A Fuzzy approach for Spam Mail Detection integrated with wordnet hypernyms key term extraction

Ms. Subodhini Gupta[1], Mr.B.S.Parekh[2], Mr. Jaimin N Undavia[3]

[1] Department of Computer Science,
V.P.& R.P.T.P.science College
Sardar Patel University,V.V.Nagar,Gujarat,India

[2] Department of Computer Science and engineering,
Maharaja Sayajirao University ,Baroda Gujarat

[3] Asst Professor, CMPICA, CHARUSAT, Changa

ABSTRACT-
The term "spam" is sometimes used loosely to mean any Message broadcast to multiple senders (regardless of intent) or any message that is undesired. Receiving spam is a common complain of many Internet users. In fact, spam email has become an increasingly bothersome problem as individuals spreading spam email find easier ways to invade users' email accounts, leading to the necessity of such tools as spam filters and spam blocker features. Email spam is a topic that requires little introduction. In 2010, it was estimated that spam compromised nearly 90% of all email sent Consuming significant resources. Many data mining and machine learning researchers have worked on spam detection and filtering , commonly treating it as a basic text classification problem. This paper proposes an efficient yet simple fuzzy based simple method applied on refined key terms set extracted from the email using wordnet and hypernyms concept to filter spam mail.

Keywords:-*Spam, fuzzy, spam filters, spam detection, spam email, Wordnet, Hypernyms.*

1) INTRODUCTION:-

Spam is also known as unsolicited Commercial Email (UCE) and unsolicited Bulk Email (UBE) or junk mail[12]. E mail is undoubtedly a very effective, cheap and easy method of communication these days. Spam is flooding the Internet with many copies of the same message, in an attempt to force the message on people who would not otherwise choose to receive it. Most spam is commercial advertising, often for dubious products, get-rich-quick schemes, or quasi-legal services. In addition to wasting people's time with unwanted e-mail, spam also eats up a lot of network bandwidth. Consequently, there are many organizations, as well as individuals, who have taken it upon themselves to fight spam with a variety of techniques. But because the Internet is public, there is really little that can be done to prevent spam, just as it is impossible to prevent junk mail. 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. They also use a practice known as "e-mail appending" or "expending" in which they use known information about their target (such as a postal address) to search for the target's e-mail address.

Improving Spam filtering is a worthy goal in itself because it faces so many challenges like

i) Skewed and drifting class distribution:-like most text classification domains ,spam presents the problem of a skewed class distribution I.e. the proportion of spam to legitimate email is uneven. In spite of claims that spam is generally increasing the volume varies considerably and non-monotonically on a daily or weekly scale. Calculating spam proportion even approximately is difficult.

ii) Unequal and uncertain error costs: - A further complication of spam filtering is the asymmetry of error costs. Viewing the filter as a spam detector, a spam message is a positive instance and a legitimate message is a negative instance. Judging a legitimate email to be spam is usually far worse than judging a spam email to be legitimate .A false negative simply cause slight irritation i.e. the user sees an undesirable message. on the other hand a false positive can be critical.

iii) Disjunctive and changing target concept:-The content of spam changes over time, as class contained feature probabilities will change as well. Some spam topics are so common as advertisement for sites, offer for mortgage re-financing, and money making schemes.

iv) Intelligent adaptive adversaries:- The spam stream changes over times as different products or scams, marketed by spam come into vogue. There is a separate reason for concept drift. Spammer has become increasingly sophisticated in their techniques for evading filtering .In its early days spam would have predictable subject lines like MAKE MONEY FAST! And refinance your mortgage. as basic header filtering become common in e-mail clients ,these obvious text markers were simple to filter upon so spam could be discarded easily. It is now common to see fragments such as:

100% mo|ney back guaran|tee

Because of these challenges spam filtering techniques becomes complex. Proposed techniques handle these challenges yet simple and flexible enough.

