ISSN: 2321-7782 (Online)
Volume 4, Issue 1, January 2016
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Spam Email Classification & Blocking

Guide Name: **K. S. Warke**
BV College of engineering for women,Pune, India

Nilam Salunkhe[1]
Department Of Computer Engineering
BV College of engineering for women, Pune
Pune, India

Shweta Kulkarni[2]
Department Of Computer Engineering
BV College of engineering for women, Pune
Pune, India

Pallavi Kamble[3]
Department Of Computer Engineering
BV College of engineering for women, Pune
Pune, India

Pooja Nage[4]
Department Of Computer Engineering
BV College of engineering for women, Pune
Pune, India

Abstract: ***Spam email has already caused many problems such as taking recipient time and wasting network bandwidth. It is time consuming and laborious to remove spam email by hand if there are too many spam email in mailbox. Thus, automatic classification of spam email from legitimate email has become very important. Decision tree and Filtering Methods are two popular and powerful techniques in machine learning community. In this study, C4.5 classification method based on decision tree and Naïve Bayes Filtering Method is introduced to filter the spam email effectively. We also block the selected category of Spam email and unblock it.***

Key words: Data Mining, Spam Email, WEKA Server, Decision tree Algorithm, Spam Filtering, Blocking and Unblocking.

I. INTRODUCTION

DATA MINING:

Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behaviour of their customers and potential customers. It discovers information within the data that queries and reports can't effectively reveal. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

DECISION TREE:

Decision tree is very popular and powerful tool in data mining community and the rules generated by decision tree are simple and accurate for most problems. The decision tree is one of the most famous tools of decision-making theory. Decision tree is a in the form of a tree structure to show the reasoning process. Each node in decision tree structures is either a leaf node or a decision node. The leaf node indicates the value of the target attribute of instances. The decision node indicates two or more branches and each branch represents values of the attribute tested. When classifying an unknown instance, the unknown instance is routed down the tree according to the values of the attributes in the successive nodes.

SPAM E-MAIL:

The problem of undesired electronic messages is nowadays a serious issue, as spam constitutes up to 75{80% of total amount of email messages. Spam causes several problems, some of them resulting in direct financial losses. More precisely, spam causes misuse of traffic, storage space and computational power. spam makes users look through and sort out additional email, not only wasting their time and causing loss of work productivity, but also irritating them and, as many claim, violating their privacy rights. Finally, spam causes legal problems by advertising pornography, pyramid schemes, etc. The total worldwide financial losses caused by spam in 2005 were estimated by Ferris Research Analyzer Information Service at $50 billion.

The ever increasing menace of spam is bringing down productivity. More than 70% of the email messages are spam, and it has become a challenge to separate such messages from the legitimate ones. We have using a spam identification engine naïve Bayesian classifier to identify spam. The classical naïve Bayesian approach was used to develop the spam filter.

II. PROPOSED APPROACH

Decision tree:

The decision tree is one of the most famous tools of decision-making theory. Decision tree is a classifier in the form of a tree structure to show the reasoning process. The following figure shows the architecture of our system.

