant probabilities are used, so that padding the text out with non-spam-related words does not affect the detection probability significantly. Also Bayesian spam filtering may be susceptible to Bayesian poisoning, a technique used by spammers in an attempt to degrade the effectiveness of spam filters that rely on Bayesian filtering.

After testing the system Different performance measures such as the precision, recall, & the accuracy etc. were observed. In future Random forest Classification Algorithm [9] implementation will be done to classify the stream data (such as emails) as spam or non-spam & the result of analysis from Classification using Bayesian filter with Random forest Classification algorithm will be compared. Random Forest algorithm is an ensemble method of predictive modeling. In 2001, Dr. Leo Breiman developed the Random Forest Algorithm which is a collection of many CART trees that are individually developed. The predictions of all trees are subjected to a voting procedure which aggregates the results. The voting determines the prediction of the final class of the algorithm. This voting is responsible for classifying Random Forest as a type of ensemble learning.

7. References

[1] M. Basavaraju, Dr. R. Prabhakar, " A Novel Method of Spam Mail Detection using Text Based Clustering Approach", Volume 5– No.4, August 2010.

[2]N.S. Kumar,D.P. Ran, R.G.Mehta,"Detecting E-mail Spam Using Spam Word Associations", International Journal of Emerging Technology and Advanced Engineering,Volume 2, Issue 4, April 2012

[3] Mehran Sahami & Susan Dumaisy, " A Bayesian Approach to Filtering Junk E-Mail ",Gates Building 1A Computer Science Department Microsoft Research Stanford University Redmond, WA 98052-6399,Stanford, CA.

[4] Ahmed Obied, "Bayesian Spam Filtering", Department of Computer Science University of Calgary amaobied@ucalgary.ca, http://www.cpsc.ucalgary.ca/amaobied

[5] Vikas P. Deshpande, Robert F. Erbacher, "An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques", Proceedings of the 2007 IEEE Workshop on Information Assurance United States Military Academy, West Point, NY 20-22 June 2007.

[6] http://www.paulgraham.com/spam.html

[7]Sebastien Gadat ,"A Stochastic Algorithm for Feature Selection in Pattern Recognition" , CMLA, ENS Cachan, 61 avenue du president Wilson, Cachan Cedex, France.
[8] Xinchuan Zeng and Tony R. Martinez, " Feature Weighting Using Neural Networks", Computer Science Department, Brigham Young University, Provo, Utah 84602.
[9] Hanady Abdulsalam, David B. Skillicorn, and Patrick Martin,"Streaming Random Forests", Quen's University Kingston, Ontario, Canada July 2008.
[10]V.Srividhya,,R.Anitha. "Evaluating preprocessing techniques in text categorization", International Journal of Computer Science & Application Issue 2010.
[11]Raju Shrestha and Yaping Lin,"Improved Bayesian Spam Filtering Based on Co-weighted Multi-area Information ",Department of Computer and Communication, Hunan University,Changsha 410082, P.R. China
[12] Denil Vira, Pradeep Raja & Shidharth Gada,"An Approach to Email Classification Using Bayesian Theorem", Global Journal of Computer Science and Technology Software & Data Engineering Volume 12 ,Issue 13 Version 1.0 Year 2012
[13] Michal Prilepok1, Jan Platos, Vaclav Snasel, and Eyas El-Qawasmeh,"The Bayesian Spam Filter with NCD", Department of Computer Science, FEI, VSB - Technical University of Ostrava, 17. listopadu 15, 708 33, Ostrava-Poruba, Czech Republic
[14]Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Georgios Sakkis,"Learning to Filter Spam E-Mail: A Comparison of a Naïve Bayesian and a Memory-Based Approach",Software and Knowledge Engineering Laboratory Institute of Informatics and TelecommunicationsNational Centre for Scientific Research "Demokritos"
153 10 Ag. Paraskevi, Athens, Greece.
[15] Grant Gross, 'Spam bill heads to the president', IDG NewsSevice,http://www.nwfusion.com/news/2003/1209spambill.html
[16]Zhan Chuan, LU Xian-liang, ZHOU Xu, HOU Meng-shu,"An Improved Bayesian with Application to Anti-Spam Email ", Journal of Electronic Science and Technology of China, Mar. 2005, Vol.3 No.1
[17] CNNIC. The 13th China Internet Development Status Report[R]. 2004
[18] R. Malathi,"Email Spam Filter using Supervised Learning with Bayesian Neural Network", Computer Science, H.H. The Rajah's College, Pudukkottai-622 001,Tamil Nadu, India, Int J Engg Techsci Vol 2(1) 2011,89-100.

Biographies

FIRST A. Bhagyashri U. Gaikwad received the B.E degree in Information Technology from the Shivaji University, Kolhapur, Maharashtra, in 2009. Pursuing MTech. degree in Computer Science and Technology from Department of Technology, Shivaji University, Kolhapur, Maharashtra, respectively.
Email Id:bhagyashrigkwd@gmail.com

SECOND B. Pratap P. Halkarnikar received B.E. from Government College of Engineering, Pune in 1986. M.E. from Walchand College of Engineering, Sangli in 1993. Presently he is working as Assistant Professor in Department of Computer Science And Engineering at D.Y. Patil College of Engineering, Kolhapur. He is consultant to many industries for development of microcontroller based products. His interest lies in microcontroller based instrumentation, computer vision, data mining and web technology. He is member of ISTE and IEI.
Email Id: pp_halkarnikar@rediffmail.com