2) RELETADED WORKS

Knowledge engineering and machine learning are the two general approaches used in e-mail filtering. In knowledge engineering approach a set of rules has to be specified according to which emails are categorized as spam or ham. A set of such rules should be created either by the user of the filter, or by some other authority (e.g. the software company that provides a particular rule-based spam-filtering tool). By applying this method, no promising results shows because the rules must be constantly updated and maintained, which is a waste of time and it is not convenient for most users. Machine learning approach is more efficient than knowledge engineering approach; it does not require specifying any rules [4]. Instead, a set of training samples, these samples is a set of pre classified e-mail messages. A specific algorithm is then used to learn the classification rules from these e-mail messages. machine learning approach has been widely studied and there are lots of algorithms can be used in e-mail filtering. They include Naïve Bayes, support vector machines, Neural Networks, K-nearest neighbor, Rough
sets and the artificial immune system.

There are some research work that apply machine learning methods in e-mail classification, Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali[14] They demonstrated that the naïve Bayes e-mail content classification could be adapted for layer-3 processing, without the need for reassembly. Suggestions on predetecting e-mail packets on spam control middle boxes to support timely spam detection at receiving e-mail servers were presented. M. N. Marsono, M. W. El-Kharashi, and F. Gebali[13] They presented hardware architecture of na¨ıve Bayes inference engine for spam control using two class e-mail classification.

Sudhakar.P, Poonkuzhali.S, Thiagarajan.K and Sarukesi.K[2]., suggested Fuzzy Logic for E-mail Spam deduction a new technique for spam categorization couple with header information and content information. However this system is under research in peer to peer networks. Even though the conceptualization is good, but the practical bottle neck will comes for identification of spam words from the global set. This will take large amount of time as it works with centralized architecture.

3) FRAME WORK FOR SPAM FILTERING:

In the proposed approach we have considered that we already have blacklisted spammer address list, blacklisted IP address list, suspicious subject word list, suspicious content word list and virus list that may be attached with document.

Our whole approach can be divided into two module .the first module is for the refinement and to reduce the volume and dimensionality of the candidate mail .we would like to reduce the volume before its further processing because many algorithm work fine on small document set but fail to deal with large document set efficiently .our candidate emails (in this paper this is referred as document sets) will be processed by first module and give extracted key terms only, as a result. And then these processed key terms will be passed to the second module as input which use fuzzy based logic to judge that the candidate email is spam or legitimate.

1) Document Analysis module:-

There are two stages in the first module, namely

i) Key Term Extraction

ii)Key Term Selection,

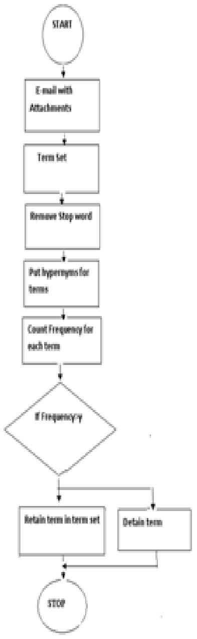
For reducing the dimensionality of the source document set so the document will contain only word which will be responsible for declaring any email as spam or ham or legitimate.

i). Key Term Extraction: the whole extraction process is as follows:

(1) First of all, each document is broken into sentences. Then, terms in each sentence are extracted as features. In this paper, a term is regarded as the stem of a single word.

(2) The terms appeared in a predefined stop word list are removed. Stop words are „a, an, the, was, „were etc. along with all removed prepositions, conjunction and articles from the data set D.

(3) Remained terms are converted to their base forms by stemming. The terms with the same stem are combined for frequency counting. Finally, the frequency of each term in each document is recorded.



ii) Key Term Selection: we understand that terms of low frequencies are supposed as noise and useless for identifying it as a spam or not spam. Thus, we apply the tf–idf (term frequency×inverse document frequency) method to choose the key terms for the document set. A term will be discarded if its weight is less than a fixed tf–idf threshold γ[15].