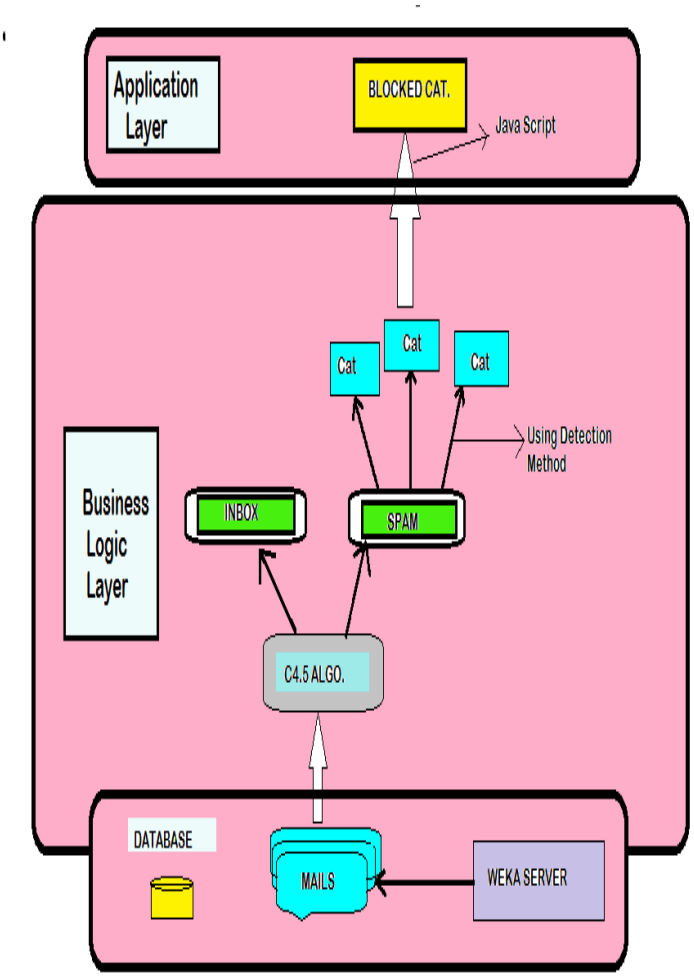

Fig . System Architecture

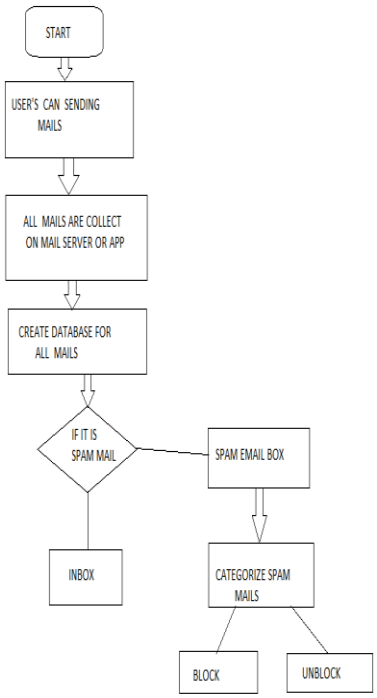

Fig . Flow Diagram

III. WEKA SERVER

Weka is a machine learning software suite developed in Java. It provides facilities for all the steps involved in solving a machine learning problem- data conversion, preprocessing techniques, classification, categorization and visualization. Weka commands can be carried out via the command line.

Out of all the other machine learning libraries, Weka was the most cited and recommended. It was the most well documented, and also since it was written in Java, understanding and customization became more convenient.

WEKA- Waikato Environment for Knowledge Analysis

Weka contains tools for Data pre-processing, Classification, Clustering etc.

Open source library or machine learning software in java.

Classification of a very large data set, and can be hosted on a cloud server.

IV. DECISION TREE ALGORITHM

We are using c4.5 decision tree algorithm for Spam email classification. C4.5 is an algorithm developed by Ross Quinlan that generates Decision Trees (DT), which can be used for classification problems. It improves (extends) the ID3 algorithm by dealing with both continuous and discrete attributes, missing values and pruning trees after construction. Data classification is a form of data analysis that can be used to extract models describing important data classes. There are many classification algorithms but decision tree is the most commonly used algorithm because of its ease of implementation and easier to understand compared to other classification algorithms**.**C4.5 is one of the most effective classification method.C4.5 is implementable in WEKA server. c4.5 is a program that creates a decision tree based on a set of labeled input data.This decision tree can then be tested against unseen labeled test data to quantify how well it generalizes. C4.5 has additional features such as tree pruning, improved use of continuous attributes, missing values handling and inducing rule set.

C4.5 is collection of algorithms for performing classifications in machine learning and data mining. It develops the classification model as a decision tree. C4.5 consists of three groups of algorithm: C4.5, C4.5-no-pruning and C4.5-rules. Now, we will focus on the basic C4.5 algorithm.

Algorithm

C4.5 is implemented recursively with this following sequence.

Generate _ decision _ tree.

Generate a decision tree from the training tuples of data partition D.

