International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol. 1, Issue 2, December 2014

Spam Mail Detection Using Fuzzy Similarity

[1]Swati Pandit,[2]Prof. Vanita Mane,[3]Prof. Rajashree Shedge
[1][2][3]Ramrao Adik Institute of Technology, Nerul, Navi Mumbai
[1]swatipandit02@gmail.com,[2]vanita.mane@rait.ac.in ,[3]rajashree.shedge@gmail.com

Abstract - **Spam is an unsolicited bulk mail or junk email. Spam emails not only waste resources such as bandwidth, storage and computation power, but also the time and energy of email receivers who must search for legitimate emails among the spam and take action to dispose the spam. So for overcoming these problems Spam filtering techniques are developed which classify messages among two categories, spam and non-spam. Different decision tree algorithms are studied for spam classification then they are compared analyzed and evaluated on the basis of various measures as Feature preprocessing, Feature Extraction, Measure of best split, Types of Attributes, and Detection rate . Finally fuzzy similarity measure algorithm is proposed which gives higher accuracy and low false positive and low false negative rate.**

I. INTRODUCTION

Spam is the use of electronic messaging systems (including most broadcast media, digital delivery systems) to send unsolicited bulk messages indiscriminately. E-mail spam, also known as junk e-mail or unsolicited bulk e-mail (UBE), is a subset of spam that involves nearly identical messages sent to numerous recipients by e-mail. Day by day the amount of incoming spam increase and, spammer attacks are becoming targeted and consequently more of a threat. Subsequently, attacks have increased further from 10 per day to approximately 60 per day in 2010 [1].

Increasingly today large volumes of spam emails are causing serious problems for users, Internet Service Providers, and the whole Internet backbone. Spam emails not only waste resources such as bandwidth, storage and computation power, but also the time and energy of email receivers who must search for legitimate emails among the spam and take action to dispose the spam [2]. The extent of the spam problem has forced many organizations to deploy a spam filtering solution. Spam filters are divided into reputation based and content based. Reputation based filters into three major's techniques which are
origin based (b) social based and (c) traffic

analyzing. Content based filters detect spam by examining the content of email messages, irrespective of the origin. There exist several families of content based filtering techniques, including (a) heuristics (b) machine learning and (c) finger printing [3]. In spam filtering process that are content based, the basic steps are preprocessing, feature extraction and the final step is classification [4]. The basic step is done in data pre-processing is stopping and stemming. Stopping is the process of removal of words that are lesser in length (i.e., words with length less than specified value like 2 or 3), frequently occurring words and special symbols. For grammatical reasons, documents are going to use different forms of a word. Stemming reduces derivation related forms of a word to a common base form. Feature extraction is the use of one or more transformations of the input features to produce new salient features. Finally, classification step is there where emails are classified into ham and spam emails. For classification different algorithms were developed which were using different techniques like heuristics, machine learning [5].

One of the categories of machine algorithms is decision tree which build the decision tree by using specific attributes list and then from decision tree rules are constructed. Then these rules are applied for classification of emails. In the report different decision tree algorithms which are studied that is Binary decision tree, Random Forest, Fuzzy Decision tree and Iterative Dichotomizer 3. All algorithms are text based have their own features and efficiency, limitations. Proposed algorithm is based on fuzzy similarity measure. In proposed method membership value of all the tokens in both the categories are considered so that accuracy is improved.

Section II explains about decision tree algorithms proposed by different authors. Section III
Describes Problem definition continues with proposed system in section IV. In section IV analysis is done and paper is concluded in section V.

II. RELATED WORK

Many algorithms are developed for spam filtering some of them are machine learning like decision tree, rule based classifier, neural network, support vector machine and naïve Bays classifier. The algorithms which are studied are

decision tree based classifier algorithm. Yun Quing Xia proposed the binary decision tree algorithm [11]. The optimal partitioning method on the whole collection of training email classes is applied and collection is divided into two groups considering each class as a whole. When this work is successfully done, the optimal partitioning is continuously applied on each of the two sub-groups. This algorithm repeats this process until each sub-group contains only one class. Binary decision tree algorithm is computationally less complex. S. Naksomboon proposed the random forest algorithm for spam email classification [12]. The first part is the preprocessing part where the words are selected which are more likely to be spam along with some predefined spammer behavior features which will act as keywords for classification. In second that is classification part dataset will be divided into many sample set and for each sample set decision tree is grown by using gini index or gain ratio. Seven behavior features taken from email header are used. Since few keywords are used in classification part it reduces the computation time, memory consumption. Wang Meizhen proposed the Fuzzy decision tree for email spam classification [14]. Fuzzy Decision Tree based spam filtering system architecture includes two main components that are feature extractor and FDT analyzer. Before data mining, we need to analyze emails behavior features from email logs. Author select Information Gain technology to analyze these features, to get the main feature, to omit some features with less information and weak correlation. Yiwen Zhang developed the Iterative Dichotomizer 3 algorithm for email spam filtering [16]. Mails are selected and put into several types such as games, education, marriage, work, entertainment, shopping and promotional products. Keyword set is defined for each category. The table of attribute keyword set and email is formed which shows the frequency of keywords in each email. By using that mutual information is calculated and decision tree is formed and according to tree rules are established to decide mail is spam or not.