Following Formula is used for the measurement of tfidfij for the importance of a term tj within a document di. In Formula, fij is the frequency of tj in di, and $\max_{tj\in di}(fij)$ is the maximum frequency of all terms in di used for normalization to prevent bias for long documents.

$$Tfidfij=0.5+0.5*[fij/(\max_{tj\in di}(fij))]*Log(1+[|D|/|(di|tj\in di,di\in D|]$$

After the weight of each term in each document has been calculated, those which satisfy the pre-specified minimum tf–idf

Threshold γ are retained. Subsequently, these retained terms form a set of key terms for the document set D, and we formally define them in Definitions

Definition-1 (Document): A document, denoted di = {(t1, fi1), (t2, fi2),…, (tj, fij),…, (tm, fim)}, is a logical unit of text, characterized by a set of key terms tj together with their corresponding frequency fij.

Definition-2 (Document Set): A document set, denoted D={d1, d2,…, di,…, dn}, also called a document collection, is a set of documents, where n is the total number of documents in D.

Definition -3 (Term Set): The term set of a document set D={d1, d2,…, di,…, dn}, denoted TD={t1, t2,…, tj,…, ts}, is the set of terms appeared in D, where s is the total number of terms and tj is the stem of a single word.

Definition-4 (Key Term Set): The key term set of a document set D={d1, d2,…, di,…, dn}, denoted KD={t1, t2,…, tj,…, tm}, is a subset of the term set TD, including only meaningful key terms, which do not appear in a well-defined stop word list, and satisfy the Predefined minimum threshold of the tf–idf method.

Algorithm-1: Basic algorithm for key term Selection/Extraction:

Input-: An email with text attachments, A well defined stop word list ,WorldNet W, the minimum tf-idf threshold.

Output-: extracted key terms

1.Extract the term set considering whole content as a set of independent words these words are referred as terms.

2. Remove all stop words from the term set.

3.Convert all term to their base or standard form using hypernyms provided by Wordnet .

4. Count the frequency of each term by evaluating tf_idf_{ij} weight

5. If the term tfidfij>=γ then retain term in the
   Else
   Consider it as noise and detain it

6. Get the extracted terms.

After applying this algorithm to the candidate Email, we can have a precise set of keyword we named it Key Term Set. now we will apply fuzzy logic to this summarized set.

2) Fuzzy Filtration Module:

In this module document set will be examined for spam or not spam if the document set contains an element that is also a member of the black list set. We will apply different rules on the document to verify it.

Algorithm -2: Algorithm for Fuzzy Filtration

Input Variables : {Content key Word (email + attachments), Subject word, Sender's Address, Sender_IP}

Fuzzy Set : {positive, Zero, Negative}

Linguistic Set : {Highpositive, highnegative, Zero}

Step 1: [Rule-1: Fuzzy filtration based on Sender Address]