Input:

Data partition, D.

Attribute _ list.

Attribute_ selection_ method.

Method:

1. create a node N.
2. if tuples in D are all of the same class, then
3. return N as a leaf node labeled with the class C and exit . Then returns the Attributes.
4. If attribute _list is empty then, return N as a leaf node labeled with the majority class in D and exit.
5. If the tuples of the given class are not same,then Calculate entropy (information gain) of database.
6. given formula-

$$\text{Info}(D) = \sum_{i=1}^{c} - \text{pi log2pi}$$

7. Calculate gain of each attribute:

$$\text{Gain}(A) = \text{Info}(D) - \text{Info A}(D)$$

8. Select the attribute with the maximum gain and store it in the 'max' variable, i.e.

$$\text{max} = \text{Info gain}(A)$$

9. Label node N with max.
10. After determining the possible outcomes(j) of the attribute A, we sort the database(D) on the basis of that attribute to form the database (Dj).
11. If the attribute A contains discrete value and then, attribute _list - attribute(A)
12. For each outcome of the attribute A, if tuples in (Dj) are all of same class, then return as a leaf node, else goto step 4.
13. Return the decision tree.

C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan to address the following issues not dealt with by ID3:

a. Avoiding overfitting the data
b. Determining how deeply to grow a decision tree.
c. Reduced error pruning.
d. Rule post-pruning.

e. Handling continuous attributes. e.g., temperature

f. Choosing an appropriate attribute selection measure.

g. Handling training data with missing attribute values.

h. Handling attributes with differing costs.

i. Improving computational efficiency.

V. SPAM FILTERING

In this paper, we are using The Naïve Bayes Classifier method for filtering Spam emails.

The Naïve Bayes Classifier method:

The classical naïve Bayesian approach was used to develop the spam filter.

The use of naïve Bayesian classifier has become highly prevalent as the ensuing system will be less complex. Naïve Bayesian classifier is a probabilistic classifier based on Bayes' theorem. The theorem assumes that each feature is conditionally independent of each other. In 1998 the Naïve Bayes classifier was proposed for spam recognition.

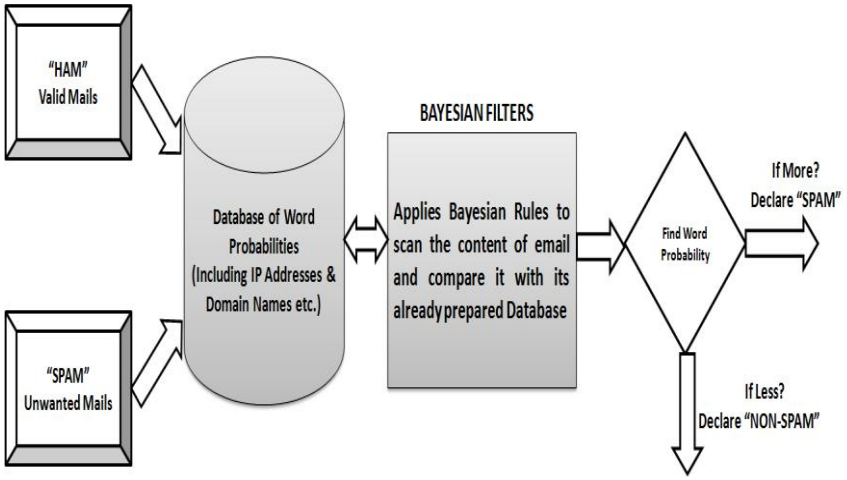

Fig. Bayesian Filter Method

Bayesian classifier is working on the dependent events and the probability of an event occurring in the future that can be detected from the previous occurring of the same event. This technique can be used to classify spam e-mails; words probabilities play the main rule here. If some words occur often in spam but not in ham, then this incoming e-mail is probably spam. Naïve bayes classifier technique has become a very popular method in mail filtering software. Bayesian filter should be trained to work effectively. Every word has certain probability of occurring in spam or ham email in its database. If the total of words probabilities exceeds a certain limit, the filter will mark the e-mail to either category. Here, only two categories are necessary: spam or ham. Almost all the statistic-based spam filters use Bayesian probability calculation to combine individual token's statistics to an overall score and make filtering decision based on the score. The statistic we are mostly interested for a token T is its spamminess calculated as follows:

S [**T**] = **C** Spam**(T)** / **C** Spam**(T)** + **C** Ham**(T)**

Where **C**Spam(T) and **C**Ham(T) are the number of spam or ham messages containing token T,respectively. To calculate the possibility for a message M with tokens {T1,.....,TN}, one needs to combine the individual token's spamminess to evaluate the overall message spamminess. A simple way to make classifications is to calculate the product of individual token's spamminess and compare it with the product of individual token's hamminess.

Stage1. Training

Parse each email into its constituent tokens

Generate a probability for each token W

S[W] = Cspam(W) / (Cham(W) + Cspam(W))

store spamminess values to a database.

Stage2. Filtering

For each message M

while (M not end) do

scan message for the next token Ti

query the database for spamminess S(Ti)

calculate accumulated message probabilities

S[M] and H[M]

Calculate the overall message filtering indication by:

I[M] = f(S[M] , H[M])

f is a filter dependent function,

such as,

I [M] = 1+S[M]-H[M]/2

if I[M] > threshold

msg is marked as spam

else

msg is marked as non-spam

VI. BLOCKING AND UNBLOCKING

In this, we are using some standard algorithms of decision tree like C4.5 and Naïve bayes for spam email classification. We are implementing these algorithms in WEKA server as we discussed earlier. The best worldwide Companies which provide email services like Google, Yahoo, Rediffmail etc. are not providing Blocking Facility of Spam email. So there is a Need of Spam email blocking in our daily routine. We are using the Java script for blocking spam emails. But some of the spam emails are important so we are providing Unblocking Facility also.

VII. CONCLUSION

In this paper, an naive bayes and decision tree based approach is proposed to classify spam emails. Extensive experiments conducted on a public spam email dataset indicate that the proposed algorithm outperforms the popular classification techniques including of C4.5, Naive Bayes.Our future work is to block the spam email and also unblock it.

ACKNOWLEDGEMENT

The project topic would not have seen the light of the day without the whole-hearted support of our guide Prof. K.S.Warke. We admire her infinite patience and understanding that she guided us in a field we had no previous experience. Mam also guided us through the essence of time management, presentation skills and how vital it is for an engineer to think from a research perspective. Whenever we approached her, she explained the concepts lucidly, so that it would be simplified and be vivid in our mind. Again, we also thank to her and all the faculty members who have made the journey of our faculty directions in this domain. We thank to all of our friends and teachers, who have attended our seminar sincerely.

References

1. L. Shi et al. /Journal of Computational Information Systems 8: 3 (2012) 949{956}
2. A. Wiehes, "Comparing Anti Spam Methods", Master Thesis, Master of Science in Information Security, Department of Computer Science and Media Technology, Gjøvi University College, 2005.
3. A. Wosotowsky, E. Winkler, "Spam Report McAfee Labs Discovers and Discusses Key SpamTrends"2009,Retrieved10th Feb2010 http://www.mcafee.com/us/local_content/reports/7736rpt_s pam_1209.pdf
4. C. Gulyás, "Creation of a Bayesian network- based meta spam filter, using the analysis Of different spam filters", Master Thesis, Budapest, 16th May 2006.
5. International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online) Vol. 2, Issue 2, pp: (8-14), Month: April - Jun 2014, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011
6. Enrico Blanzieri, University of Trento, Italy Anton Bryl, University of Trento, Italy Net Trento, Italy (October 16, 2007).
7. International Journal of Scientific and ResearchPublications, Volume 3, Issues October 2013 1 ISSN 2250-3153
8. Juneja et al., Orient. J. Comp. Sci. & Technol., Vol. 3(2), 305-310 (2010)
9. International Journal of Computer Science and Information Technologies, Spam Filter Project, Report 2014, 45345
10. International Journal of Advanced Information Science and Technology (IJAIST)