Among all above algorithms ID3 gives best detection rate. Bu its misclassification of email rate is more because keywords taken for classification part are fixed and more in number so complexity gets increases. So algorithm with fuzzy similarity measure is proposed where the membership value of all tokens in messages with respect to both the categories are calculated. Then membership degree of the token from incoming mail is calculated and with help of both membership value and membership degree parameter similarity measure is calculated for spam and ham category. Then it will be compared with some threshold value. And it will get classified in either category.

III. PROBLEM DEFINITION

ID3 and its descendants only allow testing a single attribute and branching on the outcome of that test. Another shortcoming is that ID3 relies much on the quality of training data set that is it will not properly work in case of missing values in database. Once decision tree is made it will not be changed. The keywords used for constructing decision tree are binary attributes that is they can only have values exist or not exist. In ID3 depth of tree is more because we have to consider all categories token and again false positive and false negative rate is more because spam tokens need not necessary to be present in spam mails only, they can be occurred in legitimate mails also. In fuzzy similarity measure since the membership values of keywords in training database is stored into tables and it will be updated after every mail which will increase the detection rate. Again the rate of misclassified mails will get decreased.

IV. PROPOSED METHOD

Using a fuzzy similarity approach, a classification model is built from a set of pre-classified e-mail instances. The various stages of this approach are

- Pre-processing
- Training
- Classification

All the stages are explained in detail further. Before that in figure 4.1 the framework of proposed system that is spam mail detection using fuzzy similarity measure is shown.

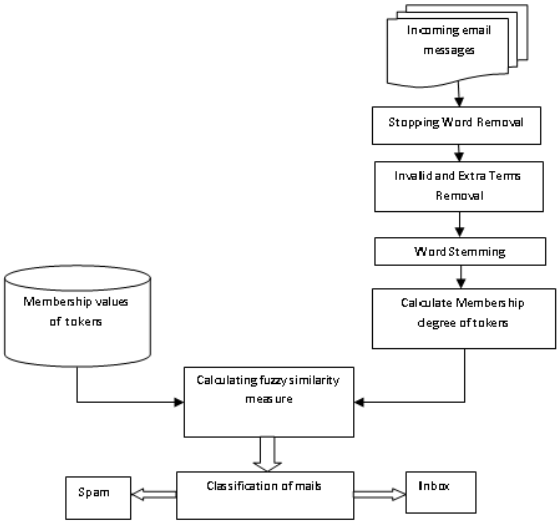

Fig. 4.1 Framework of Proposed Approach

A. *Pre-processing*

Before e-mail messages are used for training and classification, some pre-processing needs to be done in order to reach optimum results. The figure 5.2 below illustrates the pre-processing phase. First, all HTML tags are stripped off. Then, all stop words, i.e. words that appear frequently but have low content discriminating power, are removed from each e-mail message. Examples of such

words include 'a', 'an', 'and', 'the', 'that', 'it', 'he', 'she' etc. The message is then tokenized into a set of strings separated by some delimiters, e.g. whitespaces. These tokens can represent words, phrases or any keyword patterns. All mixed-case tokens are converted to lowercase. However, if a token is all uppercase, it will be treated differently than if it is all lowercase. The resulting set of tokens are stemmed to their roots (i.e. replacing each token with its base form) to avoid treating different forms of the same word as different attributes; thus reducing the size of the attribute set. For example, both "promotion" and "promoting" are converted to the same base form "promote". Also if a token appears few times in either category (e.g. less than three times), it is removed. Now, tokens from all messages are combined into one vector T= $<t_1, t_2, t_n>$ where N is the total number of tokens. Also the number of occurrences of each token, t_i, in each category, c ϵ {spam, legitimate}, is determined.