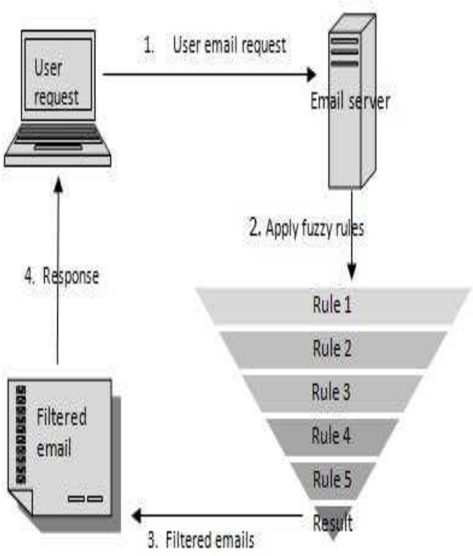
i): IF ∃ SenderAddress ∈ spammer list
Risk Factor=-0.25;

```
    ii): IF $\exists$ SenderAddress $\in$ to Ham list
        Risk Factor=0.25;
    iii) : IF $\exists$ Sender Address $\notin$ Spammerlist & $\exists$
        Sender address $\notin$ Ham address list
        Risk Factor=0

Step 2: [Rule-2: Fuzzy filtration based on Sender IP]
    i) :IF $\exists$ Sender_IP $\in$ spammerIP list
        Risk Factor=-0.25;
    ii): IF $\exists$ Sender_IP $\in$ to HamIP list
        Risk Factor=0.25;
    iii): IF $\exists$ Sender_IP $\notin$ SpammerIPlist & $\exists$
        SenderIPaddress $\notin$ HamIPlist
        Risk Factor=0;

Step 3: [Rule-3: Fuzzy filtration based on Subject Word]
    i): IF $\forall$ Subject words $\in$ Spam words
        Risk Factor= -0.50;
    ii): IF $\exists$ Subjectword $\in$Spamwords
        -0.50<Risk Factor< 0.50
Step 4: [Rule-4: Fuzzy filtration based on Content Word]
    i): IF $\forall$content words $\in$ Spam word list
        Risk Factor= -0.50;
    ii): IF $\exists$ content word $\in$Spamwords list
        -0.50<Risk Factor< 0.50
Step 5: [Rule-5: Fuzzy filtration based on Attachment]
    i): IF $\forall$ Attachment $\in$ Virus list
        Risk Factor= -1.0;
    ii): IF $\exists$ attachment $\notin$Virus list
        Risk factor=1.0;
```

4) Architecture of the proposed System:



When an e-mail is arrived, identified fuzzy input parameters are extracted and it is passed to fuzzy system for identification as per Figure1. After Fuzzyfication and Defuzzyfication categorized e-mails are send back to user. Rule 1 was applied on Fuzzy input parameter- Sender address. Based on Rule 1, Sender address was extracted from e-mail and compared against the Black list which has spammer e-mail address list. If any match was found then, Risk Factor for this rule was set to -0.25. If sender address was not found in the black list, then it was compared against the White list which contains all good and acceptable e-mail addresses. If match was found, then Risk factor for this rule was set to 0.25. If sender address was not found in both Black and White list, then attack factor for this rule was set to 0. Set this rule result in R1. Rule 2 was applied on Fuzzy Input parameter- Sender IP. IP Address of the sender was compared against the IP Address Black List. If match was found, then Rule 2 Risk Factor was set to -0.25. If not found, then Sender IP Address was compared against White List IP Address.

If match found then Risk factor of Rule 2 was set to 0.25. If not found then Risk Factor of the Rule 2 was set to 0. Assign resultant value in R2.Rule 3 was applied on Fuzzy input parameter- Subject words. An E-mail may contain one or more words in subject line. All subject word and Content words are preprocessed.

The pre-process contains the following steps i.e. stemming, stop words elimination and tokenization. Stemming is the process of comparing the root forms of the searched terms to the documents in its database. Stop words elimination is the process of not considering certain words which will not affect the final result.

Tokenization is defined as splitting of the words into small meaning full constituents. After pre processing all words are taken and compared against the Black list words. Every words impact (Risk Factor) on this subject line was calculated. From the subject line after pre-processing total words is counted and each word impact on for this rule is calculated. i.e. average impact. Now each word is compared against black and white list already available. If it is found in white list then the Risk factor for this word is set as positive. If it is found in black list then the Risk factor was set as negative. Rule 4 was applied on Fuzzy Input variable- Content Words after Pre-Processing. Every e-mail body may contain one or more words. Every word are taken and compared against the Block list words.

Rule 5 was applied to calculate Risk Factor for e-mail containing attachment. If e-mail does not contain Attachment, then Risk Factor was set to zero. If any one of the attachment content was identified in virus list then Risk Factor was set to -1. If none of the content was identified in virus list, then Risk Factor was set to 1. Rule 5 result was assigned to R5.

5) RESULT:

Result value of each e-mail was arrived by sum up previous rule results and these results are termed as decision making factors.

R1 = R1;
R2 = R2 + R1;
R3 = R3 + R2;
R4 = R4 + R3;
R5 = R5 + R4;

6) CONCLUSION AND FUTURE WORK:

In this proposed work, Fuzzy rules are constructed for 5 input parameters namely Sender's Address, Sender_IP, Subject_Words, Content Words and Attachment for common user to deduct the spam e-mails based on the attitude of the user. The proposed simplistic approach out performs in terms of accuracy in deducting spam e-mails than the existing approaches provided the Black list and White lists to be up to date. The proposed approach works only for e-mails having subject and body content as plain text. Future work aims at deducting spam emails having images and HTML also.

REFERENCES:-

[1] Cox, E., "The Fuzzy System Handbook", Academic Press, Second Edition, 1999.

[2] Sudhakar.P, Poonkuzhali.S, Thiagarajan.K and Sarukesi.K., "Fuzzy Logic for E-mail Spam deduction", Proceedings of the WSEAS 10th International Conference on Applied Computer and Applied Computational Science, Venice, Italy, March 8-10, 2011 ISBN: 978-960-474-281-3

[3] Poonkuzhali.S, Thiagarajan.K, P.Sudhakar Kishore Kumar.R And Sarukesi.K., "Spam Filtering using Signed and Trust Reputation Management", Proceedings of the WSEAS 10th International Conference on Applied Computer and Applied Computational Science, Venice, Italy, March 8-10, 2011 ISBN: 978-960-474-281-3

[4] Guzella, T. S. and Caminhas, W. M. "A review of machine learning
approaches to Spam filtering." Expert Syst. Appl., 2009

[5] H. Katirai. Filtering junk e-mail: A performance comparison between genetic programming & naive bayes. Available: http://members.rogers.com/

[6] J. Sedding, D. Kazakov, Word Net-based text document clustering, Proc. of COLING-2004 Workshop on Robust Methods in Analysis of Natural Language Data, 2004.

[7] Chun-Ling Chen , Frank S.C. Tseng , Tyne Liang An integration of Word Net and fuzzy association rule mining for multi-label document clustering" proceeding of the Data & Knowledge Engineering 69 (2010) 1208–1226

[8] Carreras, X. and Mdrquez, L., "Boosting trees for anti-spare E-mail filtering", In Proc. of RANLP, 2001.

[9] Li, K. and Zhong, Z., "Fast statistical spam filter by approximate classifications", In Proceedings of the Joint international Conference on Measurement and Modeling of Computer Systems. Saint Malo, France, 2006

[10] Cohen, W.W., "Learning Rules that Classify E-Mail.",Proceedings. of the AAAI Spring Symposium on Machine Learning in Information Access, Stanford, California,1996.

[11] Sadegh Kharazmi, Ali FarahmandNejad, Proceeding of the 9th WSEAS Int. Conference on Data Networks, Communications, Computers, Trinidad and Tobago, November 5-7, 2007.

[12] Tom Fawcett ,"In vivo. spam filtering:A challenge problem for datamining" KDD Explorations vol.5 no.2, Dec 2003. pp.140-148

[13] M. N. Marsono, M. W. El-Kharashi, and F. Gebali, "Binary LNS-based naïve Bayes inference engine for spam control: Noise analysis and FPGA synthesis", IET Computers & Digital Techniques, 2008

[14] Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali "Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification" Elsevier Computer Networks, 2009

[15] Chun-Ling Chen , Frank S.C. Tseng, Tyne Liang "An integration of WordNet and fuzzy association rule mining for multi-label document clustering " Data & Knowledge Engineering 69 (2010) 1208–1226.

[16] Hüllermeier, E.: Fuzzy methods in machine learning and data mining: Status and prospects. Fuzzy Sets and Systems. 156, 387-406 (2005).

[17] Anagha Kulkarni and Ted Pedersen, "Name Discrimination and Email Clustering using Unsupervised Clustering and Labeling of Similar Contexts", 2nd Indian International Conference on Artificial Intelligence (IICAI-05), pp. 703- 722, 2005.