B. Training Phase

During the training phase, a model is built based on the characteristics of each category in a pre-classified set of e-mail messages. The training dataset should be selected in such a way that it is varying in content and subject. Each sample message is labeled with a specific category. We first perform pre-processing to extract tokens and determine the number of occurrences of each token in each category. Let C = {spam, valid} represent the category set, T denote the set of tokens, and $f_{i,c}$ denote the frequency of token t_i in category c. From these data, we define a fuzzy token-category relation which maps each element in T × C to a membership value between 0 and 1, i.e. R: T × C →[0, 1]. Thus, the fuzzy token-category relation is given by,

$$R= \{< (t_i, c_j), \mu_R(t_i, c_j) > | (t_i, c_j) \epsilon T \times C\} \quad (1)$$

Where, $\mu_R (t_i, c_j)$ represents the membership degree of token t_i in category c_j. The value of $\mu_R (t_i,c_j)$ is calculated by dividing the total number of occurrences of token t_i in category c_j by the total number of occurrences of token t_i in all categories. More specifically, $\mu_R (t_i,c_j)$ is given by the following equation,

$$\mu_R(t_i, c_j) = \frac{f_{i,c_j}}{f_{i,\text{legitimate}} + f_{i,\text{spam}}} \quad (2)$$

Where $c_j \epsilon$ {spam, legitimate}
f_i frequency of token t_i
This implies that if a token occurs only in one category, then its membership to this category will be one, and to the other category will be zero. Typically, the membership values will range between zero and one. Having constructed a knowledge base for the token-category membership, spam filtering is now based on calculating the fuzzy similarity measure between the received message d and each category, i.e. spam or legitimate. In order to calculate fuzzy similarity, we must first determine the membership degree $\mu_d(t_i)$ of each token t_i to the message d. One way to do that is by first determining the frequency of each token in the message. The membership degree is then defined as,

$$\mu_d(t_i) = \frac{f_{i,d}}{\max_{t_j \epsilon d} \{f_{i,d}\}} \quad (3)$$

Where $f_{i,d}$ is the number of occurrences of token t_i in message *d*.
Thus, the token with the maximum number of occurrences will be assigned a value of 1, and all other tokens will be assigned proportional values.
The fuzzy similarity measure (SM) is given by,

$$SM(d, c_j) = \frac{\sum_{t \epsilon d} \mu_R(t,c_j) \otimes \mu_d(t)}{\sum_{t \epsilon d} \mu_R(t,c_j) \oplus \mu_d(t)} \quad (4)$$

Where, $\otimes$ is fuzzy conjunction operator (t-norm), and $\oplus$ is fuzzy disjunction operator (t-co norm). In this way we get two separate lists of legitimate and spam tokens. These two lists will be passed to classification process. So only spam token list will be passed to the classification process where all the spam tokens will compared against threshold value.

C. Algorithm for Fuzzy Similarity Approach

1. Pre-process the email messages.
2. Collect all the tokens in each category(spam or legitimate)
3. Count the occurrences of each token and update them in the database
4. Count the occurrence of each token in other category (in case of spam, count occurrence of token in ham and vice versa)
5. alculate the total frequency for each token $F_{total} = F_{spam} + F_{ham}$

   Where F_{total} is total frequency, F_{spam} is frequency of token in spam category and F_{ham} is frequency of token in ham category.
6. Assign membership value to token in each category using formula

   $\mu_{spam} = F_{spam} / F_{total}$
   $\mu_{ham} = F_{ham} / F_{total}$
   and update token along with membership values in database.
   Where μ_{spam} is membership value in spam category and μ_{ham} is membership value in ham category.
7. Take a test mail to decide if it is a spam or ham.
8. Preprocess the test mail.
9. Check the tokens in database to decide if its valid token or not
10. Then make a list of valid tokens from test mail and count occurrences (F_{test}) of each token in test mail.
11. Find the token with maximum number of occurrences (F_{max}).

Method Name / Parameter	Iterative Decision 3 [16]	Binary Decision Tree[11]	Random Forest[14]	Fuzzy Decision Tree [12]	Fuzzy Simlarity Measure
Feature Processing	Decision Tree Generation using code	Using Optimal Binary Partioning	Behavior Characteristics Processing	Fuzzy decision Tree Generation	Membership Value Calculation
Feature Extraction	Species of Keywords Extraction	Document Similarity	Spammer Behavior and word Frequency	Spam Behavior	All the keywords in message
Measure for Best Split	Mutual Inforamation Gain	Cosine Similarity	T Test Statistical Technique	Information gain	No Splitting
Types of Attributes	Binary Attributes	Binary Attributes	Binary and Continuous Attributes	Binary and Continuous and ordinal Attributes	Continuous Attributes
Spam Detection Rate	High	Average	High	High	Higher than ID3

12. Assign a membership degree to each token in database using formula

    $\mu_d = F_{test} / F_{max}$
    Where μ_d is membership degree of token

13. Calculate conjunction operator using formula

    Spam: $NUM_{spam} = \max \{0, (\mu_{spam} + \mu_d - 1)\}$
    Ham: $NUM_{ham} = \max \{0, (\mu_{ham} + \mu_d - 1)\}$

14. Calculate disjunction operator using formula
    Spam: $DEN_{spam} = \min \{1, (\mu_{spam} + \mu_d)\}$
    Ham: $DEN_{ham} = \min \{1, (\mu_{ham} + \mu_d)\}$

15. Calculate Similarity measure for spam: $SM_{spam} = NUM_{spam} / DEN_{spam}$
16. Calculate Similarity measure for ham: $SM_{ham} = NUM_{ham} / DEN_{ham}$
17. Calculate Threshold $\lambda = SM_{spam} / SM_{ham}$
18. If calculated λ is greater than threshold λ then mail is spam.

In this algorithm membership values of tokens are stored in database. Whenever new mail comes membership degree is calculated of each token and using both these parameters similarity measure is calculated. Membership value will get update after every incoming mail so it gives more accuracy.

D. Implementation Platform

For implementing Fuzzy similarity measure algorithm JDK and Net Beans IDE 7.2.1 platform is used for front end. For storing database SQL Server 2005 is used.

V. ANALYSIS

Analysis is done using following parameters Feature Processing, Feature Extraction, Measure for Best Split, Types of Attributes and Spam Detection Rate. All algorithms have their advantages and drawbacks but among all algorithms ID3 gives high detection rate. Spam detection using fuzzy similarity gives higher detection rate than ID3 algorithm.

VI. CONCLUSION

Since decision tree have predictive power, interpretability and computational scalability, it is taken for studied in detail. In spam filtering there are various text

based algorithms among which ID3 gives best accuracy. But depth of tree is more because we have to consider all categories token and again misclassified rate is more because spam tokens need not necessary to be present in spam mails only, they can be occurred in legitimate mails also. By using fuzzy similarity token category relationship to find the probability of occurrence of token into specific category. Again membership degree of every token of incoming mail is calculated and then by using these both parameters similarity measure is calculated. And it is compared against threshold value. Because of this rate of misclassified mails will decreased and automatically high accuracy rate is achieved. One more advantage is depth of decision tree will decrease.

REFERENCES

[1] Saadat Nazirova, "Survey on Spam Filtering Techniques", Communications and Network, 11, 3,153-60, August 2011.
[2]Bhavana Dhakare, Ujwala Gaikwad, "Spam Detection and Filtering using different techniques", International Journal of computer Applications, 2012.
[3]Hasan Alkahtani, Robert Goodwin, "A Taxonomy of Email Spam Filters", Communication and Network, 2010.
[4] Rohan M. Amin, "Detecting Targeted Malicious Email", IEEE Computer and Reliability Societies, 2012.
[5] M. Basavaraju, Dr.R. Prabhakar, "A Novel Method of Spam Mail Detection using Text based Clustering Approach", International Journal of computer Application, August 2010.
[6] http://www.spamlaws.com/spam-stats.html.
[7]http://www.windowsecurity.com/whitepapers/anti_spam/Impact-Reducing-SPAM-Part1.html.
[8] Pang Ning Tan, "Introduction to Data mining, Addison", Wesley Publication, 2002.
[9] Oded Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook", 2011.
[10] Albert Montillo, "Random Forests", Rutger University, Guest lecture, 2012.
[11] Yun Qing Xia,"A Binarization approach to Email Categorization using Binary Decision Tree", Machine Learning and Cybernetics, August 2007.
[12] S. Naksomboon, "Considering behavior of Sender in Spam Mail Detection", International Journal of computer Application, 2006.
[13] Irena Koprinska, "Learning to classify Email", Journal of Science direct, December 2006.
[14] Wang Meizen, Li Zhitang, Zhong sheng, "A method for Spam Behavior Recognition based on Fuzzy Decision Tree", IEEE Computer Society, 2009.
[15] Yonghong Peng, "Dicsretization to Enhance the continuous Decision Tree Induction", Espirit Metal Project, 2006.
[16] Yiwen Zhang, Lili Ding, Yun Wang, "Research and Design of ID3 Algorithm Rule Based Anti-spam Filtering", IEEE Computer Society, 2011.
[17] Wei Peng, "An implementation of ID3, Decision Tree Algorithm", New South Wales University, 2011.
[18] Brijain R Patel, "Use of Renyi Entropy Calculation Method for ID3 Algorithm for Decision Tree Generation in Data Mining", IJARCSMS Journal, 2014.
[19]Sarwat Nizmani, "Modeling Suspicious Email Detection using Enhanced Feature Selection", International Journal of Modeling and Optimization, Vol. 2, No. 4, August 2012.
[20]www.wikipedia.org